PHD

The plastic genome of Bordetella pertussis

Abrahams, Jonathan Simon

Award date:
2020

Awarding institution:
University of Bath

The plastic genome of *Bordetella pertussis*

Volume 1 of 1

Jonathan Simon Abrahams

A thesis submitted for the degree of Doctor of Philosophy

University of Bath

Department of Biology and Biochemistry

March 2020

COPYRIGHT

Declaration of any previous submission of the work

The material presented here for examination for the award of a higher degree by research has not been incorporated into a submission for another degree.

Access to this thesis/portfolio in print or electronically is restricted until ………………. (date).
Signed on behalf of the Doctoral College...................................(print name)…………

Declaration of authorship

I am the author of this thesis, and the work described therein was carried out by myself personally, with the exception of Chapters 3 and 4 where less than 10% of the work was carried out by other researchers. This included: sequencing, genome mapping and manually resolution of CNVs in Chapter 3 and DNA/RNA extraction, qPCR and Nanopore sequencing in Chapter 4.

Acknowledgements

It is impossible to undertake a PhD alone and as such this work was not possible without close relationships with many people, both in my personal and academic life.

First and foremost, I would like to thank Andy Preston, my supervisor, who has always made time for me and has granted me the freedom to explore any project I desire. You are a fantastic scientist and mentor. I would also like to thank Andy Gorringe, my secondary supervisor, for his guidance, depth of knowledge and kind words of support during the project.

Further thanks for academic work and support are due to Iain MacArthur, Natalie Ring and Michael Weigand. This work would not be the same without their contributions.

The project was also shaped by stimulating discussions with Enrico Gavagnin and James Horton and I am grateful for their insight. In particular, I would like to thank Enrico for introducing me to network graphs which has now become a large part of this work.

The attention, support and faith of my friends Michael Carr and Elliot Druce had in me at every step of my studies is gratefully appreciated.

To my partner, Sali Morris, I am incredibly grateful. It would not be possible to have to get to this point without your patience and support. I am looking forward to repaying the favour when it is your time to write a thesis.

Lastly, I could not have undertaken a PhD without the unwavering support of my parents, Ian and Gillian, who are excellent role models and could not be more supportive.

Abstract

The paradigm that single nucleotide polymorphisms (SNPs) are the primary metric to judge bacterial diversity is outdated. This is particularly true for the main causative agent of whooping cough, *Bordetella pertussis*- a species with limited nucleotide variation. Examined in a holistic way, however, *B. pertussis* has high potential for genetic diversity with over 250 copies of the same insertion sequence- perfect genetic material for structural variations to arise via homologous recombination. Indeed, many deletions and inversions have been described which is in contrast to the third type of structural variation: CNVs (copy number variations), which have been only infrequently described.

In this thesis, I systematically investigated the prevalence and dynamics of CNVs (and other structural variants) in *B. pertussis* using both long and short read sequence data.
I developed a reliable pipeline to predict CNVs in >2000 isolates by analysing the read-depth of Illumina sequencing samples to find regions of increased coverage. A low rate of false positives and negatives was achieved by normalising inter-sample noise. The majority of these mutations were predicted to be >50kb long and clustered at 11 hotspot loci (rather than evenly distributed throughout the genome), a phenomenon described and analysed using network graphs.

One CNV was verified by qPCR and by capturing entire tandem arrays of CNVs in single ultra-long reads generated on the Oxford Nanopore sequencing platform. Further investigation demonstrated the plasticity of the *B. pertussis* genome and it was found that multiple putative structural variants were being generated genome-wide within a single culture.
Finally, preliminarily work established the compatibility of the *B. pertussis* with the Genome Wide Association (GWA) framework. I investigated how to represent CNVs and how homoplasic deletions were, given the highly clonal nature of the species and its low mutation rate. . It was found that deletions were more homoplasic than previously thought but that there are still considerable hurdles to using CNVs in GWAS.

Abbreviations

CNV- Copy number variant

Fha- Filamentous hemagglutinin

Fim- Fimbriae

GWA (S)- Genome wide association (study)

IS- Insertion sequence

PacBio- Pacific Bioscience

Prn- Pertactin

Ptx- Pertussis toxin

PtxP3- Pertussis toxin promoter allele 3

SNPs- Single nucleotide polymorphism

SV- Structural variant

1. Introduction

1.1. The Bordetella

Bordetella pertussis (*B. pertussis*), the primary agent of whooping cough (1), is a bacterium which has evolved enhanced pathogenicity without gaining new genes. It shares a recent common ancestor with modern day *Bordetella bronchiseptica* (*B. bronchiseptica*), the extant species that is the most similar to the progenitor of the ‘classical’ members of the genus, which in addition to *B. pertussis* include *Bordetella parapertussis* (*B. parapertussis*). *B. parapertussis* can cause a milder form of whooping cough and is composed of a human and an ovine host restricted clade. As a host generalist bacterium, *B. bronchiseptica* has been found in a number of mammals (including humans) and birds although some clades are host-specific (6, 7). It is also capable of causing a variety of respiratory pathologies ranging from nearly asymptomatic to life-threatening (8, 9).

Outside of these classical *Bordetella*, there exists a wide range of host restricted and host general species. Of note is *B. holmesii* which is distantly related to *B. pertussis* but causes whooping cough that is sometimes indistinguishable from that caused by *B. pertussis* but can also be milder (and sometimes causes bacteraemia) (1) . Other species capable of colonising humans include *B. hinzii* (2), *B. avium* (3) and *B. petrii* (3–5) which can cause a whooping cough like illness or milder symptoms, but most often in immunocompromised individuals. To understand how whooping cough effects the human population and how gene loss is linked to an unstable genome in *B. pertussis*, it is important to understand the disease in the context of the evolution of both *B. pertussis* and the whole genus. As such, this is the lens by which I will describe the work I undertook and introduce it.

1.2. Whooping cough

1.2.1. The course of whooping cough

B. pertussis is a human-restricted pathogen and the primary cause of whooping cough in humans. The full course of the disease can stretch up to 5-6 months and the symptoms can be split into three symptomatic stages: catarrhal, paroxysmal and convalescent (2). Prior to the catarrhal stage there is a 6-10 day incubation period (although this can stretch up to 3 weeks). The catarrhal stage of the disease is indistinguishable from many upper respiratory tract infections as it has generic and relatively mild symptoms such as dry cough, malaise and nasal discharge (3). This stage is the most infectious period of the disease and lasts approximately 1-2 weeks. Whooping cough is rarely suspected at this stage unless exposure to a known case has been ascertained (4–6). The severity and frequency of the cough worsens during this time until the most characteristic symptoms start to develop, in the paroxysmal phase.

Progression to the paroxysmal stage of the disease is characterised by significantly more troublesome symptoms, the most recognisable of which are: post-tussive vomiting, paroxysmal cough and ‘whooping’ between coughing bouts. This stage lasts 3-6 weeks and is the most characteristic of the disease (3). Coughing fits can be so intense that in addition to post-tussive vomiting, cyanosis, bulging eyes and haemorrhage can occur. Paroxysms can also leave children vulnerable to secondary infection or complications involving hypoxia (7). In neonates, paroxysms can be replaced by episodes of apnoea in which breathing stops. Neonates are the most at-risk population and suffer the highest mortality rates in all countries, vaccine schedules and time periods (8–10). Towards the end of paroxysmal stage, the coughs become less frequent and severe. This disease gradually transitions into the convalescent phase, which can last up to 3 months. The nasopharynx is still damaged however, and regular respiratory infections can trigger paroxysmal coughing during this period (9).

1.3. Diagnosis

NICE guidelines suggest that whooping cough should be suspected if the symptoms: paroxysmal cough, inspiratory whoop, post-tussive vomiting or undiagnosed apnoeic attacks in young infants are observed in addition to >=14 days of coughing (11). The disease can be officially diagnosed when either culture, PCR or serology tests positive for *B. pertussis*. The effectiveness of PCR and culture is limited as the bacteria are present at testable levels in the nasopharynx only for approximately 2 weeks after cough onset and levels decrease over the next two weeks. However, this is when the immune system is highly active against *B. pertussis* antigens and patients begin to seroconvert. Therefore serological testing is often the most reliable testing method as patients are seropositive for approximately a year after the infection (12–14). In terms of treatment, antibiotics are effective against *B. pertussis* and the species has low antibiotic resistance levels outside of China (15). Because of the disparity between symptom onset and the presence of the bacteria, however, current guidelines recommend administering antibiotics only within 3 weeks of onset of symptoms (11).

1.4. The evolution of the *Bordetella* genus

The *Bordetella* have been found in many environmental samples as many previously undiscovered species were found in environmental metagenomic data sets. Furthermore, when put on a phylogenetic tree, the environmentally associated species were found near the root of the tree, indicating that the ancestor of the *Bordetella* genus was environmental and that adaptation to hosts was evolved later (16). In support of the environmental origins of the genus, *B. bronchiseptica* is able to grow in soil and to not only survive phagocytosis by the amoeba *Dictyostelium discoideum* but be able to 'hitch-hike' on its spores and propagate (17). Survival in amoebae is thought to be a pivotal feature in other bacterial adaptation stories as amoeba can be thought of as a 'training ground' for macrophage survival.

1.4.1. The evolution of the *Bordetella* genus

1.4.1.1. Host-associated and host-specific clades of *B. bronchiseptica* reveal the evolutionary history of *B. pertussis*

Most clades in the *B. bronchiseptica* phylogenetic tree have a number of strains isolated from immunocompromised humans, although a small number were isolated from healthy individuals (18, 19). Some *B. bronchiseptica* clades are associated with specific hosts (although not necessarily exclusively restricted) and range from seals to dogs and pigs (18,20,21). This host association, rather than total restriction, suggests that *B. bronchiseptica* has retained the genetic pathways responsible for host generalism.

Multi Locus Sequence Typing (MLST) analysis of the classical *Bordetella* species categorised the genus into 4 clades, with clade 1 and 4 comprised of *B. bronchiseptica* and clade 2 and 3 comprised of *B. pertussis* and the human specific clade of *B. parapertussis*, respectively (Figure 1.1) (18). Phylogenetic analysis by Diavatopoulos *et al* established that *B. pertussis* was most related to a human-restricted clade of *B. bronchiseptica* (clade 4), indicating that they shared a common ancestor (18). It is likely therefore that the most recent common ancestor of these two groups had begun the process of host restriction before *B. pertussis* speciated and perhaps this meant it too had an increased tropism for humans before the speciation event.

Extant members of clade 4 *B. bronchiseptica* have defined their own pathway for increased pathogenicity as it was found that some clade 4 isolates of *B. bronchiseptica* were hyper-virulent, although there was considerable phenotypic diversity in the clade. When mice are intranasally infected with the reference strain of *B. bronchiseptica* (RB50, from clade 1), colonisation (but not symptomatic or lethal infection) occurs, bacterial load peaks at 10 days and then gradually decreases. Whilst some clade 4 isolates caused a similar infection to RB50, a number of clade 4 isolates caused lethal infection in mice (despite colonisation of mice with *B. pertussis* not being symptomatic) and histological examination of lung tissue of these infections showed widespread inflammation. Most clade 4 isolates were more virulent in vitro by measuring lysis of HeLa cells during infection, although the diversity of *in vivo* lethality was also reflected in these *in vitro* tests (22). No exclusive gene gain or loss events can differentiate clade 1 and clade 4 isolates of *B. bronchiseptica*, indicating that their phenotypic differences were likely caused by more subtle nucleotide changes.

1.4.1.2. The presence of repeat elements

One of the most striking differences between *B. pertussis* and *B. bronchiseptica* is the high level of insertion sequences in *B. pertussis*, but their sporadic appearance in *B. bronchiseptica* (23). Strains from *B. bronchiseptica* clade 4 have IS1663 present in only 80% of isolates (24,25). Whilst *IS1663* is present in *B. pertussis*, the *B. pertussis* genome is most markedly dominated by the approximately 250 copies of *IS481*; but also has a number of copies of *IS1002*. It is apparent, therefore, that at some point in the speciation of *B. pertussis*, additional insertion sequences integrated into the genome and started to accumulate. It is possible that the clade 4 isolates gained *IS1663* independently and there is not yet a consensus in the literature on this.

Insertion sequences were pivotal to help shape the *B. pertussis* genome into its current streamlined form (23) which is associated with increased host-restriction in the species, throughout the genus (24) and more generally in other bacteria too (26).This dramatic rise in repetitive DNA during the speciation process is synonymous with the increased pathogenicity of *B. pertussis*. The evolutionary trajectory of *B. pertussis* has taken the species to be the dominant cause of whooping cough- a deadly disease which has historically had high morbidity, mortality and infectivity. To some extent these factors have been reduced by vaccination, but understanding the natural history of the disease can lead to a greater understanding of the disease and the causative bacterium.

1.5. The speciation of *B. pertussis*: less is more

1.5.1. B. pertussis has recently speciated

The first historical accounts of whooping cough occurred between 1100-1600 (27–30). In these records it was described exactly the same as it is diagnosed today (although of course variably by each writer) with Thomas Willis in 1674 describing one of the symptoms as: "...inspiration and expiration being suppressed for a space the vital breath can scarcely be drawn; insomuch that coughing as being almost strangled by a hoop...." (31), an early reference to the characteristic 'whoop' made by patients during paroxysms. Despite sporadic mentions of a whooping cough like illness, there were no mention of epidemics in the comprehensive manuals of medicine published that were contemporary with the first accounts of the disease. It was found however, that there was considerable mention that whooping cough epidemics were occurring in Europe and Persia in the 16th century (27,32).

Whilst it is tempting to think of historical records as prohibitively incomplete, this late emergence is in contrast to many infectious diseases which have long and rich history. Evidence of tuberculosis has been found in Egyptian mummies from 2400BC (33,34) (but no contemporary records of the disease), in references written in Indian texts 3200 years ago (35) and is mentioned in the Old testament (36). The story is similar too for plague, with the first historical accounts (which can be verified using genomics to be due to *Yersinia pestis*) in 541AD for Justinian Plague (37–39), 1000 years earlier than for pertussis.

Later historical texts, beginning in the 18th century and 19th century, document dramatically increased cases of whooping cough, deaths and epidemics thus indicating increased prevalence and mortality (40,41). This coincided with increased movement of people and rapid human population increases (32). Early data from this time surprisingly echoes modern epidemiological observations and found a cycle comprising increased incidence every 3-4 years on a backdrop of persistent cases (40). Historical records are important because they can inform us on the evolution of the bacteria and provide context to modern data sources.

A landmark phylogenetic analysis of 343 isolates revealed that the likely speciation of *B. pertussis* was far earlier than its first historic mention (Figure 1.1). Bart *et al* report that Bayesian inference estimated the divergence of the two extant lineages of *B. pertussis* to be approximately 2000 years ago (42). It was found that of the 72% of pseudogenes shared between two isolates which represent these extant lineages, all had the same inactivation mutations. It could therefore be concluded that host-restriction was also at least 2000 years old. Further molecular clock experiments by Bart *et al* could date the expansion of the lineage 2b clade of *B. pertusiss*, which comprised 98% of the study and is the vastly dominant clade in circulation in the past century, to be in approximately the 18th century which is in agreement with the historical records of the first whooping cough outbreaks (42).

This is manifest in the population structures of these different bacterial species. TB is ancient but with a low SNP rate and therefore there is low diversity, but each isolate has many unique mutations (43). *B. pertussis* is clonal but all strains diverged recently, giving a population with very little variation (42). This is consistent with a tight bottleneck event during the host-restriction process leading to very little diversity. *Y. pestis* is not as clonal as either TB or *B. pertussis* and has a more diverse population (44).

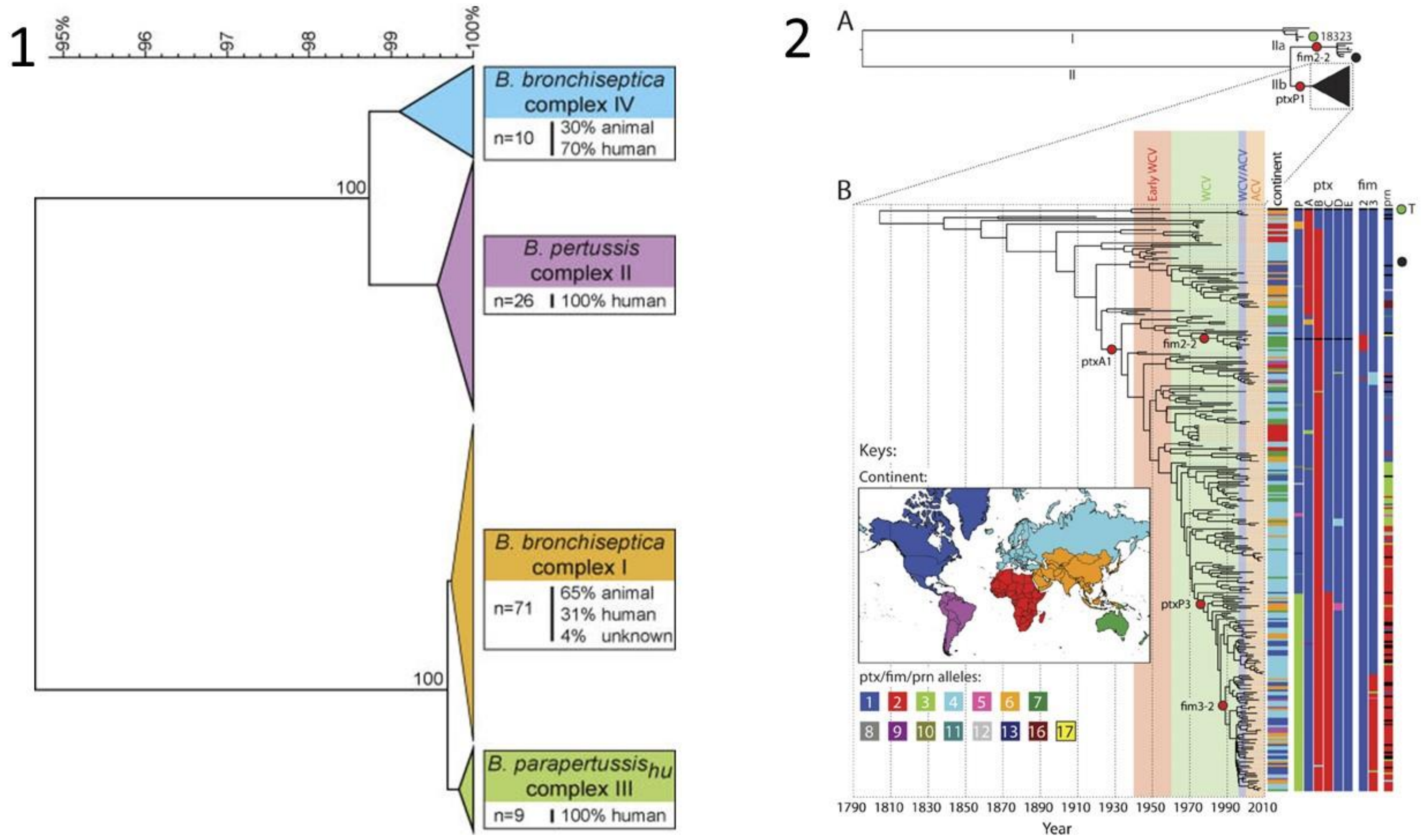


Figure 1.1. [1] MLST tree showing that *B. pertussis* is most related to *B. bronchiseptica* isolates from the human associated clade 4. [2] Bart *et al* studied the core genome SNPs (generated from Illumina sequencing) to show two deep clades of *B. pertussis*. The expanded section shows lineage 2b, which makes up the vast majority of circulating isolates. The vaccine eras are noted on the 2b subtree. Reproduced from Diavatopoulos *et al* and Bart *et al* (18,42).

1.5.2. Evolution by genome reduction: less is more

Whilst phylogenetic dating shows *B. pertussis* is a recently speciated, host-restricted pathogen, it does not show by what mechanism this has happened. This becomes clear when the size of the genomes are compared. The first published genomes of *B. pertussis* (4.1Mb) and *B. bronchiseptica* (5.3Mb) indicated that *B. pertussis* contained 1Mb less DNA than *B. bronchiseptica,* and *B. parapertussis* (human associated) (4.7Mb) contained 600kb less DNA (23). Genome reduction in *B. pertussis* has occurred primarily by recombination between homologous insertion sequences, of which there are approximately 250 copies of *IS481*, 17 copies of *IS1663* and 6 copies of *IS1002* in each *B. pertussis* genome (23). As *B. pertussis* shares a common ancestor with the human-associated *B. bronchiseptica* clade 4 isolates, the host restriction process likely predated the speciation of *B. pertussis* and was occurring in their common ancestor. Genome reduction was therefore likely a catalyst for host specificity rather than the sole driver (18).

In addition to gene loss, gene inactivation can occur when either a frameshift, premature stop codon mutation or when an IS translocates to disrupt the open reading frame. Both *B. pertussis* and *B. parapertussis* genomes have a considerable number of such pseudogenes, with *B. pertussis* having approximately 300 and *B. parapertussis* having 200 despite the reference *B. bronchiseptica* genome having only 13 (23). Recent RNA-seq analysis of *B. pertussis* does show that many pseudogenes are still transcriptionally active however (45, 46), thus indicating that despite being divergent from their original structure, the gene still may be contributing to the phenotype of the cell in some cryptic way. Disruption by frameshift mutations or insertion sequences may therefore be modifying the transcriptome in some cases, rather than just reducing it (expanded on below).

The maintenance of insertion sequences, particularly when they get to high copy numbers, is costly for the host bacterium (47). Insertion sequences can be seen as parasitic, existing purely to maximise their own success. It is likely that the initial copies of insertion sequences in *B. pertussis* were detrimental to the cell and were acting in a parasitic capacity. This is because there is likely no benefit to a single random insertion of an insertion sequence. Beyond genome streamlining, homologous recombination between insertion sequences can also cause rearrangements and CNVs (copy number variants) - processes can be beneficial to bacteria (reviewed in ‘The unique biology of structural variations’). In addition to structural changes, it has been found that some insertion sequences in the *B. pertussis* genome have outward facing promoters and therefore affect the transcription of neighbouring genes (46). It is likely that the co-existence of insertion sequences and bacterial cells is a multi-factorial relationship that has both benefits and costs to both sides.

It is likely that the transition to the human host was a bottleneck event that was pivotal to the expansion of insertion sequences and to the loss of genes in the species. This is because bottleneck events lower the diversity of the species and thus the power of purifying selection. This allows the ‘selfish’ insertion sequences to proliferate. This occurs for two main reasons. Firstly, living in the host as compared to the environment means a drastic reduction in the effective population size of the species. This decreases the competition and allows genetic drift to erode the genome via frameshifts or for the genome to tolerate IS expansion. Secondly, the change of environment means many genes are superfluous and therefore have reduced purifying selection acting on them and as such, IS transposition into these genes is tolerated. The transposition of insertion sequences can also occur into seemingly useful genes and it is viewed that these events are weakly deleterious and cannot be purged by the weak purifying selection. This means that a mixture of weak purifying selection and genetic drift allows the proliferation of insertion sequences, either because they provide useful benefits or simply because they cannot be purged (48–50).

All the same key virulence factors are present in *B. bronchiseptica, B. pertussis* and *B. parapertussis* and there is very little DNA that is unique to either *B. pertussis* or *B. parapertussis* (24). The increased pathogenicity and host restriction of these species are not due to the gain of new factors. *B. pertussis* has lost approximately 1000 genes whereas *B. parapertussis* has lost approximately 600 genes. Gene presence/absence in these species was compared to a manually curated database of gene functions and it was found that many of these genes were involved in membrane transport, metabolism, regulation and cell surface structures (23).

The idiom that less is more is certainly true for the pathogenicity of *B. pertussis*: a slimline genome is associated with increased pathogenicity. Other members of the genus have also shown similar (although not identical) evolutionary paths. Members of the pathogenic *Bordetella* genus can therefore be seen as a case study for a number of factors which are pivotal to the success of *B. pertussis*: evolution of host restriction and subtle changes to the function and regulation of virulence factors (18, 23, 24, 42). For example, whilst *B. holmesii* is not closely related to *B. pertussis*, it has also evolved by genome reduction and in fact has a smaller genome than *B. pertussis* by approximately 400kb. This has also been driven by IS expansion (with a different complement of IS, but including horizontally acquired *IS481*). Additionally, *B. parapertussis* appears to have a host restricted, ovine clade in addition to the human restricted clade and has undergone genome reduction.

1.5.3. The factors that drove the evolution of whooping cough caused by *B. pertussis*

Multiple host restricted pathogens emerged from a common ancestor in the *Bordetella* genus. It is hard to define exactly what factors allowed host restriction in these species or why *B. pertussis* appears to be the dominant causative agent of whooping cough, however. Although mechanisms of pathogenesis in *B. pertussis* are known, they are often shared between the classical *Bordetella* species and their function in other *Bordetella* species is less well described.

Common to all *Bordetella* species, is the two-component system BvgAS. The expression of many virulence factors in the classical *Bordetella* are under control of this system. It is often described that the system has three states: off (Bvg-), on (Bvg+) and intermediate (Bvgi) which respond to temperature (25^{o}C, 37^{o}C, intermediate respectively) and $MgSO_4$ concentration (51).

The Bvg+ phase also has early, middle and late genes (52–54). Whilst these categories and sub-categories are useful to describe observations, in reality a continuum of gene expression is induced by a continuum of these environments.

The Bvg+ state is clearly a virulent state, due to its induction at human body temperature and that it turns on expression of all key virulence genes. The role of the Bvg- state, which activates expression of genes which include motility and urease operons, has been linked to transmission, survival and persistence (55,56), such as in aerosolised droplets, the key transmission route of the bacterium (57,58). In *B. bronchiseptica*, the role of Bvg- is related to the environmental survival or environmental growth of an ancestor (17).

Animals prevent bacterial growth on mucosal surfaces by reducing free iron levels (Fe^{3+}), a key component in cell growth. In general, bacterial growth is supported at concentrations of free iron as low as 10^-7 M but in mucosal surfaces this is restricted to levels of approximately 10^-24 M (59). Therefore, the acquisition of iron is critical to the success of a pathogen. As such, the classical *Bordetella* have at least 3 systems: alcaligin, enterobactin and heme utilisation to scavenge iron from the host- the former two being siderophore pathways. *B. parapertussis* appears to lack the enterobactin iron acquisition pathway, however (60). In general, iron acquisition can be seen in all of the animal-associated isolates in the genus (in various pathways) and is absent from *B. petrii* (an environmental species).

On the mucosal surface, many factors are used by *B. pertussis* for adherence and cytotoxicity. Factors which are implicated in the adherence of cells include pertactin (PRN), fimbriae (fim) and filamentous hemagglutinin (FHA). In brief the function of these factors is as follows. Pertactin is an autotransporter which has a role in attachment (61), although it likely also has a role in immunomodulation (62). Fimbriae bind to monocytes and epithelial cells in the upper and lower respiratory tract (63). FHA binds heparin and thus allows binding to epithelial and macrophage cells (63).

Bacteria which cause severe disease, such as *B. pertussis*, often have more advanced systems to evade host immunity than bacteria that are opportunistic or only cause mild symptoms (64). For example, one significant difference between *B. pertussis* and the other classical *Bordetella* is the lack of O-antigen- an important factor in pathogenicity. *B. parapertussis* and *B. bronchiseptica* both have and express the LPS O-antigen genes whereas they are deleted from *B. pertussis*. The O-antigen is highly immunogenic and how this affects *B. pertussis* is not clear, but appears important for *B. parapertussis* to colonise hosts convalescent for *B. pertussis* by inhibiting antibody binding (65). In addition, systems have developed to evade complement killing in *B. pertussis*, such as binding the host complement regulator (C1-inh) which activates complement. *B. pertussis* sequesters this molecule on its cell surface, avoiding complement killing. This activity is absent in the other *Bordetella* species (66, 67).

B. pertussis encodes a number of toxins which cause damage to the host epithelium and contribute to the development of whooping cough. Adenylate cyclase causes haemolysis but also can bind to a variety of cells, bind to calmodulin and cause an overproduction of cAMP, depleting the cells of ATP and impairing their function (68). One of the main toxins known in *B. pertussis,* and one that it exclusively produces within the classical *Bordetella*, is pertussis toxin. Pertussis toxin is secreted by a type 3 secretion system and once inside host cells, blocks inhibition of adenylate cyclase activity, leading to an increase in cAMP in the cell. The factor responsible for the bacterium to trigger such intense paroxysms in the host has not been found (68), however. Other toxins include dermonecrotic toxin and tracheal cytotoxin, which also play roles in pathogenesis, although their contribution is not well known. Pertussis toxin is so key to the pathogenesis of *B. pertussis*, that acellular vaccines containing just this toxin are of comparable efficiency to 5 component vaccines (69).

Beyond the mechanisms thought to be key to the development of whooping cough, the function of genes in *B. pertussis* is under-studied and this is also true of the other human-restricted pathogens in the genus. Therefore, considerable work needs to be undertaken to ascertain which genes (and mutations) are critical to human pathogenicity of the genus and full development of whooping cough. Further investigation into genotype-phenotype links, as is needed in the *Bordetella*, can be undertaken on a large scale using the framework of Genome Wide Association Studies (GWAS) (70). This is a statistical framework that strives to link genotypes to phenotypes whilst controlling for the underlying population structure. This is complicated by the fact that bacteria reproduce largely asexually and thus have a very strong phylogenetic relationship with most mutations occurring together, since there is no 'scrambling' of the genome during sexual reproduction. Traditionally, GWAS is undertaken using the SNPS or gene presence or absence as genotypes. Technological and theoretical advancements have meant that recently, other mutation types such as small indels can also be included by analysing the genome as K-mers (K- length sequences) which present a unified framework to study all mutations. The work in Chapter 3 includes an investigation into GWAS for *B. pertussis* using structural variants as a genotype in order to test the suitability of the framework for use with the species.

1.6. The evolution and epidemiology of B. pertussis in the era of vaccines

Evolution is clearly not a static process and as such is continuous. The previous 100 years have seen dramatic changes in the incidence of whooping cough and the evolution of *B. pertussis* in response to vaccination schedules (71). Whilst the incidence and epidemiology of the disease is linked to the vaccine schedule, the link between genetics and vaccination schedule is not clear (72). I review these factors briefly in the pre-vaccine era and in-depth in the whole cell and acellular vaccine periods which are most relevant to understand the disease of today.

1.6.1. Pre-vaccine era (1800's-1940's)

The 19th and early 20th century saw drastic increases in the number of whooping cough cases and epidemics (31). By 1910 it was a huge public health burden with mortality at 10% for children under 5. At approximately the same time *B. pertussis* was confirmed to be the causative agent and within just 30 years, by 1940 (73), the first chemically inactivated whole cell vaccine had been extensively rolled out for children. This ushered in a new era of the way whooping cough was considered by society and altered its epidemiology.

1.6.2. Whole-vaccine era (1940’s-1990’s)

1.6.2.1. A drastic reduction in disease

Whole cell vaccines were incredibly effective at reducing the incidence of whooping cough in countries with good vaccine coverage. The incidence of whooping cough fell from approximately 115,000 to 270,000 cases a year (with 5000-10000 deaths) in 1940 to 1200 to 4000 cases (with just 5-10 deaths) a year in the 1980’s in the US (74). This was therefore a highly successful programme and representative of many other vaccination success stories, although the success fluctuated over the years.

1.6.2.2. A shift in demographic

Whilst the vaccine schemes were effective in reducing the number of cases of whooping cough, the disease was not eradicated and carriage of the bacterium likely still persisted. Early studies suggested that carriage was still ongoing as the 3-4 year spikes in incidence, clearly seen in the pre-vaccine era, were still ongoing (75). This is in contrast with measles, where an increased cycle time was noted.

In this era, efforts to reduce the incidence of the disease focused on rolling out vaccines to as many as possible, and 79% of British children were being vaccinated by 1973. There were, however, vulnerable parts of the global population who could not be vaccinated effectively, including neonates and developing countries in which vaccine coverage was not optimal. It is interesting to note that a change in demographic was recorded in this period. It was noticed that adolescents were becoming more vulnerable to the disease as early as the 1960's (76). The longer the time since last vaccination was an important factor to determine the attack rate of the bacterium. There were also an undetermined amount of cases in adolescence and adults (77–79) and mild cases of the disease (78).

1.6.2.3. Vaccine scares and an aversion to 'crude' vaccines

As the cases of whooping cough dropped by 99%, the public lost the fear of the disease, despite the era of extremely high prevalence still being within living memory. A lot of attention was instead paid to both real and perceived side effects of the vaccine. Real side effects (although temporary) included fever, limb swelling and persistent crying (80–82) whilst notable and tenaciously associated perceived side effects were sudden infant death syndrome (SIDS), neurological disorders (84, 85) and febrile seizures. Many of the real side effects were associated with the highly antigenic cell-wall components of the whole-cell vaccine. These effects caused increasing concern about the vaccine and thus vaccination rates fell UK-wide to just 30%, a trend that was echoed in many developed countries with some stopping vaccination entirely (85). Predictably, this caused outbreaks almost immediately with Sweden (with a fully suspended vaccine programme) experiencing two outbreaks in 1983 and 1985. There were more than 4 times more cases in 1981 compared to 1985, which corresponded with 2 and 6 years after the cessation of vaccinations and reflected the demographic of whooping cough in unvaccinated/semi-vaccinated populations (86, 87). The UK suffered similar outbreaks between 1978 and 1982 (87). With the public distrustful of the whole cell vaccine and whooping cough cases on the rise there was an urgent need for a new vaccine with fewer side effects.

1.6.3. Acellular vaccine era (1990- present)

Work began on making acellular vaccines composed of varying combinations of up to 5 key antigens: pertussis toxin (PTX); filamentous haemagglutinin (FHA), pertactin (PRN) and fimbrial protein 2 and 3 (FIM2/3). Early efficacy trials demonstrated such vaccines provided similar short-term protection to whole cell vaccines and fewer side effects and they were rolled out in most developed countries between 1990 and 2005. At first these acellular vaccines replaced the whole cell vaccine booster vaccinations in a hybrid whole cell and acellular vaccine schedule and then a few years later acellular vaccines replaced all use of whole cell vaccines in many developed countries (88). In developing countries, whole cell vaccines were still used and many still use them today.

1.6.3.1. A demographic change to adolescence and neonates

The switch to the acellular vaccine is correlated with an increase in cases of whooping cough and a shifting of the demographic. The ACV is known to grant a shorter period of immunity than the WCV or natural infection. A meta-analysis showed that for every year after an ACV booster, the odds of contracting the disease increased by 1.33 and that after 6-8 years from the last booster, only 10% of children would have sufficient immunity (89). In addition, it was found that adolescents who had received solely the acellular vaccine were 6.6x more likely to get the disease than those who had received only the whole cell vaccine and that those receiving a mixture had directly proportional risk to how many acellular boosters they had received. This fits well with the epidemiological data which shows that since the introduction of the acellular vaccine, there has been a large increase in the number of adolescent whooping cough cases. The waning immunity provided by the vaccine was not related to how many antigens are included in the vaccine, how many booster shots are given or the method by which the vaccine was produced and is supported by a number of studies (91, 92).

1.6.3.2. The resurgence of whooping cough

The introduction of the acellular vaccine coincides with global increases in incidence of whooping cough. For example, in 2012 there were large outbreaks in both the UK and the US, the worst, in some cases, for 60 years. An holistic examination of the data indicates that cases started to increase 15 years prior to the introduction of the acellular vaccines, despite the prevailing thought being that the acellular vaccine is the main factor in the resurgence of the disease (93, 94). It is therefore possible that the increased awareness of the disease and improved testing by PCR and serology was contributing to an increase in reported cases in addition to the waning immunity from the ACV.

One of the problems of the ACV is that they do not prevent colonisation and transmissions. Infant baboons were given a standard course of ACV and when challenged with *B. pertussis* did not show any symptoms of whooping cough, but were colonised by *B. pertussis* and able to transmit the bacterium to other baboons (94). Further evidence for asymptomatic or cryptic transmission of the bacteria is that between 12% and 30% of adults reporting prolonged cough are seropositive for *B. pertussis* (94). This means that *B. pertussis* is likely in wide circulation within the human population which allows transmission to at-risk populations such as neonates. In the UK, however, neonates are protected by immunisation of pregnant mothers, a strategy that seems to be highly effective (96, 97).

1.6.3.3. The weakness of acellular vaccines

Unfortunately, antimicrobial drugs do not enjoy the long lifespan that vaccines do as most vaccines have remained effective for many decades. Some vaccines, however, have not had long life spans and it appears there are two key factors that in influence vaccine escape of a bacterium or virus: the ability of the vaccine to stop colonisation/transmission and the ability of the vaccine to raise a sophisticated and complex immune response against the pathogen. The acellular pertussis vaccine fails both of these tests (97).

Vaccine escape is not often seen because vaccines often raise an immune response to a wide range of antigens. Antimicrobial resistance is much more common partly because its use is not prophylactic (allowing large populations to exist and evolve mutations) and partly because they target either a single residue or small number of residues. Vaccine escape has been seen previously in a few rare cases where vaccines have been designed to (sometimes inadvertently) target a small number of bacterial or viral molecules(97). For example, in *Yersinia ruckerii*, a pathogen of salmonids, an inactivated whole-cell vaccine was produced using currently circulating flagellated isolates. Isolates could evolve resistance to the vaccine that had multiple mechanisms of action by a small number of mutations: Single mutations in the bacterium could turn off expression of the flagellum, which did not affect pathogenicity (98, 99).

The jettison or inactivation of genes by the pathogen that are included in an acellular vaccine can be seen patently in *B. pertussis*. In particular, the pertactin gene has undergone homoplasic deletion/inactivation mutations and these mutations are clearly under positive selection as many countries report a rapid rise in Prn-negative isolates. For example, in the US it was found that 85% of isolates were Prn-negative (99) and in Europe it could be shown that pertactin deficiency rose from 1.9% in 1998-2001 to 25% in 2012-2015 (100). This is clearly alarming but it is known that Prn is not crucial to pathogenicity of *B. pertussis* and as such has been omitted from many acellular vaccine formulations that comprise only two or three components.

Pertussis toxin has been considered to be important in the pathogenicity of whooping cough, but it is possible to isolate strains that have deleted or inactivated the *ptx* operon (responsible for pertussis toxin production) and do not produce the toxin (101–103). These isolates are rare however, and less than 5 have been previously isolated. A screen of over 300 strains for pertussis toxin production did not see any isolates beyond these few to be defective in pertussis toxin production (104). Filamentous haemagglutinin deficient strains have also been described (105). There is therefore considerable potential for isolates to evade the immunity generated by acellular vaccines. It appears that the prn-deficient isolates are fitter in acellular-primed hosts but at a disadvantage in naïve hosts (62, 106, 107).

Crucially, it could be seen that the acellular vaccine could not prevent colonisation or transmission (94,108). This was also inferred for the whole cell vaccine too, due to seroprevalence and mild cases of whooping cough (109,110). Colonisation allows the bacteria to accumulate to high levels in the host giving ample time for genetic diversity to accumulate and the selective forces of the vaccine primed immune system to select for vaccine-escape mutations.

1.7. The genetics of modern *B. pertussis*

An understanding of both the natural history and the evolutionary history of *B. pertussis* can shape the knowledge of its evolution in the current era. The first *B. pertussis* genome sequence was generated in 2003 from the reference Tohama strain (23). Tohama had been the reference strains for some time and was isolated from a Japanese patient in the 1950's. Because of its age, it is now not representative of circulating isolates and may also be adapted to laboratories due to passaging (111). The use of Tohama continues, but a closed genome sequence is available for B1917 which is now being promoted as a new, more representative reference genome (112) as it was isolated from a Dutch patient in 2000. B1917 has the promoter allele for pertussis toxin (ptxP3) which is found in the majority of many modern circulating isolates and is therefore representative of such isolates, although the ptxp3 allele may be a marker for other mutations (see Chapter 5 and 'Key alleles associated with vaccination schedule' below).

When *B. pertussis* entered the human population, which was at least 2000 years ago (42), it is thought that this was associated with a dramatic bottleneck event that decreased genetic diversity and gave rise to the current population structure of the species today. This is one of the main factors that influence the population structure and genome of *B. pertussis*, in addition to a low SNP rate and a small accessory genome. Factors which are detailed here.

1.7.1. Low SNP rate

The low SNP rate of *B. pertussis* was shown by Bart *et al* (42) who studied a diverse set of 343 isolates from different time periods. A SNP rate of 2.24x10^ -7 per site per year was found and reanalysis by a study of mutation rates in 38 diverse bacterial species (including the Bart data) supported this (115, 116). Duchêne et al also found that in comparison to many other species (such as *Enterococcus faecium*, *Staphylococcus aureus* and *Streptococcus pyogenes*), the mutation rate of *B. pertussis* was found to be an order of magnitude lower (113).

The ratio of synonymous (SM) to non-synonymous mutations (NSM) gives an insight into the selection pressures acting on the species. This is because genetic drift alone would cause both mutation types to occur at the same frequency whilst purifying selection would remove non-synonymous mutations, allowing synonymous mutations to occur at a higher frequency. Sealey et al used the ratio of SM and NSM in a dataset of 100 UK isolates to identify that genes encoding proteins included in the acellular vaccine were evolving at faster rates than other cell-surface genes (a comparable control group as cell-surface proteins are generally under positive selection to avoid host immunity) (114). The mutation rates in these genes were highest after the introduction of the acellular vaccine in many parts of the world. Overall, the findings of the study supported the idea that the genes encoding proteins in the ACV were under positive selection as the ratio of NSM to SM was higher than in cell surface genes, although it wasn't clear if the higher rate of mutations in the ACV era was due to selection pressures or enhanced mutability.

This work by Sealey et al was extended by Etskovitz et al who compared the SM/NSM ratios between only the 5 genes encoding products of the acellular vaccine (115). They found that FHA and PRN were under positive selection but that PTX genes were under purifying selection. This sheds light onto the Sealey et al study by showing that overall, these genes may be under positive selection, but that masks a more subtle pattern between the 5 regions.

It therefore appears that some of the genes involved in coding for the proteins contained in the acellular vaccine are under positive selection, likely in response to the increased selection pressure of the acellular vaccine. It is unclear, however, the selection pressures acting on other sites in the genome or how much of this effect can be explained purely by time (42, 119). There is therefore more work to be undertaken to establish what selection pressures shape the modern population of *B. pertussis* (in particular, outside of the well-studied virulence genes) and how tolerant the genome is to deleterious or neutral mutations.

1.7.2. Key alleles associated with vaccination schedule

Pressure acting on key virulence genes was described when it was found that novel alleles rapidly displaced older alleles in selective sweeps at least 4 times in the period of 1949-2010 in the Netherlands (71). This result was echoed by other studies in other countries, including Australia (117–119). These findings are consistent with the theory that the *B. pertussis* population is under selection by vaccine-induced immunity for key virulence genes.This theory is not universally supported and many speculate that many other factors could be partially or wholly responsible.

One of the alleles to become dominant has been ptxP3, caused by a single SNP (compared to ptxP1, the previously dominant allele) in the promoter region of the genes encoding the pertussis toxin. PtxP3 is now the dominant allele in nearly all countries with the acellular vaccine and whilst it was first found in the 1980's it gradually replaced ptxP1 in the 1990's (71). Whether this allele is the *ptx*P3 mutation has occurred only once on the phylogenetic tree and as such is associated with many other mutations and their contribution to the phenotype attributed to *pt*xP3 is not clear. *Ptx*P3 is associated with both enhanced virulence in vivo (increased colonisation), in vitro (increased toxin production and macrophage killing) and statistically (increased hospitalisations) but when the *ptx*P3 allele was created in the genetic background of a *ptx*P1 isolate, the in vivo virulence associated with *ptx*P3 was not recapitulated- an elegant experiment that demonstrates the importance of other genomic features (120).

1.7.3. An accessory genome by gene loss

Many species have highly variable gene contents between isolates. This can be described as the core genome, the part of the genome found in 99% of all isolates, and the accessory genome, genes which are variably present. The combined core and accessory genomes are known as the pangenome. Well known examples of species with extensive accessory genomes include *Klebsiella* and *E. coli,* which have core genomes of 17% (121) and <10% (122,123), respectively. In these species, the creation of large accessory genomes is driven by horizontal gene transfer, although this is not the case in *B. pertussis*.

B. pertussis has a large core genome of approximately 90% that is driven purely by gene loss. On-going genome reduction causes considerable variation between circulating isolates in their gene complements. The categories of genes that were deleted ancestrally (membrane transport, metabolism, regulation and cell surface structures) are the same categories to which genes that are being deleted among recently circulating isolates belong, indicating that the same selection pressures that gave rise to the species are also acting now. In addition, it can be shown that pseudogenes are over-represented among those genes in recently occurring deletions (18%) compared to the whole genome (10%) (71,120) and that recently isolated strains have smaller genomes that older isolates. This indicates that the *B. pertussis* genome is undergoing continual streamlining, perhaps to purge defunct pseudogenes. Gene deletion therefore creates genetic diversity in the species, although modest.

1.8. The unique biology of structural variations

Structural variants (SVs) are mutations which alter the order of the DNA in large contiguous blocks. Whilst these mutations can cause minor alterations to the nucleotide composition at the points at which the mutation occurs, their effect is mostly on the intervening DNA. There are three main classes: deletion, CNV and inversion. The formation of structural variants relies on the intrinsic mechanisms of DNA and DNA replication, an appreciation of which is required to understand how SVs are formed.

1.8.1. How SVs are formed

1.8.1.1. Homologous recombination as a chromosome maintenance mechanism

Structural variants are formed through errors in routine DNA maintenance and replication pathways, much the same as single nucleotide polymorphisms or small insertions/deletions and therefore are classed as mutations. It is common to class structural variants as mutations (124,125) and as such I will refer to them this way in this thesis. The mutational pathways that lead to the formation of structural variants are described here.

DNA damage leading to single stranded breaks or double stranded breaks in the DNA double helix can be repaired by a variety of pathways, including homologous recombination, non-homologous end-joining, nucleotide excision repair and mismatch repair. Each pathway is useful for repairing different types of DNA damage whilst also being able to non-specifically repair any damage (mismatch repair, nucleotide excision, non-homologous end-joining) or repair only specific sections of DNA (homologous recombination) (126–129).

Homologous recombination is responsible for the genesis of most of the known structural variants in *B. pertussis* and so is reviewed extensively here (130–132). Homologous recombination is a mechanism of DNA repair primarily but secondary effects include the transfer of DNA inter-chromosomally or intra-chromosomally. The majority of structural variations are generated by the process of homologous recombination, despite it playing a minor role in DNA maintenance overall. But the process is also involved in horizontal gene transfer (133,134).

There are two paths of homologous recombination in bacteria that are used as an ‘average’ mechanism (Figure 1.2), although there is considerable diversity in this process. They deal with single (ss) and double stranded (ds) breaks in DNA. Both involve the Rec genes with dsDNA breaks being repaired by the RecBCD genes and ssDNA breaks repaired by the RecFOR genes. Both pathways use RecA to bind homologous sequences (125, 126) but have subtly different mechanisms.

Double stranded DNA breaks can be caused by a variety of mechanisms, such as UV light or chemical mutagenesis. DNA lesion or single stranded breaks can also be turned into double stranded breaks through the DNA replication process. During replication, the replication fork progresses from the origin of replication to the terminus but will encounter DNA lesions or single-stranded breaks en-route which cause a stall in the replication complex. Because the DNA replication process involves synthesising a new DNA strand for each of the parental strands, when the DNA replication fork reaches the single stranded break the leading strand (and its newly synthesised complementary strand) will disassociate causing: a double stranded break, replication fork demise and the creation of two separate dsDNA molecules (129, 130, 135). When a double stranded break occurs, the RecBCD complex is recruited (128). RecBCD binds, unwinds and degrades the DNA until it reaches an ‘X’ (Chi) site (analogous to a checkpoint) (135). At this point RecBCD synthesises an extension to the leading strand in the form of a new single stranded DNA coated in RecA (136,137). The reaction then proceeds to the RecA process (128,138,139).

Alternatively, when a single stranded break has occurred the enzymes RecFORJQ are involved in recruiting RecA. The RecJ exonuclease enlarges the single strand break thus creating an exposed ssDNA portion of the DNA molecule. The RecFOR complex then recruits RecA to the exposed ssDNA and the reaction proceeds to the RecA process. Although more complex events can lead to further modifications and repair pathways (140,141).

The RecA coated DNA (either newly synthesised DNA in the case of dsDNA breaks or newly degraded DNA in the case of ssDNA breaks) binds to a homologous region. This may be the site where the DNA replication fork stalled (See ‘replication restart’ in Figure 1.2), homologous region of the same chromosome or a sister chromosome generated during DNA replication that is still within the cell. The RecA coated single stranded molecule invades the target DNA and binds to it (130, 144, 145). This forms a complex junction (Holliday junction) between two double stranded DNA molecules (if recombination between two different DNA molecules) or section of the same DNA molecule (if intrachromosomal recombination). The junctions are then cut to resolve the junction (144–146). Whilst homologous recombination is an effective pathway for DNA repair, a by-product of this results in the generation of genetic diversity (124,147,148).

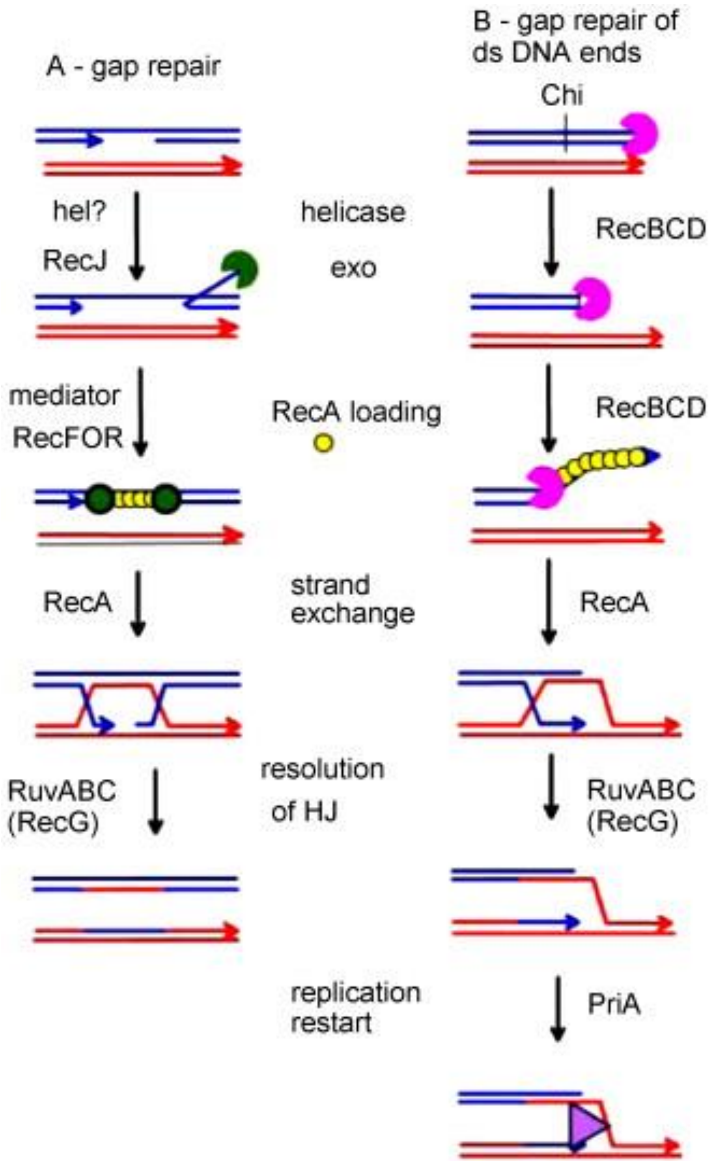


Figure 1.2. Schematics of ssDNA (gap) repair and dsDNA repair by homologous recombination via RecA. Reproduced from: “Recombination proteins and rescue of arrested replication forks” (144).

1.8.1.2. How deletions, duplications and inversions are formed by homologous recombination

If the two homologous sequences are in the same orientation, then a deletion or CNV can occur (Figure 1.3). If this occurs between two molecules of DNA (such as between sister chromosomes during DNA replication), one molecule will donate the recombined region to the other and experience a deletion whilst the recipient molecule would receive the extra DNA and have a CNV. If a deletion occurs between different sites on the same chromosome then the intervening DNA is ‘looped out’ and forms a non-replicating circular DNA molecule (124,148). If the two DNA regions are in the opposing orientation on the same molecule of DNA then an inversion can occur. It is not possible for inversions to occur between different DNA molecules (Figure 1.3).

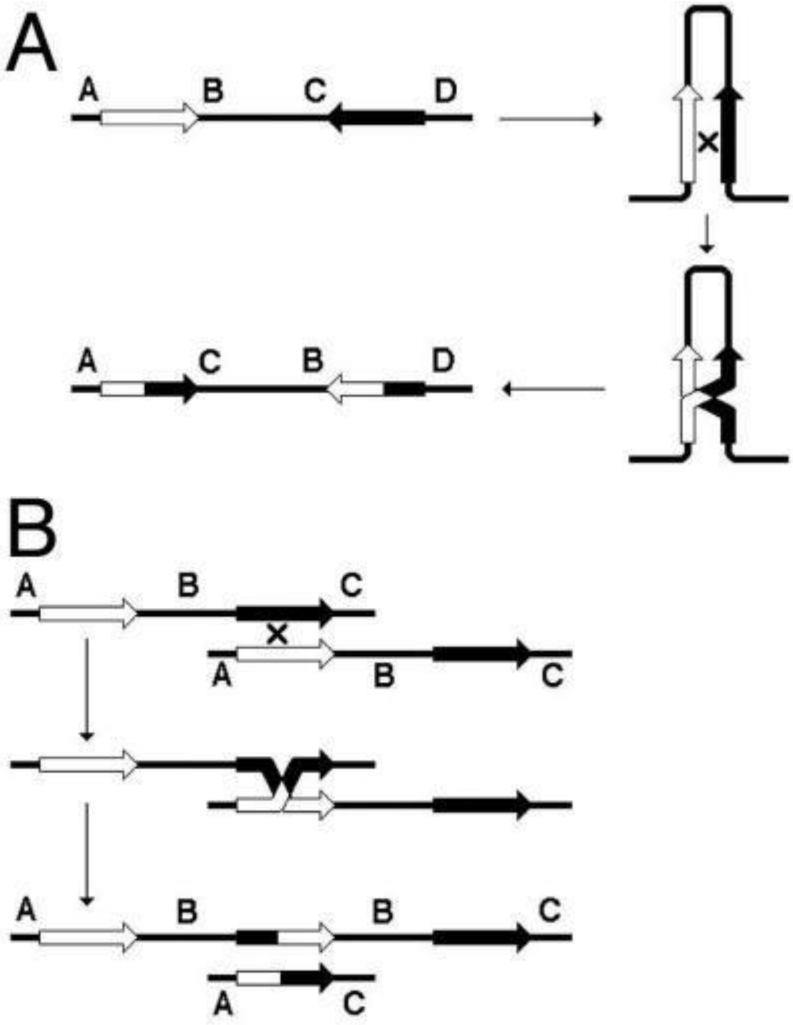


Figure 1.3. Repeats in opposite orientation on the same molecule recombine to form an inversion (A). Repeats in the same orientation on different molecules recombine to form a Holliday junction which is then resolved to give one molecule a duplication of the intervening sequence and the other molecule a deletion (B). Adapted from: “Homologous Recombination—Experimental Systems, Analysis and Significance” (149).

1.8.1.3. Rarer mechanisms of SV formation

SVs can also form by processes which are collectively known as illegitimate recombination. This occurs when aberrant DNA complexes form from incorrect annealing during DNA processing, rather than from a break in the DNA. This type of recombination may require a small section of homology or could be carried out with no homology. This can lead to deletion, duplication or inversion of the DNA (150).

When only one copy of the bacterial chromosome is present (precluding homologous recombination for most areas of DNA in the chromosome), non-homologous end joining (NHEJ) can be used to repair dsDNA aberrations. This process requires homology of only 4-12 base-pairs and sometimes can be undertaken on areas with no homology. The system is not present in all bacteria, however and its impact in the bacterial kingdom is unclear. Structural variants can occur from this process when more than one double stranded break occurs in the genome and the breaks become incorrectly ligated to one another (150–153).

1.8.1.4. Amplifications from duplications

Once a gene duplication has taken place, the tandem array can further amplify to greater copy numbers. This can occur by the standard homologous recombination pathway or through the related rolling circle amplification pathway. In the homologous recombination scenario, amplification from a tandem array can proceed using homologous recombination between random copies in the array, the full array or only part of it. This can mean that the array can amplify nearly exponentially, if a second copy of the full array enters the recombination process. This will occur at the rate of 2n-1, as the second array must share an overlap of at least 1 repeat with the first array. At this maximal rate, where the whole array recombines, a duplication can turn into an array of 100 repeats in 7 generations. Taking into account ‘sub-optimal’ recombination between random subunits of the array means 100 copies could be obtained in approximately 11 generations (124,154–156).

It has been observed that amplification of tandem arrays can occur faster than the maximum speed that Rec-A dependent mechanisms can provide. Petit *et al* found that arrays can be amplified by the rolling circle amplification pathway. As the DNA replication fork progresses through the tandem array, recombination between the leading strand-based molecule and the lagging strand-based molecule can cause a circular DNA molecule to form with a trapped replication fork. The replication fork then replicates this circular molecule which is highly unstable and non-replicable. This intermediate molecule can be stabilised and made heritable through further homologous recombination with the remaining array (or potentially single locus) left on the main chromosome of the cell (155,157). It has been found that large tandem arrays are commonly made from sequences that are less than 40kb long, although it is unclear why this might be, considering that an array of 100 copies of a 40kb region would extend the genome by the same amount as 10 copies of an 400kb region (124,156).

1.8.1.5. Selection acting on CNVs

Whilst the work in this thesis concerns mainly the highly unstable tandem arrays of duplications, these events are the most frequent source of new genes with new functions. The early models of gene duplication proposed that following the duplication of the gene, the second copy is free to evolve a new function because the original copy of the gene continues to fulfil the original role. This model (known as the neo-functionalisation model) makes sense only if the second copy of the gene is under neutral selection pressure from its inception so that it can occur at a sufficient frequency and time to acquire new mutations. For example, if a perfect second copy of the gene was deleterious at its inception but this fitness cost could be ameliorated by a single SNP that was under positive selection for a secondary function, the second copy of the gene would be selected against before the new function could evolve. This is known as Ohno's dilemma, after it was Ohno who proposed the neo-functionalisation model (158–160).

A more sophisticated model was proposed to explain the fate of duplications which takes into account the continuous selection pressure that acts on all genes-the Innovation, Amplification and Divergence model (IAD) (158). Genes have primary functions which they are under selection for, but also secondary functions which may have residual activity and/or are neither positively or negatively impacting the cell before the IAD process starts. When a change in environment occurs and these secondary functions of the gene are beneficial to the cell (or alternatively, secondary functions of the gene occur by point mutations during a constant positive selection for this other function), they are now under positive selection (158). This is the innovation stage. In this Amplification stage of IAD, because gene duplication events are much more common than point mutations (see below), an increase in the activity of these secondary functions of the original gene is likely to come from gene duplications rather than point mutations(158).

The new selection pressure the cell is under can also select for point mutations in any of the copies of the original gene. This is the divergence stage of the IAD process. Because of these extra copies of the gene there are also extra chances for the mutation to occur due to increased mutation targets. Once beneficial mutations accumulate in one of these copies, the selection on the secondary function of the original gene and its identical copies reduces. This is because tandem arrays are a fast but crude way to respond to the change in environment which can be easily out-competed by subsequently occurring more elegant solutions. For example, amplification of an array of genes to 50 copies may cause 50x as much activity in the secondary function but a single SNP may mean the enzyme coded for has 50x more activity in this secondary function. In this scenario both mutations cause the same effect, but the SNP is more efficient as less product is produced. Subsequently, the array may reduce in copy number or additional copies may become inactivated (161, 164–166).

Experimental studies have found that most gene duplications are under purifying selection given that, over a long period of time and a stable environment, the function and level of expression of the gene is generally well adapted and thus a disturbance to this is deleterious (164). It has been found that duplication (or tandem array) size is not necessarily correlated with fitness cost but instead it is the disturbances to gene expression and regulatory networks which are under strong purifying selection themselves. In this regard, it is not the cost of replicating and maintaining extra DNA, but instead the cost is associated with superfluous protein synthesis which is an energy highly intensive task (165). The work in this thesis focuses on mainly the short-medium term tandem arrays of identical genes, the ‘amplification’ stage of the IAD model (158).

1.8.1.6. SV formation occurs more frequently than other mutations

From experimental systems using *Salmonella enterica* and *E. coli*, species with a low number of repeats, the frequency of duplications has been tested (150, 169, 170). It was found that the rate of duplication was highest in areas of high density of repeat sequences, such as loci near two rRNA operons. In a seminal paper, an *E. coli* strain was created containing a lactase gene (*lac*) which encoded an enzyme with only 2% of the activity of wild type via a frameshift mutation. This activity was not sufficient to allow the strain to grow when lactose was the only carbon source. It could be demonstrated that this gene could revert back to the wildtype at a frequency of 10^-8, yet when 10^8 cells were plated on minimal media supplemented with lactose, 100 colonies were observed, 100 times greater than was expected. This was due to gene amplifications encompassing the *lac* gene and was explainable under the AID model (166, 171, 172).

The standing variation of the initial population meant there were multiple cells with duplications of the *lac* gene which could survive initially. As these populations grew there was selection for higher copy numbers of the *lac* gene giving a higher chance of a *lac* revertant gene forming. As such it could be noticed that some of the colonies growing on lactose plates, when grown in the absence of lactose would revert back but a smaller minority had gained stable lactase function. This smaller population had amplified the tandem *lac* array leading to a *lac* revertant and a collapse of the frame-shifted *lac* tandem array. These experimental systems are exactly that, however- experimental (125). Organisms with higher repeat content will experience higher rates of SV formation. This means that structural variations may be the primary response to environmental change as their rapid formation allows the population to quickly change due to the high standing variation (124).

1.8.1.7. Selection acting on inversions

Circular bacterial genomes are composed of two halves (known as replichores) which start at the origin of replication and end at the terminal region. The replichores are normally evenly sized and as such the terminus is 180 degrees from the origin of replication. An upset to replichore balance, such as a large CNV in one half of the chromosome, is associated with fitness costs. Espe *et al* found that an imbalance of 50 degrees in *E. coli* (168) was associated with fitness costs but other studies have found that this could also be as low as 16 degrees (169). This is because the unbalanced time to replicate each replichore can slow down the total time taken to complete genome replication and as such replichore imbalance correlates proportionally to slower growth rates. This may depend on how fast the cells are replicating, however (170).

When population growth is fast, cell division can outpace DNA replication and as such multiple DNA replications must occur in order to keep pace. As DNA replication starts at the origin of replication and proceeds to the terminus, multiple simultaneous replication forks mean that as replication proceeds round the genome, there will always be more copies of genes near the origin than near the terminus. This means that genes near the origin will have higher expression (as there are more copies being transcribed) than genes near the terminus, this is termed the gene dosage effect. Due to this, some genes are located near the origin and are less likely to be near the terminus. Most commonly these genes are the essential genes for growth, but may be limited to genes involved in transcription and translation (176, 177).

The gene dosage effect likely only affects fast growing bacteria, which excludes *B. pertussis* which has a generation time of 5-8 hours. It is thought that in slow growing species, there are fewer replication forks needed as DNA replication and cell division are well matched and thus the gene dosage effect is reduced (170). However, it is also known that in slow-growing organisms DNA replication can be slower (and therefore still outpaced by cell division) and that multiple forks can still exist (173).

A considerable challenge to cells is the concurrent replication and transcription of the genome. Fundamentally, these two processes are bound to collide as DNA replication and transcription are co-occurring at the same time in cells and occur at a rate of approximately 1000nt/sec and 80nt/sec, respectively (in *E. coli*) (174). Collisions can cause replication fork arrest or slowing and as such there is a selection pressure to reduce these events. Due to their different speeds, collisions of the two complexes can occur even when a gene is transcribed on the leading strand in the direction of replication (5‘ to 3’) but the head-on collisions if genes are transcribed in the opposite direction to replication slow down the replication fork more (175).

Because of these interactions between the replication and transcription process, it was initially thought that highly expressed genes would be preferentially located on the leading strand (176) but later it was shown that essential genes were preferentially located on this strand (172). It was thought this was because replication fork stalling due to head on collisions created a higher chance of mutations locally and mutations in essential genes are often lethal (172). Evidence from later studies with more data extended the categories of genes found preferentially on the leading strand, revealing that transcription factors were favoured on the lagging strand and that slower growing species tended to have a lower strand bias (177).

1.9. How single molecule sequencing has revolutionised the study of SVs in *B. pertussis*

The rapidly expanding knowledge of the *B. pertussis* genome is catalysed by technological innovations in sequencing. The latest advancement is single molecule sequencing, described as the third wave of sequencing (after Sanger sequencing and short-read sequencing) and has enabled the mass production of closed *B. pertussis* genome sequences. Single molecule sequencing, which is undertaken most frequently on the two competing platforms of Pacific Biosciences (PacBio) and Oxford Nanopore (Nanopore), generates long reads which can span nearly all the repeat regions found in a genome. This makes closing genome sequences a feasible task.

1.9.1. Pacbio SMRT DNA sequencing enables closures of *B. pertussis* genome sequences

Nearly all bacterial species sequenced using just short reads will produce an assembly which is fragmented. It is therefore not clear the order in which these fragments (known as contigs) occur. This is particularly true for *B. pertussis* as reads on the Illumina platform are a maximum size of 300 bp and thus cannot bridge the 1kb *IS481*. Assemblies of *B. pertussis* genome sequence data therefore contain at least as many contigs as there are repeat regions. In the reference Tohama I genome there are >270 repeats (23) and so an assembly of this data with short reads gives at least 270 contigs. When analysing a fragmented assembly, therefore, it is not clear if any genomic rearrangements or CNVs have occurred. In contrast, platforms which produce reads longer than 1-3kb (the range of most repeat sizes in *B. pertussis*) are able to produce closed assemblies in which the position of all genes is known (23).

To fully resolve CNVs the reads must be larger than the tandem array and have adequate coverage (normally at least 30x) which often is not possible on most platforms (and with most protocols) for large CNVs (over 5-10kb in length) (106, 132, 134, 183). Due to the transient nature of the tandemly duplicated loci they are nearly always identical and therefore during assembly they are often collapsed into a single copy of the locus and can remain invisible to short-read sequencing or long-read sequencing, and are hard to spot even with genome mapping (106, 134).

Molecular epidemiology at the CDC has focused on the production and analysis of closed genome sequences. These are assembled by a combination of enzyme mapping (a computer analysed restriction digest using an infrequent cutting endonuclease), PacBio sequencing and Illumina sequencing. This facilitates the generation of hundreds of closed genomes a year for *B. pertussis* and at the moment there are 470 uploaded to the Sequence Read Archive (SRA). Multiple technologies are used for these assemblies because long-read sequencing platforms, such as PacBio and Nanopore, offer long read lengths but higher error rate than Illumina sequencing which offers short but accurate reads. Usually, they are both used on the same sample so that each of the dataset's strength complements the other's weakness. Enzyme maps help in this regard in that they contain very long DNA fragments, often showing the DNA order in fragments as long as 750kb (184, 185). Enzyme maps do not contain the base composition of the fragments however and thus must be combined with sequencing platforms to produce closed genome sequences.

1.9.1.1. Resolving CNVs and complex SVs

When the resolution of a genome sequence using automated tools has been impossible with short-read, long-read and enzyme mapping data, this implies there is a complex or very long structural variation. This, at the moment of writing, requires manual resolution which involves finding areas of altered read depth coverage and cross examining this with the enzyme mapping before manually inserting the sequence of the suspected structural variant. This is then checked to be true by mapping long reads back to this manually altered assembly- if the junctions between the proposed structural variants have good coverage then it can be said that the reads support the hypothesised structure (106, 134).

1.9.1.2. The study of inversions in *B. pertussis*

It has been seen previously using PFGE that the *B. pertussis* genome was fluid in gene order (181) but the huge number of closed genomes generated by long-read platforms allows the study of genome inversions in *B. pertussis* on a large scale. The landmark study by Weigand et al could demonstrate that there were considerable gene order differences amongst a cohort of over 200 *B. pertussis* isolates (130). It was found that trees made from these gene orders approximately matched the phylogenetic tree made using SNPs meaning that there was not a random assortment of gene orders but that they were changing in a clock like fashion (similar to

SNPs). This meant that despite the large potential for rearrangement, there was limited and incremental change.

How these inversions were shaped by the forces of selection remains unclear. The inversions found by Weigand et al, however, did fit in with the fundamental selection pressures that face all bacterial genomes (130). It was found that the inversions were largely symmetrical around the origin or terminus, thus preserving replichore balance and any potential leading-strand bias. However, a number of asymmetrical inversions were also found and certain gene orders appeared more conserved than others (132, 176). These are both potential evidence of selection and this was expanded on in a second study in which Weigand et al found that there were patterns of inversions that were more common than others (131). The four most frequently observed gene orders were highly similar and when put onto a core-genome SNP-based tree it was observed that there were cyclical repeated inversion mutations that meant isolates were cycling through these 4 gene orders. This may have meant that these gene orders were conserved because of purifying selection, positively selected for under certain fluctuating environmental conditions or were fluctuating because they were under neutral selection and were drifting between conformations. The study was ultimately inconclusive in this regard.

Elegant studies have shown that genomes can undergo remarkable structural changes without significant fitness costs, as long as replichore and genome organisation rules are adhered to. A key study by Cui et al demonstrated that cells with circular genomes which had been linearised were just as fit as cells with the natively circular genome, if the origin of replication was central in the molecule (preserving replichore balance) (182). Another study by Itaya et al could show that randomly inserting 3.4Mb of DNA from *Synechocystis* into the *Bacillus subtilis* genome did not affect its fitness, as long as replichore balance was preserved (183).

It is interesting to note that although *E. coli* and *Salmonella* differ considerably in their nucleotide composition, their genomes have an almost identical gene order and structure. This is likely because of their large population sizes in which greater purifying selection effectively purges any deleterious (or marginally deleterious) mutations such as inversions. Following this, it is possible that the low effective population size of *B. pertussis* means that the species is more prone to genetic drift as there is not the population size needed to have strong purifying selection (175).

1.10. CNVs in B. pertussis

The study of CNVs in *B. pertussis* appears in the literature a number 14 times, although their study has been serendipitous and sporadic (106, 134, 189–192). One of the earliest and most comprehensive descriptions of CNVs in *B. pertussis* described a CNV of the adenylate cyclase/hemolysin gene (185). An isolate produced mixed colonies with either high or low haemolytic activity. The low haemolysis phenotype was stable whilst high haemolysis was unstable and could go onto produce both high and low haemolytic colonies. PFGE results showed that the higher haemolysis was due to a 350kb CNV. Phenotyping of this strain was undertaken by a variety of ways. The enzyme responsible for haemolysis had higher expression but a number of other virulence factors had the same expression between the two colony types. This indicated that the CNV was not disrupting expression genome-wide. Studies in a mouse model of respiratory infection revealed that high and low haemolysis clones did not show a significant difference in colonisation and revealed that the CNV was unstable in vivo, although the spectrum of mutants recovered after in vivo challenge was not reported. In addition, epithelial cell invasion models and macrophage survival assays did not show significant differences between colony types. This has been the only study to investigate the phenotype arising from CNVs in *B. pertussis*. The results showed that beyond the most fundamental level the CNV did not affect any broad in vitro or in vivo phenotypes (185).

Other studies of CNVs in *B. pertussis* have been limited. These studies relied on either CGH, PFGE or long-read sequencing (combined with enzyme mapping) technologies to describe the structure of the CNVs and were not able to phenotype them (106, 134, 189, 191, 192). No systematic analysis of CNVs in B. pertussis has been undertaken however. Systematic analysis of these mutations are rare in the study of bacterial species, although there are many case studies of them.

1.11. Structural variations as an overlooked class of mutation in bacteria

In humans a variety of CNVs are known to directly cause diseases or be associated with complex diseases. Complex diseases such as autism (188), schizophrenia (189) and Parkinson's (190) have been known to associate with structural variants. SVs have been implicated in the evolution of cancers with high grade ovarian cancer (191) and invasive breast cancer (197, 198) being two types that have a particularly close relationship with SV formation.

Due to the known importance of SVs in human populations they are routinely tested for, not only for diagnosis but also for epidemiological investigations. It is therefore equally important to describe the diversity of structural variants that can be observed but also to make specific genotype-phenotype links. The known significance of structural variants in humans means that micro arrays and analysis of whole-genome sequencing is routinely included in their study (188,194–196).

Whilst there are a multitude of case studies on structural variants in bacteria, the surveillance and systematic description of them for whole species is rare. This is in stark contrast to human genomics. I believe this means the prevalence and impact of CNVs in the bacterial kingdom is vastly under-appreciated and I hope to contribute to describing this dimension of bacterial genomics throughout the thesis.

1.12. Aim of study

Structural variations have been established as a ubiquitous type of mutation in all forms of life, including viruses. Their study, however, is predominantly limited to eukaryotic organisms (188–196). I therefore sought to investigate the prevalence of these mutations in *B. pertussis*, a species with a highly repetitive genome that has been known to be prone to structural variation (Aims 1 and 2).

Further to this, I wanted to explore the suitability of Genome Wide Association Studies to study CNVs and deletions in *B. pertussis*. In the absence of phenotypic data for *B. pertussis* I studied two potential problems of a future GWAS in the species: how to represent CNVs and the level of homoplasy (and therefore linkage) of deletions. GWAS is in high demand as CNVs are a mutation type which is heavily understudied in the bacterial kingdom and the impact of deletions on the species is unknown.

Aims

1. To investigate the prevalence of CNVs in *B. pertussis.*

2. To characterise the genome plasticity of *B. pertussis*

3. To define a reliable method to represent CNVs for a future Genome Wide Association Study

4. To investigate the level of homoplasy of deletions in *B. pertussis*

2. Methods

2.1. Sequence read mapping

Short-read data originating from the Illumina platform were retrieved from the National Centre for Biotechnology Information's (NCBI) Sequence Read Archive (SRA). One run was chosen at random for each BioSample, totalling 2709 runs including 94 locally provided runs. Reads were mapped to the *B. pertussis* B1917 genome sequence, which is broadly representative of the modern circulating strains (112) (RefSeq ID: NZ_CP009751.1), using BWA (197) implemented in Snippy (available: https://github.com/tseemann/snippy).

2.2. CNV prediction

CNVnator (198) was used to predict CNVs from read depth data generated from the mapping process. Statistical tests for significance within CNVnator discriminate high and low confidence calls. To further increase specificity, we implemented a very low P-value cut-off (p<0.0001). Abyzov *et al* empirically tested CNVnator to determine that ratios of the average read depth to the standard deviation of 4-5 produce the best balance between sensitivity and specificity (198). In accordance, samples exhibiting ratios < 3 were discarded as CNV calls were unreliable on such variable data (198). Window length was optimised for each genome, testing window sizes 500 -1000bp at intervals of 100bp to evaluate which gave a ratio closest to 4.5 as to minimize the effect of stochastic and/or artefactual fluctuations in read depth across the genome. Copy number estimates were rounded to the nearest 0.1. Code is available: https://github.com/Jonathan-Abrahams/Duplications.

2.3. Control data for CNVnator

As a negative control, short reads were simulated from the B1917 reference genome using ART to simulate the error profile of Illumina HiSeq paired-end 150 bp data (-ss HS25 -p -l 150 -f 20 -m 200 -s 10) (199). Simulated reads were mapped back to the reference genome using Snippy and CNVnator was used to call any spurious CNVs, as described above (198).

2.4. Heatmap

The read depth-based predictions were hierarchically clustered based on the similarities of CNV profiles (including deletions) of samples using the R package Hclust. This therefore meant that strains with similar complements of CNVs and deletions were clustered together on the heatmap which was plotted using the R package Plotly (200).

2.5. Networks

Overlapping gene content among CNVs was evaluated by constructing undirected network graphs which quantified the relationships (edges) between each CNV (nodes). An edge was constructed between nodes if both CNVs had a 75% overlap (non-reciprocal). Network analysis was undertaken in R using the Igraph package (201) and networks layout was generated by the Fruchterman algorithm (202).

2.6. qPCR

Bacteria were grown on charcoal agar for 3 days at 37^{o}C before inoculation into Stainer-Scholte (SS) broth (203) and grown overnight at 37 ^{o}C with shaking at 180 rpm. These cultures were used to inoculate fresh media at an OD_{600} = 0.2. Bacterial cells were harvested (1ml for DNA and 10ml for RNA extraction) at OD_{600} = 1.1$\underline{+}$0.1 by centrifugation (4000xg for 10 min) and resuspended in 700 µl of Tri-reagent (Invitrogen, ThermoFisher, Loughborough, UK), vortexed vigorously, and frozen at -80ºC. DNA was purified using QIAamp kit (Qiagen, Manchester, UK) in accordance with the manufacturer's instructions. The concentration of DNA was determined using Qubit broad range DNA quantification kit (Fisher Scientific).

qPCR was run on a StepOne Real-time PCR System (Applied Biosystems, ThermoFisher) using TaqMan™ Universal PCR Master Mix (Applied Biosystems), in a total reaction volume of 20 µl with 100pmol of DNA and with primer and probe concentrations as described in Table 2.2. Triplicate reactions were run for each sample. Reaction conditions were: 10 min at 95ºC followed by 40 cycles of 15 sec at 95ºC and 1 min at 60ºC. Copy number was quantified by using the $2^{-\Delta\Delta CT}$ method. Three biological repeats were used for determination of copy number in UK54.

To isolate RNA, nucleic acids were precipitated with ethanol, residual DNA was removed by incubation with 4U of Turbo DNase (Ambion, ThermoFisher) for 1 hour at 37 ºC, and RNA was purified using the RNeasy kit (Qiagen, Manchester, UK) in accordance with the manufacturer's instructions. The concentration of RNA was determined using Qubit broad range RNA quantification kit (Fisher Scientific). RNA integrity was determined by agarose gel electrophoresis. Finally, RNA was confirmed as being DNA-free by PCR using 50 ng of RNA as template in PCR with *recA*F and *recA*R primers. First strand cDNA was synthesised using ProtoScript II (NEB) with 1µg of total RNA as template and 6 µM random primers and incubated for 5 min at 25ºC, 1 h at 42ºC. The reaction was stopped by incubating at 65ºC for 20 min. cDNA was diluted 1/30 in H_2O for use in qPCR.

RT-qPCR was run on a StepOne Real-time PCR System using SyberGreen Turbo Master mix (Applied Biosystems), in a total reaction volume of 25 µl with primers at 300 nM. Triplicate reactions were run for each sample. Reactions conditions were: 95ºC for 10 min and 40 cycles of

95ºC for 15sec and 1 min at 60ºC. The housekeeping gene *recA* was used as a stably expressed control gene (Table 2.2). The ΔCT and ΔΔCT were calculated by determining the difference between the reference condition and experimental condition. Relative expression was represented as fold change (fold change =$2^{-\Delta\Delta CT}$). Significance was determined with one-way ANOVA. I undertook the initial qPCR to quantify the copy number of the locus in the original sample of UK54 and Iain MacArthur undertook all subsequent qPCR experiments.

Table 2.2. Table of primer and probe sequences and their optimal concentrations for two experiments: DNA and RNA quantification.

Name	Role	Sequence (5' to 3')	Optimal concentration (nM)	Role
CNV_fw	Forward primer	TCTGGGGAGTCGAAAGCAAT	300	DNA
CNV_rv	Reverse primer	TCTTGAGGGTGGCGAAGAAT	900	DNA
CNV_probe	Probe	FAM-ACGCCCCTTGCTGACGTCGC-BHQ	200	DNA
BP283_fw	Forward primer	CAGGCACAGCACTATTGCG	500	DNA
BP283_RV	Reverse primer	GACGATTACCAGCGAGATTACGA	300	DNA

BP283_probe	Probe	FAM-CCGCCATCGCAACCGTCGCATTCA-BHQ	200	DNA
RecA_fw	Forward primer	AACCAGATCCGCATGAAGAT	300	RNA
RecA_rv	Reverse primer	ACCTTGTTCTTGACCACCTT	300	RNA

2.7. Phylogenetics

To investigate the phylogenetic relationship between strains containing CNVs, a core genome SNP alignment was created using Snippy (available: https://github.com/tseemann/snippy). Phylogenetic trees were constructed using RAxML-ng (204,205). RAxML was used with the GTRgamma model, 10 starting trees and 500 bootstraps. Itol (206) was used to display the tree. HomoplasyFinder was used to find ancestral states (207).

3. Chapter 3: A systematic investigation in *Bordetella pertussis* reveals 273 CNVs

3.1. Introduction

At the time of writing there are over 1.2 million bacterial whole genome sequencing runs archived on the European Nucleotide Archive (ENA), the majority of which have not been analysed for large deletions, inversions or tandem CNVs, collectively known as structural variants. Despite this, there is a rich and diverse literature describing structural variants and their phenotypes (124) including antimicrobial resistance (208,209) and increased virulence (210,211) – topics of major public health concern.

As many structural variants are formed through homologous recombination, bacterial species with highly repetitive genomes are likely to experience increased burden of SVs (124). In this study we focus primarily on *Bordetella pertussis*, the main causative agent of whooping cough , genomes of which have approximately 200 copies of IS*481* (15,23). Speciation of *B. pertussis* from a *B. bronchiseptica*-like ancestor was synonymous with the accumulation of insertion sequence (IS) elements and subsequent large-scale rearrangements and deletions, likely though homologous recombination between repeats. As a result, genomes of *B. pertussis* encode at least 1000 fewer genes than *B. bronchiseptica* (23).

Homologous recombination between repetitive IS elements still plays a major role in the genetics of *B. pertussis* and as such it is described as having a plastic genome (23,120,130). *B. pertussis* genomes experience deletions, inversions and amplifications of large tracts of DNA, although these distinct forms of SV have been described to widely varying levels. Deletions mediated by homologous recombination continue to streamline the genome in extant *B. pertussis* lineages and have enjoyed systematic description (120,212). Similarly, inversions have been subject to recent study (213), catalysed by advances in long read sequencing. This is in contrast to amplifications, which have been found 11 times previously, primarily using techniques that predate whole genome sequencing (178,184–187,214,215). There exists now a wealth of whole genome sequencing data suitable for studying amplifications, yet there has been no systematic investigation of their contribution to genomic diversity within the *B. pertussis* population.

The highly repetitive genome of *B. pertussis* is known to be capable of a high frequency of structural variants, such as deletions and rearrangements (130,212,216). This contrasts with the third type of structural variant, CNVs, of which only twelve in *B. pertussis* have been serendipitously discovered (105,132,184–187). I therefore hypothesised that the *B. pertussis* population contained vastly more than twelve CNVs. In this chapter I aimed to define an accurate method to describe, categorise and compare CNVs in *B. pertussis*.

3.2. Methods

3.2.1. Manual assembly methodology

28 *B. pertussis* isolates, each containing 1 CNV were assembled (Table 2.1). This involved short read data from the Illumina platform, long read data from the PacBio platform and either Opgen or Nabsys genome maps. The genomes were assembled with PacBio reads using Hierarchical Genome Assembly Process version 3 (Pacific Biosciences) (HGAP) software.

The Opgen and Nabsys enzyme mapping technologies are evolutions of pulsed-field gel electrophoreses (PFGE) as both platforms rely on the analysis of physical patterns of DNA. These patterns are generated from either tagging (Nabsys) or enzyme cutting (Opgen) at regularly spaced sites on the genome. Both platforms create maps of individual DNA molecules which are then assembled to produce a consensus map (but not base sequence) of the input DNA. The advantage of these platforms is that the DNA molecules are long, often over 500kb, and as such provide long-range information which is outside of the normal range of long-read DNA sequencing platforms.

A hybrid assembly process is used on these isolates and produced closed genomes for the vast majority of strains. This process normally results in a single closed genome, but if a genome was not closed, it was checked against genome maps. Genome maps were combined with increased read depth indicating a copy number variant (CNV). CNVs were resolved by manually altering the assembly to match the data from the optical maps. This DNA structure could be verified by mapping the PacBio reads back to the assembly-gapless coverage meant the reads supported this configuration. Illumina reads were then used to polish this assembly (217–219).

If genomes were resolved using Nabsys genome maps, the protocol was as follows. Genomic DNA isolation from *B. pertussis* strains was performed at the CDC according to a Nabsys solution-based protocol modified from the bacterial DNA protocol for AXG 20 columns and Nucleobond Buffer Set III (Macherey-Nagel, Bethlehem, PA). Purified DNA was sent to Nabsys for nicking, tagging, coating and data collection on an HD-Mapping instrument. Nicking enzyme Nb.BssSI (NEB) was used for strain D236 and the nicking enzyme combination Nt.BspQI/Nb.BbvCI (NEB) was used for strains D800, H624, J085, J196, and J321. Resulting *de novo* assembled HD maps, raw data, and data remapped to PacBio *de novo* assemblies were provided by Nabsys for further analysis and sequence assembly comparisons at the CDC using NPS analysis (v1.2.2424) and CompareAssemblyToReference (v1.10.0.1).

If genomes were resolved using Opgen genome maps, the protocols was at follows (and is reproduced from (218)). Optical maps for each isolate were prepared from cells of single 1-mm colony equivalents following growth on Regan-Lowe agar without cephalexin using the Argus system (OpGen, Gaithersburg, MD) according to special company protocols. Briefly, high-molecular-mass bacterial DNA (205-kbp average size) was isolated with minimal shearing and applied to a chemically modified glass surface with fabricated microfluidic channels. The stretched DNA on the channels was digested *in situ* with *Kpn*I in a partial digestion mode and stained with a JoeJoe fluorescent dye on an automatic MapCard processor. To confirm the unusual insertions and CNVs that were revealed, restriction enzyme *Bam*HI was used. The digested DNA molecules were imaged using an Argus fluorescence microscope and Path-Finder automated image-acquisition and tiling optical map assembly software (OpGen). The resulting single-molecule restriction maps were assembled into consensus whole-genome maps with Gentig software (OpGen) that recurrently aligned overlapping DNA molecules with similar fragments to calculate a concluding map. Final whole-genome maps in this study are composites from at least 32 single fragmented molecules at every point and typically represent an average depth of 50 to 300 molecules. Restriction map alignments between different strains were generated using MapSolver software (v.2.1.1; OpGen, Gaithersburg, MD).

All manually resolved genomes were generated by Michael Weigand and collaborators at the CDC (132,218,219).

Table 2.1. Accession numbers for the genome sequence data of the 28 isolates for which genome sequences were manually resolved genomes.

Alias	BioSample	Isolation Location
J448	SAMN05770316	India
D236	SAMN08200080	USA: UT
J737	SAMN11822393	USA: CO
J196	SAMN10161199	USA: CO
J767	SAMN11822404	USA: CO
J085	SAMN07352199	USA: CO
J085	SAMN07352199	USA: CO
J029	SAMN07352195	USA: CO
J385	SAMN11822230	USA: CO
J083	SAMN07352198	USA: CO
A639	SAMN11821629	USA: OH
D800	SAMN11821631	USA: PA
D800	SAMN11821631	USA: PA
J742	SAMN11822397	USA: CO
J741	SAMN11822396	USA: CO
J739	SAMN11822394	USA: CO
J740	SAMN11822395	USA: CO
D665	SAMN08200081	USA: NV
H624	SAMN08200082	USA: OR
J412	SAMN11822239	USA: VT
J447	SAMN05770315	India

J299	SAMN07352224	USA: CO
J299	SAMN07352224	USA: CO
J139	SAMN08200079	USA:TX
J733	SAMN11822392	USA: CT
J349	SAMN11822226	USA: OR
J318	SAMN10161200	USA: MN

3.3. Results

3.3.1. CNVs in a small set of genomes

The US Centers for Disease Control and Prevention (CDC) conducts routine molecular epidemiology of pertussis. This involves using data from PacBio and Illumina sequencing platforms and enzyme mapping from the Nabsys and Opgen platforms (105,130,213,220) to produce closed genomes (see chapter methods). Including retrospectively sequenced samples and prospectively sequenced samples, 725 isolates were sequenced during the years 2014-2018 . Of these, 45 assembled isolates could not be resolved by this analysis and contained evidence of CNVs (mis mapping reads, increased coverage etc). This indication, in addition to the availability of high-quality long-range data from long reads and long DNA fragments in enzyme maps, meant these isolates were good candidates for further analysis.

Assemblies which may have had CNVs were analysed using the manual assembly method which resulted in closed genomes being obtained for 28 isolates (Figure 3.2), including two used for the production of vaccines against pertussis. Each of these strains had one resolved CNV. Conflicting genome orders from multiple data sources and/or inadequate coverage of the CNV junction when reads were mapped to the hypothesised genome order led to the remaining 16 genomes not being closed using the manual assembly method. Further automation utilising enzyme maps and long reads may mean these isolates can be fully assembled in the near future.. Analysing these CNVs in conjunction reveals there was several recognisable characteristic: A conserved and clustered distribution of CNVs primarily at 3 loci; the variable length of CNVs which were often over 50kb (Figure 3.2) and all these CNVs were flanked by repeat sequences over 1kb in length, primarily IS481.

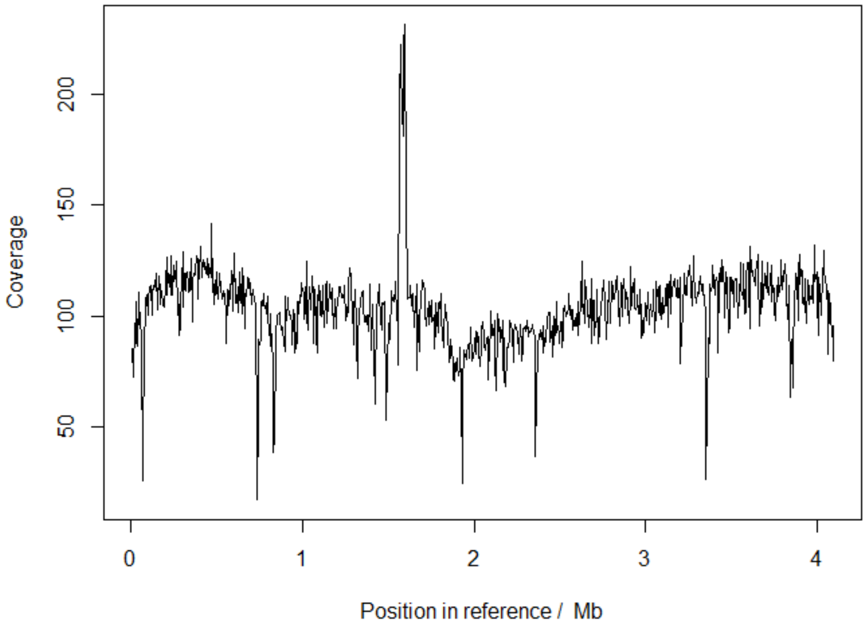


Figure 3.1: Reads were mapped against the B1917 reference genome (X axis) and how many reads covered each base was plotted (Y axis). This revealed a region with twice as much read coverage as other parts of the genome-an example of a large spike in read coverage. Such spikes were accompanied by a failure of PacBio assemblies to close the genome sequence and further investigation into these genomes identified that each had a CNV at the location of the spike in coverage.

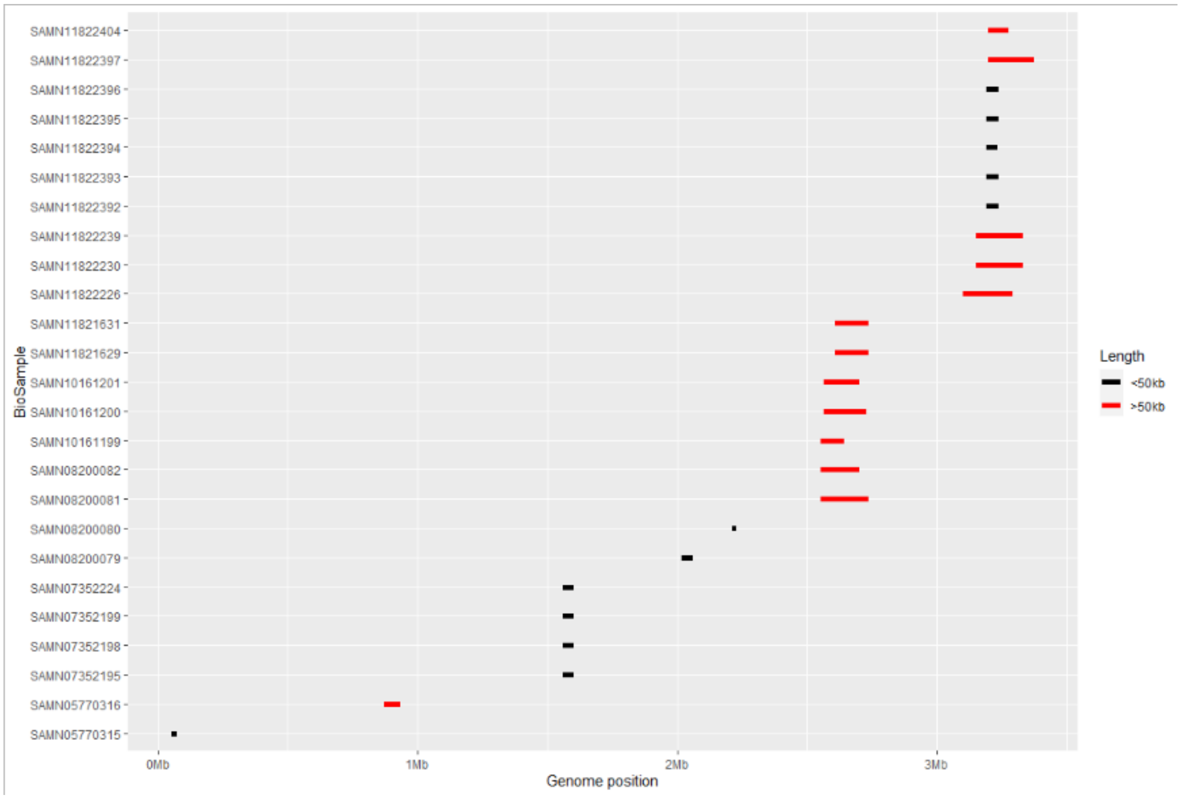


Figure 3.2: CNVs (represented by horizontal lines proportional to the size of the resolved CNV) were resolved in 24 isolates (Y axis) in a variety of genomic loci (in relation to B1917, X axis). It can be seen that CNVs appeared frequently at 3 loci and had overlapping but not identical start/end locations. There was also diversity with the size of CNVs, with many CNVs over 50kb long.

These results (Figure 3.2) were foundational to the thesis. It was plausible that there were many isolates containing undiscovered CNVs and that there was evidence of this in their sequencing data. However, most isolates do not have long read data or enzyme maps available for them, but just short read data. Using the manually resolved dataset as a benchmark, we sought to develop a prediction and screening tool to identify CNVs in *B. pertussis* within the 1000's of isolates in the Sequence Read Archive (SRA), which is formed mostly of short read data.

3.3.2. Establishing a methodology

3.3.2.1. Read depth: pros and cons of different read depth tools

As short-read sequencing platforms are popular and a lot of data exist in this format, I sought to define a method to predict CNVs from this type of data. The read length (<300 bp) of this platform means that it is inadequate to resolve these CNVs (which are commonly ~100kb) or span the junction between tandem arrays which are bound by repeat elements >1kb in size. As the copy number of a locus is beyond the resolution of short read data, it is necessary to use other sources of data as proxies. There are a number of these that can be used, but the most fundamental and informative is the read depth (Figure 3.3). The logic is simple: increased copy number of a locus (in comparison to the reference genome) will mean more reads map to this region in the reference genome than would be expected if the locus was at single copy (Figure 3.4).

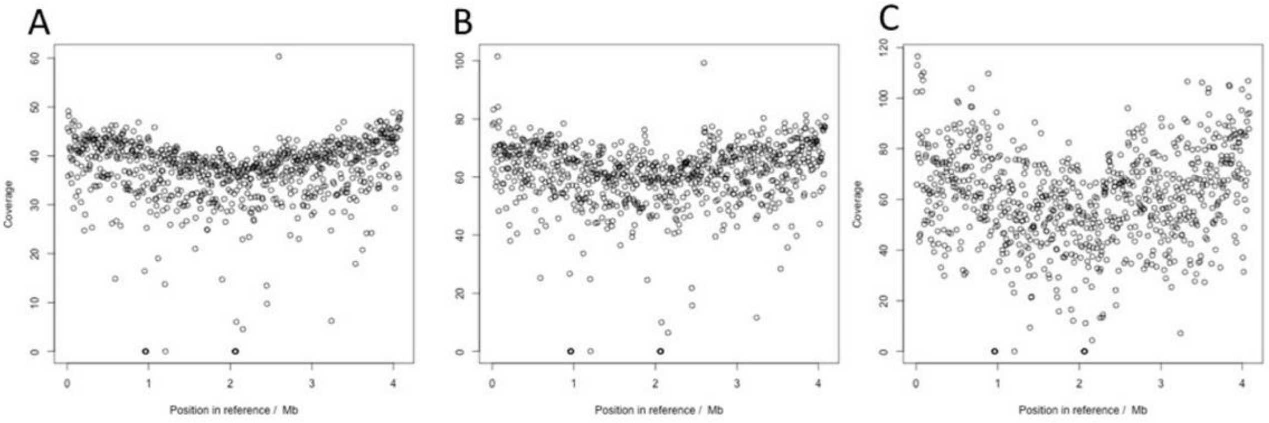


Figure 3.3: Reads from 3 isolates were mapped to the reference genome (X axis) (see ‘Sequence read mapping’ section in methods) and the average read depth (Y axis) plotted in 5kb windows. Genomes A, B and C show low, medium and high read depth noise respectively, as can be seen by the spread of read coverage in each graph. This proves problematic for downstream analysis as high read depth noise leads to false positive CNVs being predicted.

Other signals from short read data include split read signals (a read spanning an SV junction) and read-pair data (measuring the distance between read pairs). These two sources of data can provide base pair resolution to the junction sequences and thus infer the structure of the tandem array. Both methods, however, rely on the CNV junction to not be in a repeat region that is larger than the read size. This is not applicable to the CNVs studied here which have been formed by homologous recombination between large repeats. This data is absent in the CNVs described here.

3.3.2.2. How window length effects CNV predictions

Predicting CNVs in data sets of bacterial genome sequences, which frequently consist of sequences of hundreds to thousands of small genomes, rather than fewer but larger sequences for eukaryotes, presents a challenge- one that is not often mentioned in the literature. Comparing read depth coverage data across thousands of samples which vary by a multitude of factors such as sequencing chemistry, sequencing instruments and read lengths is complex and causes fluctuations in coverage. It is not known, however, exactly how these factors contribute to the inter-sample differences in read coverage.

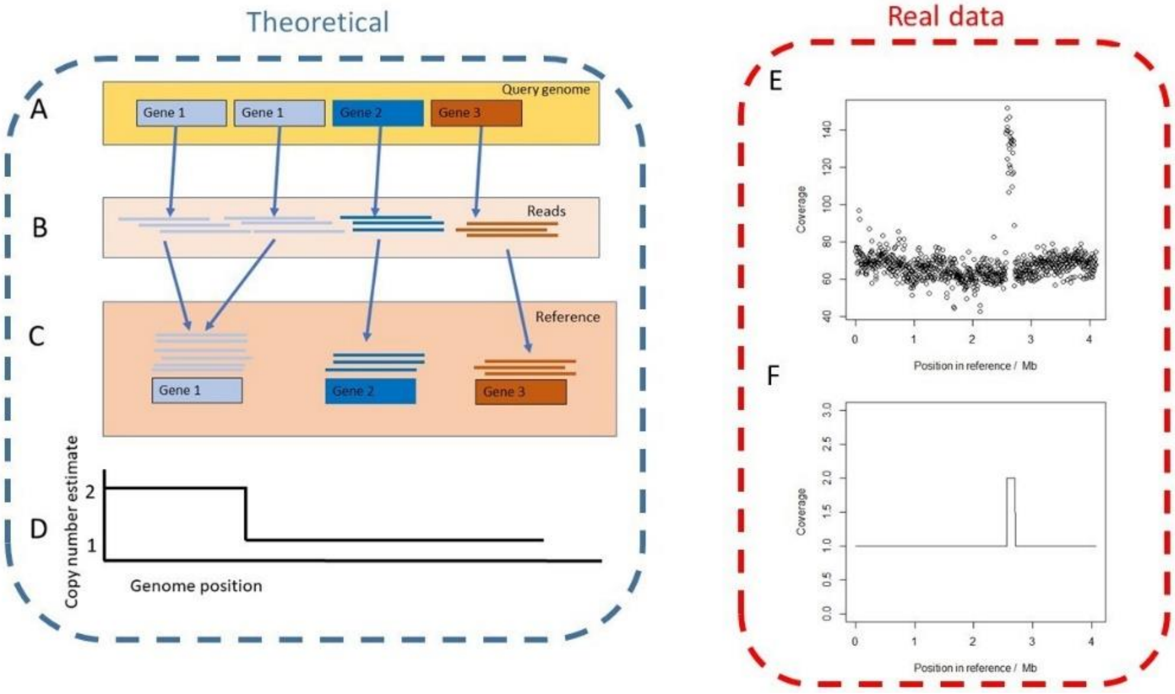


Figure 3.4: Schematic overview of prediction of CNVs from sequencing read depth. In the theoretical example (purple box, left), the query strain contains a perfect tandem CNV of gene 1 whilst gene 2 and 3 are at single copy (A). Short reads from the query strain are generated (B) and mapped to the reference genome, that contains all genes at single copy (C). Reads from both copies of gene 1 in the query strain map to this locus in the reference sequence and thus twice as many reads map to this gene compared to genes 2 and 3. This data must be processed to obtain estimates of copy number and to avoid technical bias (D). Using an example with real data (red box, right) the strain SAMN08200079 was analysed using the CNVnator method (see Methods). Read coverage was graphed to reveal a CNV at 1.4Mb (E, analogous to theoretical graph C). An example of a statistical analysis of these data was then graphed (F, analogous to theoretical graph D).

As read depth is the proxy used here for predicting CNVs, artefactual fluctuations can appear as false positive CNVs in the analysis. It is therefore necessary to normalise these fluctuations

(198). All read depth-based prediction tools will analyse the read depth in windows to analyse how read depth changes across the genome. However, the size of this window influences the results. Larger windows are less sensitive to fluctuations in read depth but are also very specific (low false positives) whilst the inverse is true for smaller windows. Therefore, larger windows can be used on genomes with high read depth fluctuations and smaller windows for less noisy genomes. CNVnator provides extensive supplementary data and methods pertaining to this and notes a heuristic: the optimum ratio of average read depth to the standard deviation is generally between 4-5. Choosing window size, therefore, is a balancing act of these factors (198).

CNVnator (198) is one of the most highly cited read-depth based CNV prediction tools and as such was an attractive choice to use in this analysis. In addition, it was easy to use in a parallel way as it is available containerised on the Docker platform. A recent benchmarking study (221) noted CNVnator was highly sensitive but lacked specificity but this study was run with CNVnator on its default setting- not having optimized window length.

3.3.2.3. Why use B1917 for mapping?

The choice of reference is important when using a mapping pipeline. Although the pangenome of *B. pertussis* is small, any gene that is not present in the reference will not be analysed in the mapped data. Additionally, and much more relevant for *B. pertussis* (130), when data is mapped to a reference the true gene order of the sample is masked. Therefore, strains with CNVs in rearranged loci may appear as discontinuous stretches of duplicated DNA in the reference-obscuring the true genomic structure of the CNV and distorting the number of CNVs predicted for strains.

To minimise the side-effects of read mapping, therefore, an isolate that was broadly representative of the global population of *B. pertussis* in terms of gene content and gene order was needed. Additionally, the strain must be widely used so that any results can be tested and replicated by the scientific community. It is known that the strain B1917 is generally representative of the gene content of recent circulating isolates and has been recently established as a modern reference genome (112), thus leading to many labs having a stock of it. It is therefore a viable alternative to the traditionally used reference strain Tohama I (23), which was isolated in the 1950's.

It was not clear, however, if B1917 had a representative gene order and therefore this was investigated. A whole genome alignment was conducted with 3 of the genomes with manually resolved CNVs (isolated during outbreaks in the US), B1917, Tohama I and 2 additional closed genomes isolated from other countries using Mauve (222). As the vast majority of the closed genomes contained on the SRA were isolated in outbreaks in the US, it is therefore important to include diverse isolates from other countries. More genomes could not be used as the progressiveMauve (222) algorithm, aligns genomes in a pairwise fashion which means that computational time scales exponentially with extra genomes. This experiment resulted in a high quality phylogenetic tree (Figure 3.5) which showed that B1917 was separated by less gene order changes from all isolates in the dataset compared to Tohama. B1917 is therefore a more appropriate reference genome than Tohama.

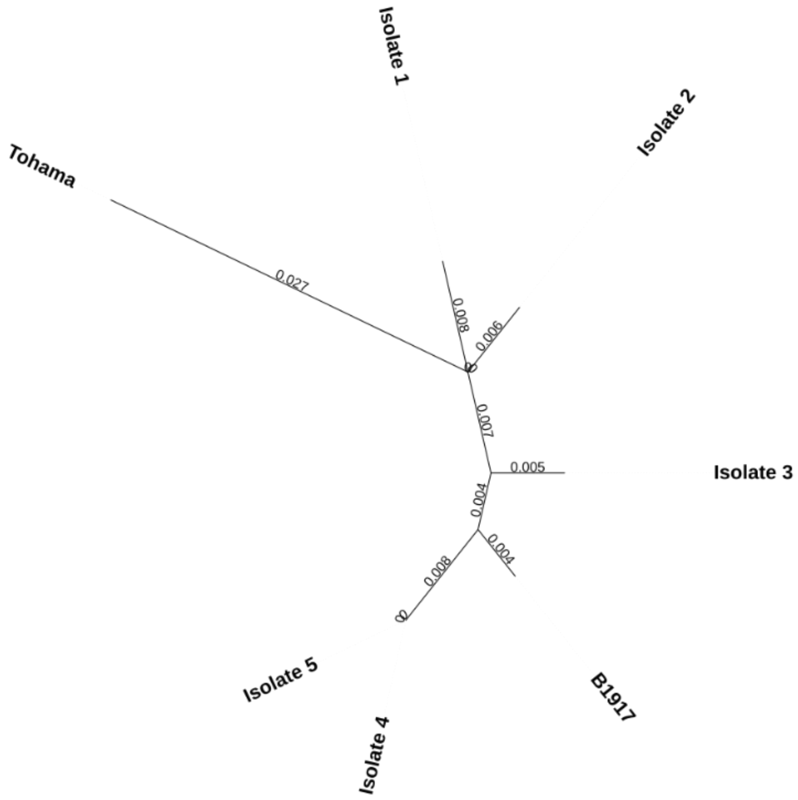


Figure 3.5: A tree showing that Tohama is separated by roughly 3x as many unique gene order changes (branch length: 0.027) from the rest of the phylogenetic tree as compared to B1917 (branch length: 0.004). This demonstrates that B1917 has a gene order that is closer to modern isolates than Tohama.

3.3.3. Establishing accuracy

3.3.3.1. Training dataset- B1917

The first test of CNVnator on *B. pertussis* data was to simulate short read data for the B1917 reference genome and map it against itself. When this data is analysed with CNVnator, all genes should be identified as being present and at single copy. As expected, analysis with CNVnator returned no false negatives or false positives in this experiment as all gene were correctly identified as being at single copy. The approach was subsequently evaluated by analysing the set of 28 manually resolved genomes.

3.3.3.2. Manually resolved genomes

Read depth is only a proxy for copy number, so it is best suited to predict simple CNVs rather than complex CNVs. Here I define simple CNVs as a single stretch of DNA that is contiguous in both the reference genome and the genome from which the data was generated. A complex CNV is defined as a stretch of DNA that is contiguous in the reference but non-contiguous in the genome from which the data was generated. A complex CNV may be formed by tandem arrays of different genes in close proximity or by a tandem array of a rearranged segment of DNA, for example.

When establishing the accuracy of the pipeline I considered the 25 simple CNVs and two of the three complex CNVs separately. Only one (J321) of the 25 data sets containing simple CNVs failed the quality control checks (see Methods) for high read depth noise and was excluded. This left 24 high quality strains with simple CNVs. Whilst the 24 simple CNVs mainly occurred at three distinct loci (Figure 3.2), their beginning and ending coordinates, as well as overall length, varied between strains. Thus, three measures of accuracy were tested: the correct prediction of the 24 CNVs, the quantity of false positives and the discrepancy between the predicted start and end coordinates of the predicted CNVs and the true coordinates from the manually resolved genome sequences (breakpoint accuracy). Here a breakpoint is defined as the start and end of the CNV locus and these two points form the junction between the two tandem copies.

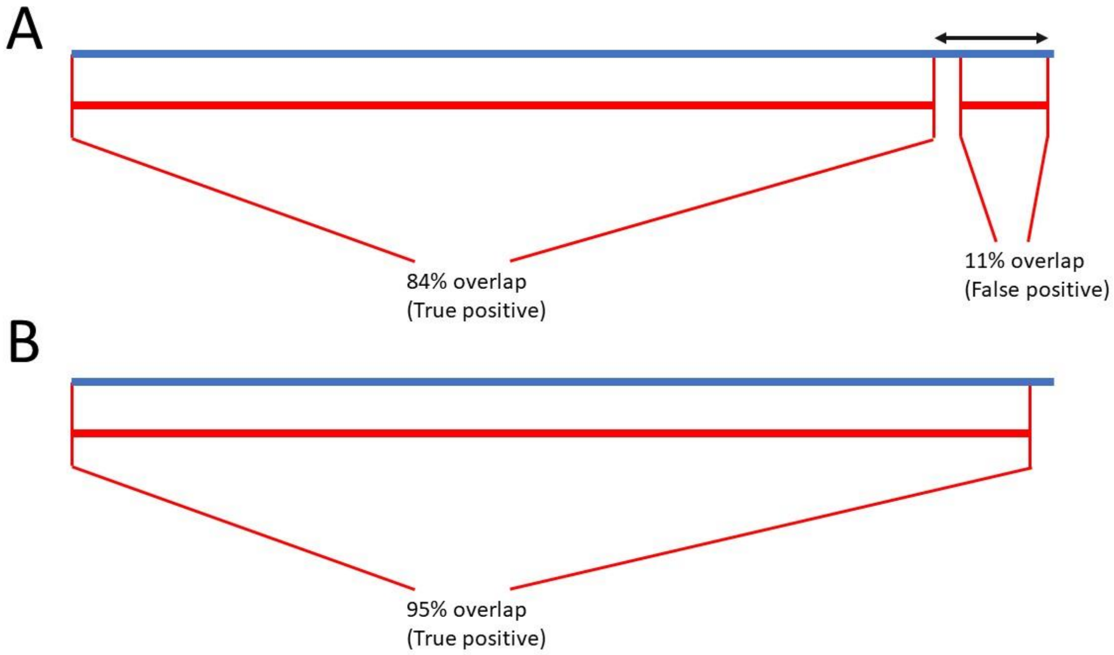


Figure 3.6. A schematic detailing how predicted CNVs(red) were compared to true CNVs(blue), in order to evaluate the accuracy of the pipeline. (A) The strain J085 had two CNVs predicted (red) despite only 1 confirmed in the manual assembly (blue). One prediction had an 84% reciprocal overlap with the true CNV and was considered a true positive whilst the other predicted CNV had only an 11% overlap and was considered a false positive. This also caused the predicted end of the true positive fragment to be further from the true end, impacting the breakpoint accuracy, denoted by a black arrow. (B) A stereotypical CNV was predicted which had a 95% overlap with the true CNV in strain D236.

Of the 24 resolved, high quality and suitable CNVs, 23 were correctly predicted (defined as >=80% reciprocal overlap between the predicted and true CNV regions) (Table 3.1). One CNV was counted as a false negative as no prediction was made for this isolate. Three false positives were detected in three different strains. Two of these were due to one gene within the CNV locus being predicted as single copy, causing the true, single CNV to be predicted as two, separated by the falsely predicted single copy gene (described in Figure 3.6). In the third false positive, a second locus was predicted as a CNV. Interestingly, mapping PacBio reads generated from this strain to the reference also showed increased coverage at the same locus, however enzyme mapping showed no evidence of a second duplicated locus in this isolate. To be conservative, this was therefore counted as a false positive, despite the mixed evidence that this locus contained additional copies.

The breakpoint accuracy of estimates was calculated (Figure 3.7). As a true positive was counted if it had >80% reciprocal overlap with the true CNV, this meant that a true CNV could have both false positive and true positive predictions associated with it (Figure 3.6). To prevent false positive fragments skewing further analysis, false positives predictions were excluded. The median distance between the true end point coordinates and the CNVnator-derived estimates was 0.5 genes. There were five estimated start/end points which were considerably (>=5 genes) less accurate than the rest of the dataset, mainly arising from the two strains in which the CNV was predicted as two separate loci- a false positive and a false negative. Thus, the pipeline correctly predicted, and with good breakpoint accuracy, the CNVs for 20 of the 27 resolved genomes (74%), with 2 further correctly CNVs predicted (11%) but as two adjacent but separate loci.

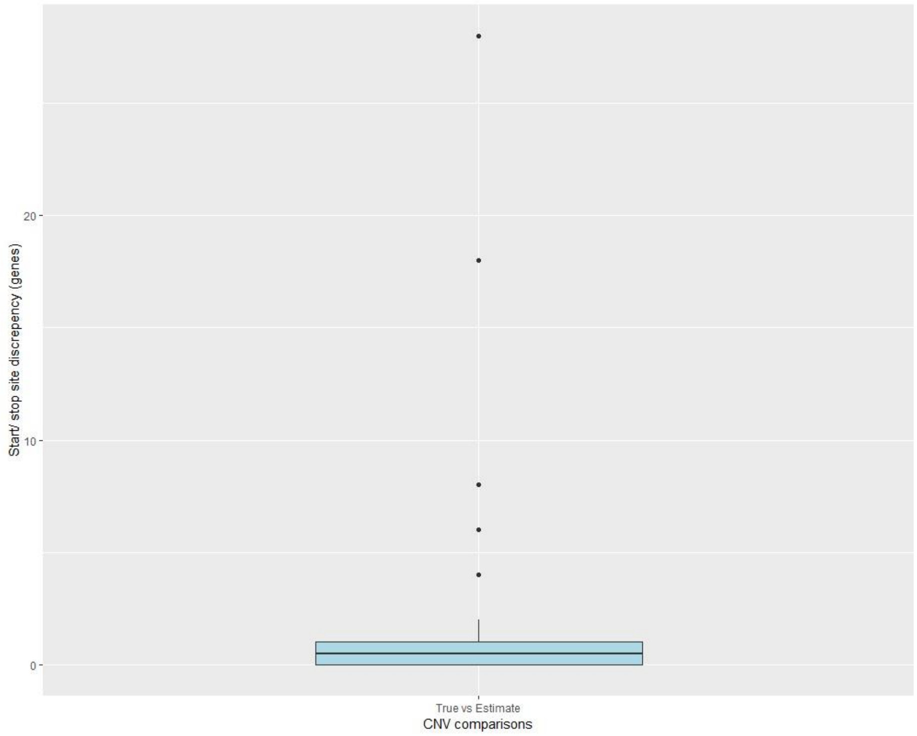


Figure 3.7: For each CNV, the start and end coordinates of the predictions were compared to the true coordinates and the distribution of these discrepancies was plotted(Y axis). This showed a tight distribution around the median distance of 0.5 genes discrepancy.

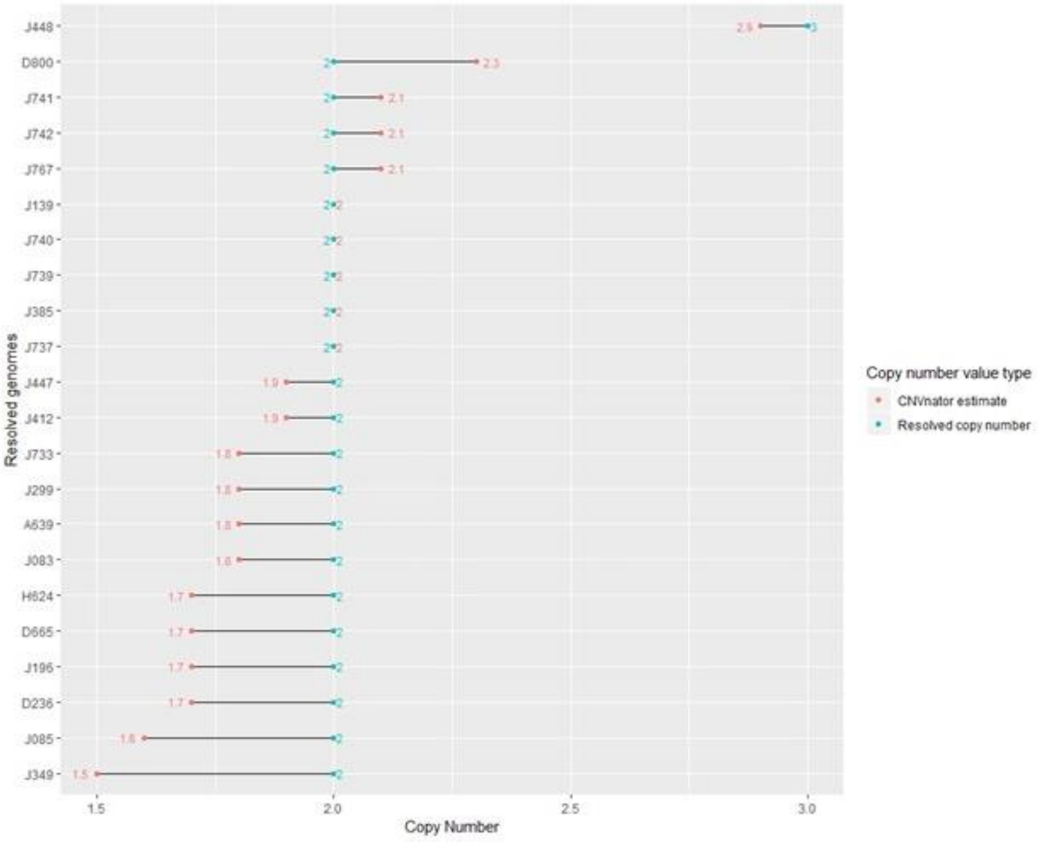


Figure 3.8: The true (orange) and estimated (blue) copy number (X axis) of CNVs was plotted for the manually resolved cohort (Y axis). Large discrepancies (black bars) between the estimate and the true CNV copy number state can be seen for the majority of isolates.

Table 3.1. The 27 predicted CNVs compared to the true CNVs.

Alias	Estimate start (B1917 gene index)	Estimate end (B1917 gene index)	Estimated Length	Copy number estimate	True start (B1917 gene index)	True start/Estimated start discrepancy	True end (B1917 gene index)	True end/Estimated end discrepancy	True copy number	Copy number discrepancy	Reciprocal overlap
J448	2331	2440	109	2.9	2331	0	2440	0	3	+/-0.2	>=0.8
D236	2798	2947	149	1.7	2799	1	2947	0	2	Lower	>=0.8
J737	2798	2947	149	2	2799	1	2947	0	2	+/-0.2	>=0.8
J196	2840	2999	159	1.7	2840	0	3000	1	2	Lower	>=0.8
J767	2840	2900	60	2.1	2840	0	2900	0	2	+/-0.2	>=0.8
J085	2752	2769	17	1.5	2752	0	2915	146	2	Lower	FALSE
J085	2770	2907	137	1.6	2752	-18	2915	8	2	Lower	>=0.8
J029	2830	2907	77	1.6	2830	0	2871	-36	2	Lower	FALSE
J385	2831	2871	40	2	2830	-1	2871	0	2	+/-0.2	>=0.8
J083	2830	2866	36	1.8	2830	0	2867	1	2	Lower	>=0.8
A639	2830	2870	40	1.8	2830	0	2870	0	2	Lower	>=0.8
D800	778	834	56	2.3	779	1	834	0	2	Higher	>=0.8

D800	2098	2362	264	1.3	779	NA	834	NA	2	Lower	FALSE
J742	1403	1444	41	2.1	1403	0	1446	2	2	+/-0.2	>=0.8
J741	1403	1445	42	2.1	1403	0	1446	1	2	+/-0.2	>=0.8
J739	1403	1445	42	2	1403	0	1446	1	2	+/-0.2	>=0.8
J740	1403	1444	41	2	1403	0	1446	2	2	+/-0.2	>=0.8
D665	1790	1827	37	1.7	1791	1	1828	1	2	Lower	>=0.8
H624	1965	1978	13	1.7	1966	1	1978	0	2	Lower	>=0.8
J412	50	67	17	1.9	50	0	68	1	2	+/-0.2	>=0.8
J447	2331	2439	108	1.9	2331	0	2440	1	2	+/-0.2	>=0.8
J299	2269	2303	34	1.7	2276	7	2440	137	2	Lower	FALSE
J299	2304	2440	136	1.8	2276	-28	2440	0	2	Lower	>=0.8
J139	2289	2403	114	2	2289	0	2405	2	2	+/-0.2	>=0.8
J733	2277	2357	80	1.8	2276	-1	2363	6	2	Lower	>=0.8
J349	2276	2401	125	1.5	2276	0	2405	4	2	Lower	>=0.8
J318	2291	2376	85	1.4	2288	-3	2432	56	2	Lower	FALSE

3.3.3.3. Predicting complex CNVs

In the dataset of 28 manually resolved genomes, three genomes were excluded from the previous analyses for having complex CNVs. As Weigand et al had demonstrated that the *B. pertussis* population has undergone diverse rearrangements (130), it was unclear how this would affect the prediction of CNVs. This is because multiple SVs occuring at the same locus is a challenging signal for read-depth based approaches to correctly predict CNVs. Rearrangements have no impact on read depth and adjacent CNVs or CNV/deletion combinations leave an amalgamated read depth signal, because it is not clear that there is a separation between the SVs. For example, if there was a two-copy region next to a 3-copy region, it is possible that CNVnator would not be able to identify that this was two distinct SVs and would predict a long CNV of copy number 2.5.

It is therefore important to investigate how these events appear in practice in the analyses. Two samples with resolved, complex CNVs were examined. In strain B199 (105) there is a triplication followed by a CNV of a larger region (which includes the original triplication region) leading to some parts of the locus to be at 2 and some at 3 copies (Figure 3.9). In strain F701, the CNV locus had a different gene order than B1917 (Figure 3.10), having likely been affected by an ancestral inversion mutation.

For both isolates the predicted CNVs did not satisfy the strict >=80% reciprocal overlap rule for any of the true CNVs and thus could not be classed as successfully predicted. Merging the true CNVs into a single CNV also did not satisfy the strict 80% reciprocal overlap rule. These results indicate that CNVnator struggles to accurately estimate the boundaries of such CNVs. Whilst gene order changes are common in *B. pertussis*, it is unlikely that a rearrangement will affect a CNV given that only 2 out of the 28 manually resolved genomes presented here were affected by rearrangements (7%).

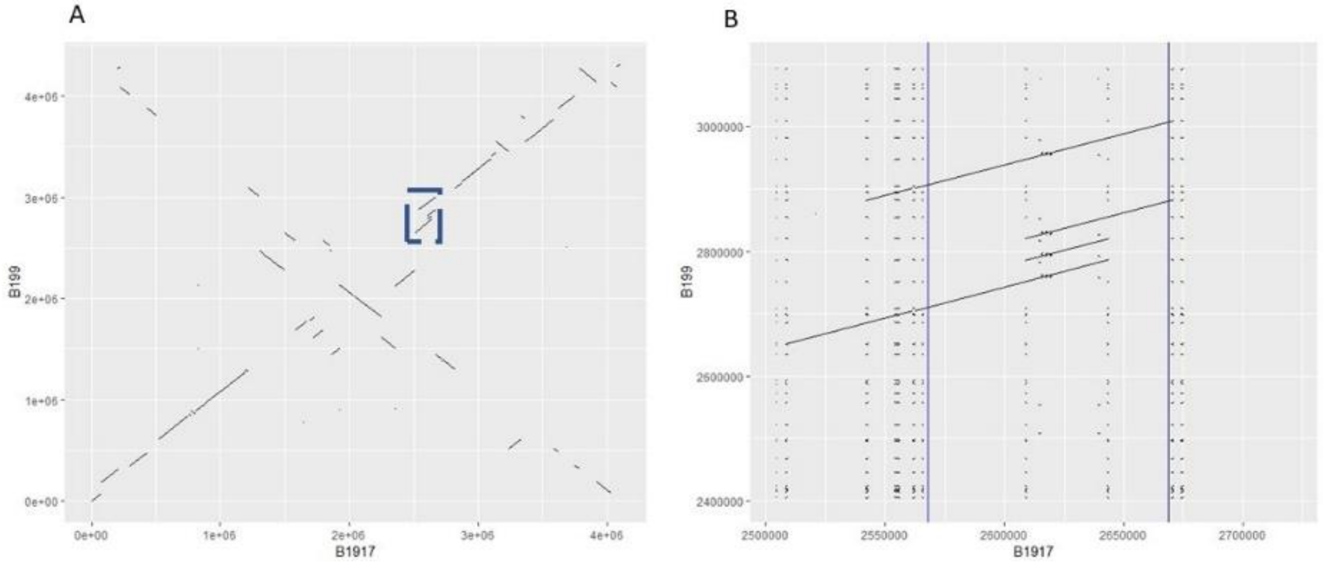


Figure 3.9: A: The genome of B1917 (X axis) compared to strain B199 (Y axis) . The section that has a resolved CNV is highlighted in the blue box and expanded in (B). The DNA corresponding to 2.54Mb to 2.67Mb in B1917 has a complex arrangement of CNVs. The structure is characterized by an initial triplication and then followed by a CNV of a larger region (which includes the original triplication region). This isolate therefore has some loci at 4 copies, and 2 copies. The longest region estimated as a CNV by CNVnator is highlighted using blue bars (for start and end positions).

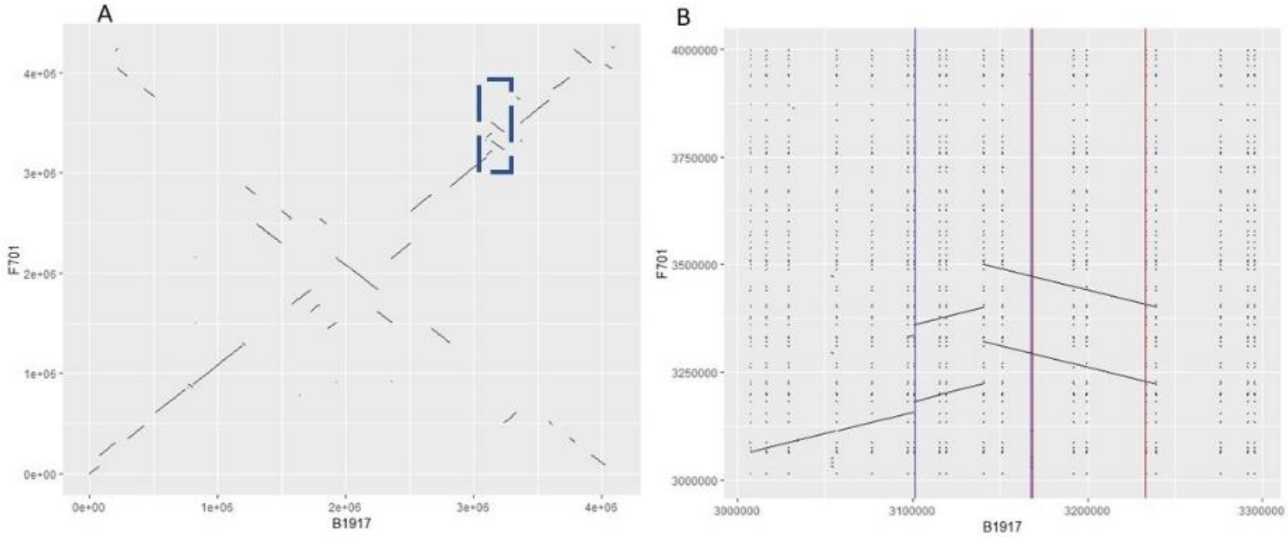


Figure 3.10: A: The genome of B1917 (X axis) compared to strain F107 (Y axis). The section that has a resolved CNV is highlighted in the blue box and expanded in (B). The DNA corresponding to 3.1Mb to 3.24Mb in B1917 is duplicated in F107 but partway through it has been disrupted by an inversion. The two relatively evenly sized estimates by CNVnator are highlighted using two colours (blue and red) at both the start and end points. Where the two estimates meet in the middle, the line appears purple.

3.3.4. Predicting CNVs in a cohort of 2709 isolates

3.3.4.1. Cohort statistics

The pipeline was applied to predict CNVs in 2803 *B. pertussis* isolates for which short-read sequence data was available in the Sequence Read Archive (SRA). Of the 2709 total *B. pertussis* samples, 94 exhibited 30x average coverage and 185 had high read coverage noise. Therefore, the final dataset included 2430 *B. pertussis* isolates. Due to its size, this dataset was dubbed the 'large cohort dataset' and is referred to as such throughout the thesis.

Of the 2430 studied isolates, 1711 had all genes predicted at single copy and therefore did not make any further appearances in the analysis apart from in the phylogenetic tree. This confirmed that B1917 had a highly representative gene complement as these 1711 isolates contained all of the genes present in B1917 at single copy. This left 719 strains with at least one deletion or CNV. Of the 719 strains with CNVs: 191 isolates contained 272 CNVs as some isolates had more than 1 CNV predicted. In summary, therefore, 7.8% of the studied *B. pertussis* strains contained at least one CNV.

It was found that 43 isolates had more than 1 CNV, containing 143 CNVs in total. The median number of CNVs for this group was 2. This data was used to estimate how many predictions may have been false positives due to CNVs being split into two, as had occurred in the manually resolved dataset (Figure 3.6) in 10% of the cohort. I calculated the distance between CNVs in this group and found that 23 isolates had CNVs within 3 genes of each other. In total this comprised 27 CNVs that may have been false positives due to CNV splitting- which was 10% of the predictions. The oversensitivity of the method therefore likely did not significantly impact the findings.

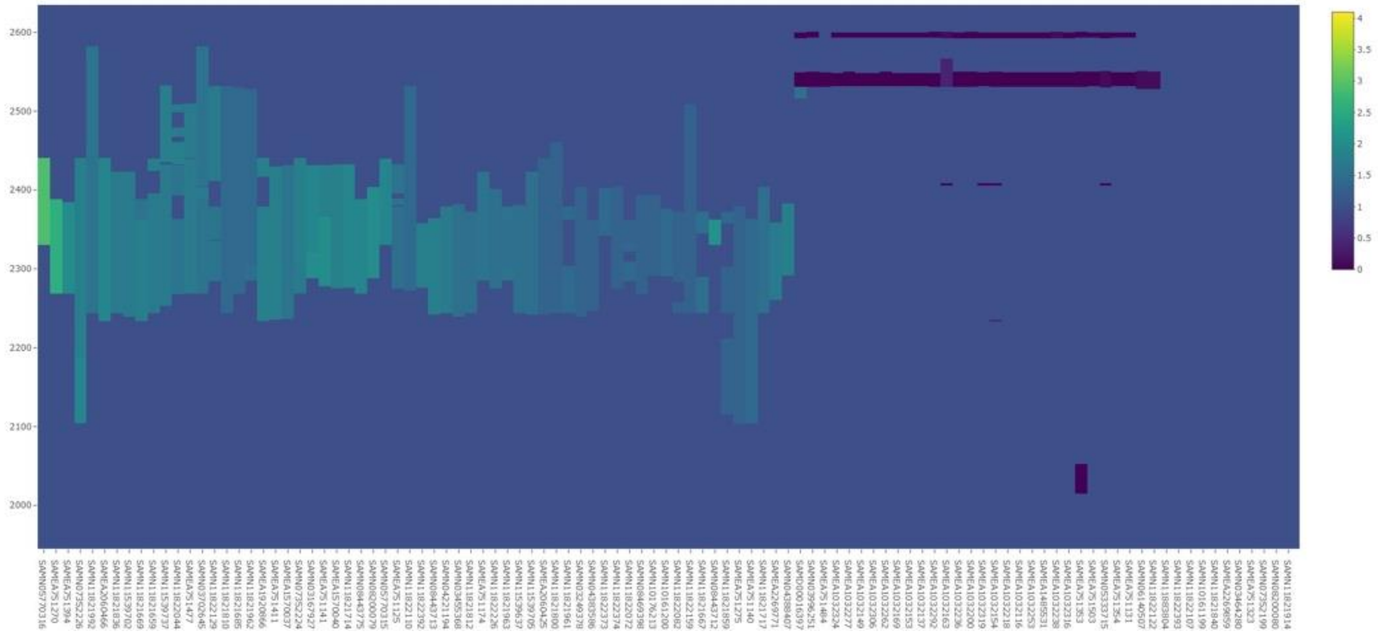


Figure 3.10: Reads were mapped to the reference genome for 2430 isolates, of which approximately 100 are shown here (X axis). The copy number prediction for each gene is displayed (Z axis) for every gene in the genome (Y axis), of which 600 genes are shown here. Sections of higher copy number (yellow colour) are visible. This section of the genome corresponds to the locus of Network 1. A legend of the colour scale is on the far right.

A visual analysis of this data revealed that the key results observed in the manually resolved dataset are confirmed in the global cohort of 2709 *B. pertussis* strains, namely: CNVs clustered at specific loci, varying gene content within these clusters and many long CNVs (Table 3.2 and Figure 3.11). It was clear from this analysis that the vast majority of CNVs were found at a small number of locations, which I therefore termed ‘hotspot’ loci.

3.3.5. Leveraging network graphs to analyse hotspots

Hotspot formation has been previously described in bacteria (209,223,224). Hjort et al showed that when *Salmonella enterica* clones were exposed to colistin, a region was amplified and this conferred resistance. It was determined that a single gene involved in regulating lipidA biosynthesis conferred colistin resistance when amplified and positive selection for this drove formation of the hotspot, despite the amplification being >50kb in length. Similarly, Domenach et al showed that multiple isolates of *Mycobacterium tuberculosis* (*M. tuberculosis*) had CNVs which were under selection *in vitro* but not *in vivo* but did not determine which genes were under selection in this condition. Both studies found these amplifications arranged in hotspots which appeared to be similar to those described here (large size and overlapping but not identical gene contents). Given the similarities between the hotspots identified here and those in the literature, I hypothesised that positive selection was driving hotspot formation in *B. pertussis*. Furthermore, If hotspots were driven by positive selection, it was likely that the genes common to CNVs in the hotspots were driving this (209). I therefore developed a quantitative framework in order to define hotspots, analyse their properties and find which genes were in the core of the network.

3.3.5.1. Using networks graphs to analyse hotspots

In microbiology, trees are frequently used to compare differences (normally SNPs) in a DNA sequence that is shared between multiple individual strains (the core genome), yet trees are a generic concept and form part of the mathematical field of graph theory. A tree is a type of graph in which all elements are related to one another other, in this example, by sharing the same core genome DNA. Graphs are constructed using nodes and edges, which represent data points and relationships, respectively. A phylogenetic tree aims to quantify the relationships between the datapoints (isolates) by creating the simplest graph that explains the observed DNA patterns.

In order to create the network graph, the relationship between all CNVs was quantified as the proportion of gene content overlap between all pairwise comparisons. Network graphs were constructed between CNVs ('nodes') that overlapped with each overlap coded as a line ('edges'). The 272 identified CNVs formed 24 network graphs, representing 24 distinct genomic loci. Only 11 network graphs, corresponding to the hotspot loci, included three or more isolates and contained 254/272 (93%) of the predicted CNVs (Table 3.2).

3.3.5.2. Using networks to leverage new data

Once CNVs were coded within a network structure, the core genes could be investigated. It was hypothesised that most networks would centre on a small group of genes, as can be visually seen in the heatmap (Figure 3.12). To define the network core, I determined the genes contained in at least 90% of the amplified genes in each network. A number of network cores contained genes with varied predicted functions (Table 3.2). For example, Network 1 contained the genes involved in flagella assembly and function (Figure 3.11) (225); Network 2 contained the *nuo* operon which codes for NADH dehydrogenase, a key component of the electron transport chain (Figure 3.13) (227, 228) and Network 3 contained the fim3 gene involved in the pathogenesis of *B. pertussis* and present in some acellular vaccine formulations (Figure 3.13) (228).

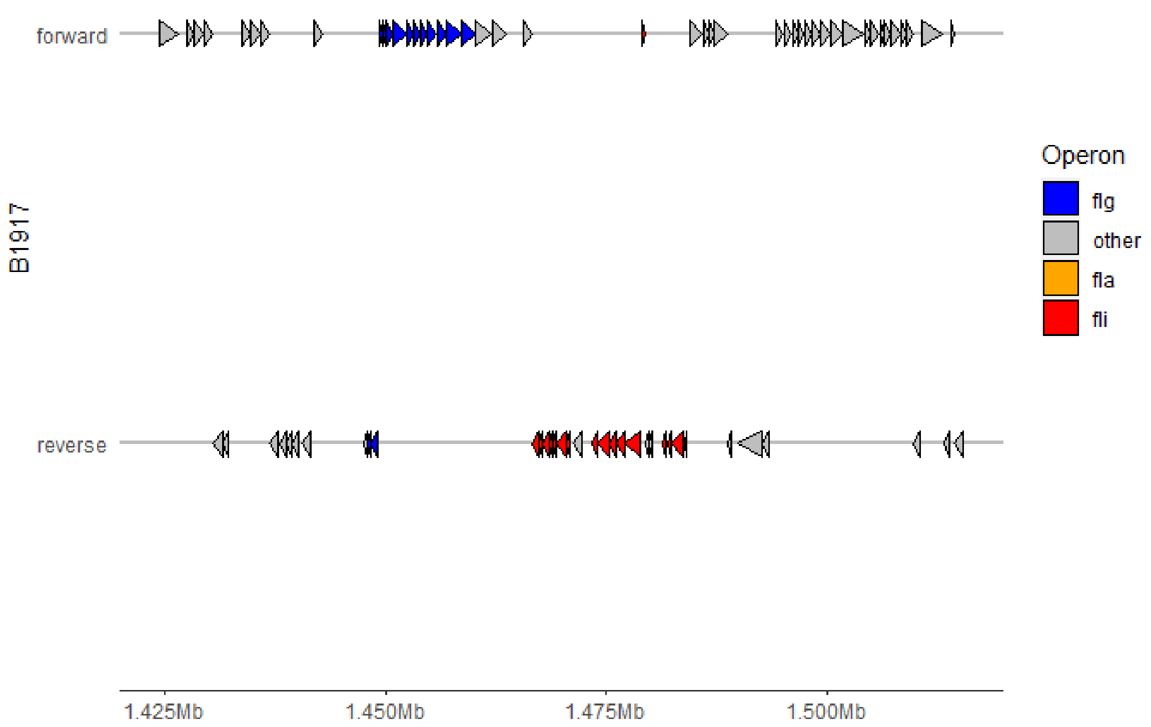


Figure 3.11. A schematic of the genes (on the forward and reverse strand) contained in the core of Network 1 in B1917(X axis). The genes involved in flagella assembly and function are highlighted.

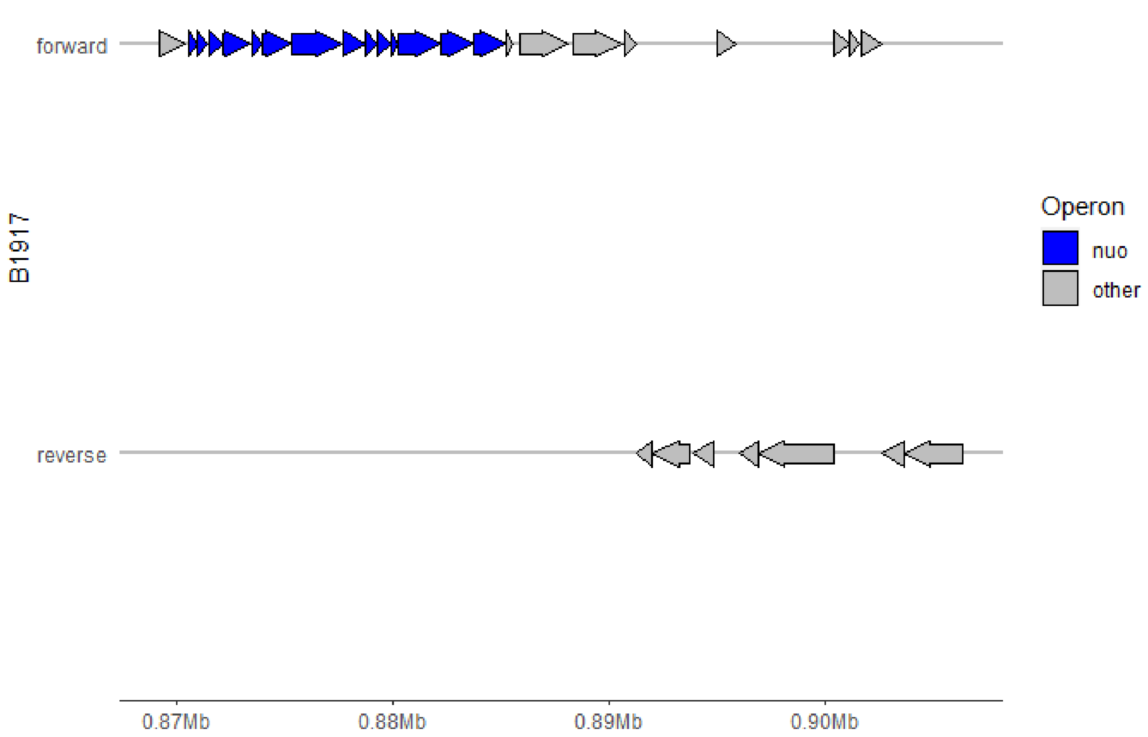


Figure 3.12. A schematic of the genes (on the forward and reverse strand) contained in the core of Network2 in B1917 (X axis). The *nuo* operon, involved in respiration, are highlighted.

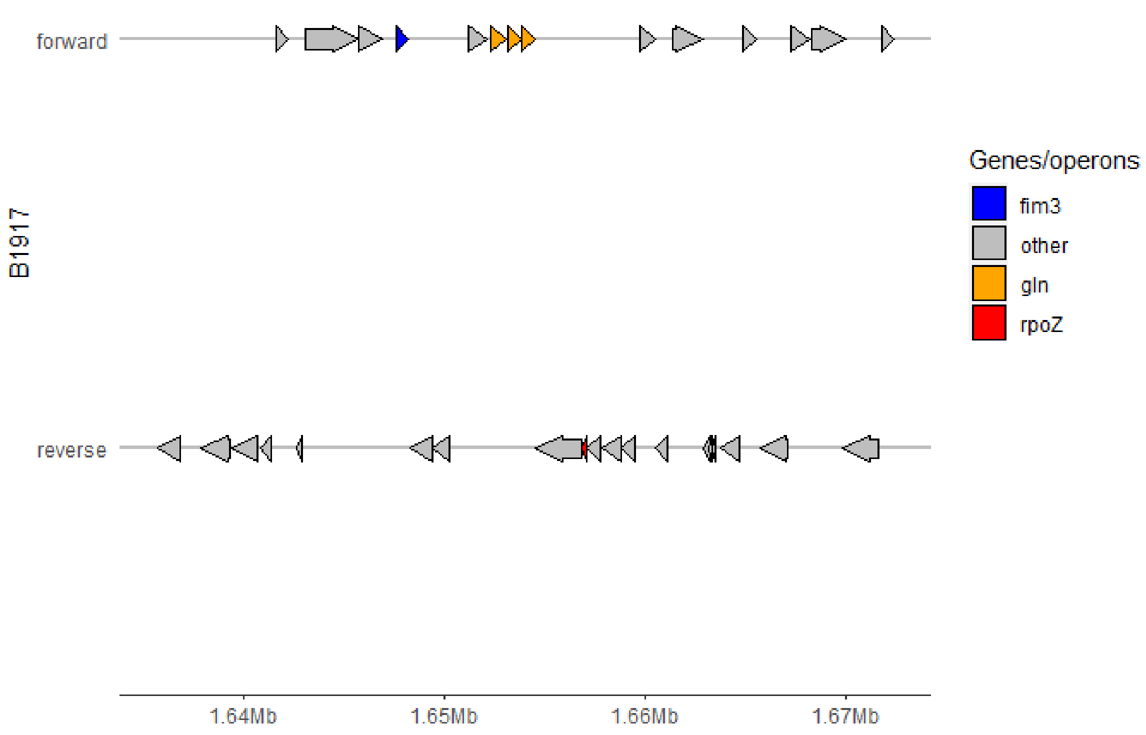


Figure 3.13. A schematic of the genes (on the forward and reverse strand) contained in the core of Network 3 in B1917 (X axis). A number of genes of interest are highlighted.

Table 3.2 Core genes of the three most frequent hotspots

Index	Network 1 core genes	Network 2 core genes	Network 3 core genes
1	BP1350	BP0840	BP1558
2	BP1352	nuoA	BP1560
3	BP1353	nuoB	BP1561
4	BP1354	nuoC	BP1562
5	BP1355	nuoD	BP1563
6	leuD	nuoE	BP1565
7	BP1358	nuoF	mutS
8	BP1359	nuoG	BP1567
9	BP1360	nuoH	fim3
10	BP1361	nuoI	BP1569
11	BP1362	nuoJ	dapA
12	BP1363	nuoK	BP1572
13	BP1364	nuoL	glnH
14	BP1365	nuoM	glnP
15	flhB	nuoN	glnQ
16	BP1370	BP0855	spoT
17	flgM	bfrD	rpoZ
18	flgA	bfrE	gmk
19	flgB	BP0858	BP1579
20	flgC	fabG	BP1580
21	flgD	BP0860	BP1581
22	flgE	BP0861	BP1582
23	flgF	BP0862	amn
24	flgG	serB	BP1584
25	flgH	mfd	BP1585
26	flgI	ispD	BP1586
27	flgJ	ispF	BP1587

28	flgK	BP0867	rph
29	flgL	fbp	BP1589
30	tsr	pepN	BP1590
31	tar		BP1591
32	BP1388		BP1592
33	fliR		BP1593
34	fliQ		
35	fliP		
36	fliO		
37	fliN		
38	fliM		
39	fliL		
40	BP1397		
41	fliJ		
42	fliI		
43	fliH		
44	fliG		
45	fliF		
46	fliE		
47	BP1405		
48	BP1406		
49	fliT		
50	fliS		
51	fliD		
52	flaG		
53	folC		
54	BP1413		
55	cvpA		
56	purF		
57	dsbB		
58	glnD		

59	map		
60	rpsB		
61	tsf		
62	pyrH		
63	frr		
64	uppS		
65	cdsA		
66	dxr		
67	BP1426		
68	BP1427		
69	BP1428		
70	lpxD		
71	fabZ		
72	lpxA		
73	lpxB		
74	rnhB		
75	BP1434		
76	BP1435		
77	ppsA		
78	BP1437		
79	BP1438		
80	BP1439		

In addition to the *Nuo* operon, a number of genes found in the core hotspots appeared to be essential in-vivo and in-vitro in broth culture and the murine model of colonisation, respectively (229). Amplified essential genes included RNA polymerase coding genes (*RpoB*, *RpoC* & *RpoZ*) and lipid biosynthesis (*LpxA*& *LpxB*). Essential genes are often preferentially near the origin of replication and experience increased copy number as multiple replication forks start at the origin, leading to those genes having increased copy number and expression. It was possible that amplification was a second path to increasing gene dosage of essential genes.

Statistical analysis (Fishers exact test) of the association between essential genes and amplifications found that 44% of amplifications (123/272) were enriched for in vitro essential genes and 3% (8/272) for in vivo essential genes. This may have indicated that essential genes were under positive selection for increased copy number. Analysing hotspot cores resulted in 5 of the 6 not enriched for essential genes of either type, with only the core of hotspot 2 being significantly enriched for in-vitro essential genes. On the premise that hotspot cores are the genes which are driving the positive selection of the amplification, the association between amplifications and essential genes appears to be largely incidental, although network 2 may be driven by gene essentiality. Essential genes in the core of network 2 included the *nuo* operon (detailed above) in addition to genes involved in terpenoid synthesis (*ispD* & *ispF*) and gluconeogenesis (*fbp*). The relationship between essential genes and amplifications appears nuanced and complex and may be just one factor that influences amplifications in *B. pertussis*.

Network name	Frequency (CNVs)	Mean length (genes)	Median start (B1917 gene name)	Median end (B1917 gene name)	Mean copy number	Core (>=90%) proportion (%)	Network density (%)
1	102	106	RS12140	RS12755	1.6	67	55
2	57	82	RS15100	RS15490	1.7	37	63
3	21	80	RS07175	RS07660	1.68	44	60
4	18	20	RS00010	RS00130	1.35	10	100
5	13	67	RS19230	RS19625	1.93	40	50
6	11	75	RS05505	RS05935	1.6	33	73
7	8	49	RS04185	RS04430	1.88		71
8	8	74	RS09665	RS10290	1.82		43
9	7	13	RS19965	RS10580	2.49		100
10	6	23	RS19465	RS19565	1.32		100
11	3	45	RS01035	RS01300	1.63		67

Table 3.3: Table of network statistics. Columns correspond to the network name, the frequency of CNVs in each network, the mean length of the CNVs, the median start and end genes of the network, mean copy number, how big the core network is in relation to the mean length and the network density. The core network was defined as the genes contained in >90% of the CNVs in the network. Network density is defined as how interconnected the network is and therefore how overlapping the CNVs are.

3.3.5.3. qPCR verification of a CNV

In order to validate predictions made via CNVnator, a CNV was chosen for further analysis. The initial verification method was undertaken by qPCR but later verification was achieved by Nanopore sequencing, which forms a substantial amount of work described in Chapter 4. It was therefore necessary to choose a CNV with a tractable size that could fit, in its full tandem configuration, into an ultra-long Nanopore read. The genome of UK54 (SAMEA1920853) was predicted to have a 16 kb long CNV at a copy number of 4; short enough to observe the CNV locus in a single sequence read on the Nanopore platform (as its tandem length was predicted to be up to 64kb), assuming that each copy occurred in tandem as observed in both our data and previous reports. This was the highest copy number CNV in the dataset. The CNV was part of Network 9 (Table 3.3 and Figure 3.13) which was comprised of 7 other CNVs, one of which was also predicted at a copy number >2 (3.3, Strain SAMN11822098).

The copy number of this locus in UK54 was validated using qPCR to ensure the copy number prediction was correct. A probe and primer set were designed to quantify the DNA copy number of a DNA segment inside (in gene B1917_RS10525) of the CNV and outside of the CNV, using a2^ΔΔCT analysis. This method is most often used to quantify the changes in the expression levels of genes, but is equally suited to comparing copy number between two samples.The relative copy number of the B1917_RS10525 gene within the CNV compared to a single- copy gene encoded outside the CNV locus was 4.38 +/- 0.4 which matched the read depth-based prediction, supporting the ability of CNVnator to predict CNVs in *B. pertussis*. Further analysis of the CNV in UK54 is presented in Chapter 4. This method successfully verified one CNV, but required primer/rpboes to be designed for each CNV. I therefore sought to verify CNVs in an alternative way.

3.3.6. Predicted CNVs are highly associated with repeats

Verification was possible by investigating the association of predicted CNVs with repetitive elements in comparison to all genes. All previously resolved CNVs were adjacent to repetitive sequences in line with homologous recombination between large (>1kb) repeats being the driving mechanism of SV formation in *B. pertussis*. Suggesting this was a clear marker for true CNVs. Analyses were restricted to isolates for which closed genome sequence information was available (excluding 28 isolates with manually resolved genome sequences were analysed above), as it is possible to locate accurately IS elements only in closed genome sequences. This left 16 CNVs in 13 isolates remaining for analysis.

The predicted boundaries of these 16 CNVs were significantly (p<6^-08) closer to repeat genes (median distance of +/- 1 gene) than non-CNV genes (median distance of +/- 5 genes) (Figure 3.15). This, in conjunction with our stringent quality control steps and the previously accurate predictions (of which the median distance to true CNV starts and ends was 0.5 genes- marginally higher), supports the accuracy of the prediction of 272 CNVs.

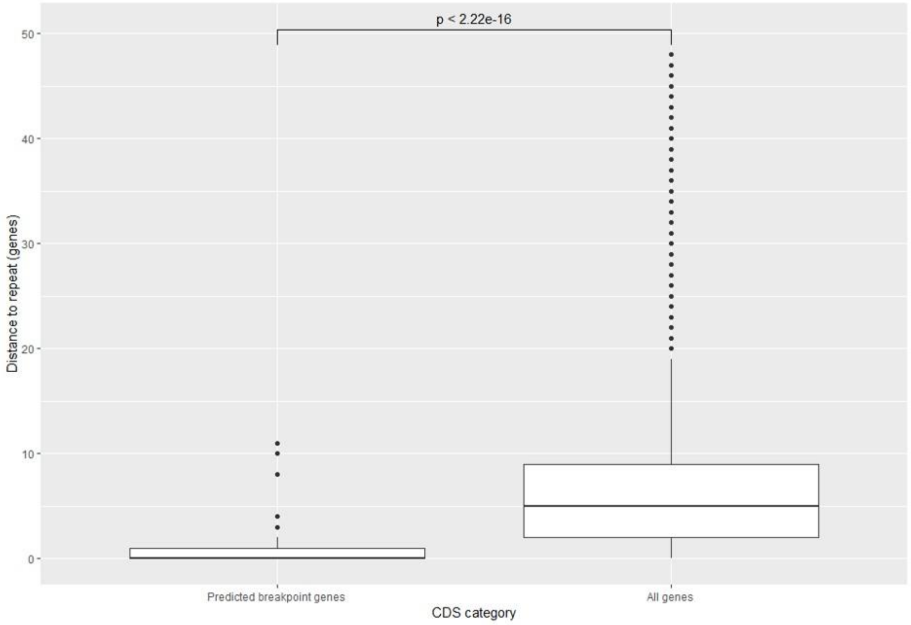


Figure 3.15: The distance (measured in genes) between CNVs and repeat genes (Y axis) was identified in closed genomes. The genes at breakpoints were compared to all genes in the genome (X axis). The ends of CNV loci were found to be significantly closer (median: 0 genes) to repeats than the average gene (median:5 genes).

3.3.7. CNVs occur as homoplasies throughout the phylogenetic tree

It was demonstrated that while CNVs did overlap at hotspot loci, often they had varying gene contents-strongly indicating each arose from an independent mutation. It was possible, however, that there could have been a single CNV event and subsequent remodelling gave rise to the hotspot like effect. It has been previously shown that hotspots arise by mutations in independent lineages and therefore I hypothesised that this was the case in the dataset presented here.

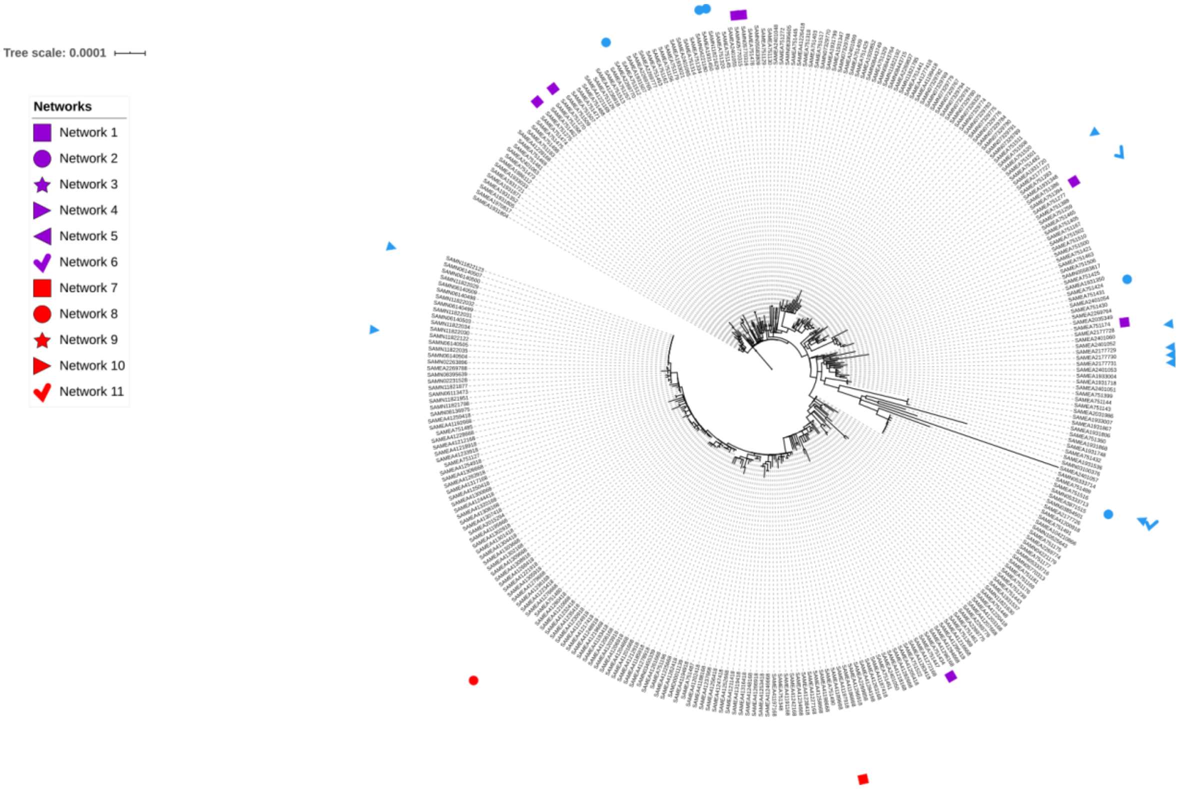


Figure 3.16. Maximum likelihood phylogenetic tree of a sub-population (n=317) of the large B. pertussis cohort studied here. All branches had >=99% bootstrap support. Mapping CNVs to the tree demonstrated strains containing CNVS belonging to the same hotspot had distant phylogenetic relationships. CNVs are therefore highly homoplasic mutations.

3.4. Discussion

B. pertussis is described as a monomorphic bacterium that has evolved as a human-specific pathogen through gene loss via homologous recombination between direct repeats (23). However, homologous recombination can also cause multi- gene CNVs. Although 12 multi-gene CNVs had been described previously(105,132,184–187), no systematic analysis of CNVs in *B. pertussis* had been carried out. In this chapter, short-read genome sequence data generated on the Illumina platform for 2430 strains was analysed using read depth as a proxy for copy number. The results revealed 11 clusters consisting of 272 CNVs, some of which comprised hundreds of genes, revealing a novel aspect of genetic variation among *B. pertussis*. This contributes to a growing literature that demonstrates that quantifying *B. pertussis* diversity requires a comprehensive view of mutation types, not just the quantification of DNA base changes.

3.4.1. Evaluation of the method

3.4.1.1. Manually resolved dataset construction and evaluation

Analysing read depth data for the presence of CNVs is an indirect method of finding CNVs and as such its use carries its own inherent strength and weaknesses. To quantify these, estimates were generated for a dataset containing known CNVs. A set of 28 isolates which had been analysed by a combination of Pacbio and Illumina sequencing and genome mapped by Opgen and Nabsys provided an excellent set of known CNVs in *B. pertussis*. The goal of the method presented here was to predict these CNVs using read depth signals from short read data alone to simplify the process of screening for CNVs.

Read depth based CNV prediction are best suited to predicting CNVs that occur as a single stretch of contiguous DNA in both the reference and the isolates being analysed. This is in addition to being affected by highly fluctuating read depth in a sample. As such, the dataset fell into three groups: ideal for analysis (n=24); not ideal for analysis (n=2) and poor quality (n=1). Two genomes with complex CNVs were separated into the ‘non-ideal’ group due to having CNVs composed of multiple SV events (Figures 3.10 and 3.11). A single genome failed quality control tests (testing for low read depth noise and <30x coverage) and was excluded from further analysis.

It was not possible to express the results of this comparison experiment formally as sensitivity and specificity because a true CNV could be predicted correctly (at a minimum 80% reciprocal overlap) but a false positive also predicted for that CNV (for example at 20% of the length of the CNV) (Figure 3.7). This therefore meant that a contingency table where each element of the test (here a CNV) could only fall into one category (true positive; false positive; true negative or false negative) was not suitable. This also meant that whilst one true negative was simulated here, it was not directly comparable to the true positives.

Establishing true negatives from real data was not carried out. This is because it cannot be verified that a closed genome truly did not have a CNV unless it had been heavily scrutinised for indirect evidence of CNVs. Only genomes which did not assemble into a single contig were heavily scrutinised for indirect evidence of a CNV and then the genome maps were consulted for further evidence. It was not possible to verify that the high levels of scrutiny that were applied to the true positives was applied to these ‘true negatives. In support of unresolved CNVs remaining in apparently fully resolved closed genomes, 16 CNVs were predicted in publicly available closed genomes and were found to be adjacent to repeats-a hallmark of true CNVs (Figure 3.15).

By varying the window length in response to the read depth noise of each sequencing sample, it was possible to normalize CNV predictions between samples. However, there were limited chances to test this methodological step as the manually resolved dataset was of a generally high quality. Flexible window lengths likely enabled the study of 1000’s of bacterial isolates of varying quality generated using a range of different instruments. This is a scale of CNV prediction in prokaryotes that is not often achieved, likely due to high false positive predictions.

3.4.1.2. A Low incidence of false positives and false negatives

A considerable strength of the pipeline was the low incidence of false positives and false negatives (Table 3.1). Out of the 24 CNVs which were judged to be highly suitable for this analysis, 23 were correctly predicted (95%), as defined here by a strict minimum 80% reciprocal overlap. An 80% reciprocal overlap is superior to many published methods which use 50% overlap between a predicted and a true CNV to be sufficient for a prediction to be considered correct (198,230). There was, however, one false negative and it was unclear why this CNV had been failed to be predicted in this case. Overall, it was demonstrated that this method was comparable to many studies.

A small number of false positives were predicted. Of the 24 suitable CNVs, 3 (13%) of these CNVs were predicted as two CNVs: a major fragment and a minor fragment (explained in Figure 3.7). In these 3 cases the major fragment was big enough to satisfy the reciprocal overlap rule and so therefore the true CNV had been predicted. The three minor fragments, although part of the true CNV, were reported as false positives (Figure 3.7). The pipeline is therefore too sensitive and I hypothesise that in certain datasets small drops in sequencing coverage are associated with an adjacent CNVs being predicted. This is a small issue, however, given that the vast majority of the predictions were highly accurate. The balance between sensitivity and specificity was achieved through testing various window lengths, to select the lengths which gave a mean over the standard deviation of read coverage between 4 and 5 (198). It could be that there are extra heuristics or noise-smoothing tools that can reduce the over-sensitivity of this analysis by further tweaking the window length. Over-sensitivity of the method is likely to inflate the number of CNVs predicted in the large cohort analysis.

A false positive CNV was predicted at a second locus in the D800 sample. The existence of a CNV was investigated in this isolate and whilst both PacBio and Illumina showed a moderate rise in coverage at this locus, genome maps did not corroborate this. It is therefore unclear exactly what was causing this rise in coverage, but it may have been indirectly caused by a mixed population of cells (see Chapter 4). Seeding cultures for DNA extracting from single colonies derived from a mixed population containing different genotypes would mean that analysis of differing genotypes was undertaken on each sequencing instrument. This would give conflicting results, as was observed for D800. The aim of this dataset was to evaluate this pipeline in the most conservative way and therefore this result was counted as a false positive, despite mixed evidence that it was a CNV.

3.4.1.3. Achieving high breakpoint accuracy

The predictions were also highly accurate, having a median distance of 0.5 genes between the predicted and the true start/end positions. However, there were predictions that had considerably worse break-point accuracy. Two of these estimates were related to the over-sensitivity of the pipeline leading to a CNV being correctly predicted but ending a large distance before the true CNV end. It is to be expected, therefore, that the CNVs predicted in the larger cohort of *B. pertussis* isolates are highly accurate.

3.4.1.4. Indirect evidence of breakpoint accuracy from novel predicted CNVs

Beyond the 28 isolates with known CNVs, an indication that the predictions were accurate was found in the analysis of the global cohort of isolates. As homologous recombination is thought to be the primary driver of SV formation in *B. pertussis,* I sought to determine the relationship between predicted CNVs and repeats. I found that the predicted CNVs were significantly closer to repeats (median: 0 gene) than the majority of genes in the genome (median 5 genes). The closer association of CNVs to repeats indicates the predictions were not spurious results with no association to the known underlying biological processes of CNV formation. As this evidence is indirect, however, this association could be caused by a different mechanism which is as of yet, unknown. Additional direct and indirect evidence of CNVs is provided in Chapter 4. The best source of direct evidence of CNVs is by capturing whole CNV arrays in single long sequencing reads, but failing that, long reads that span the end of one copy of the tandem CNV and the start of the second copy (junction sequences) provide indirect evidence of CNVs.

3.4.1.5. Copy number discrepancies

Many of the estimates presented here were non-integer values such as 1.2 or 2.3. Naively, this appears to be purely a technical artefact. An interesting feature of read-depth mapping, however, is that it is an amalgamation of information from many different cells in the population that was sequenced. This was key to the work in Chapter 4 and was one of the driving forces that led to the investigation of mixed populations of cells. It is evidenced in Chapter 4 that CNV instability is likely to account for many of these ‘intermediate’ copy number estimates.

The accuracy of the copy number prediction was demonstrated by qPCR which verified the copy number of a CNV in UK54. UK54 was chosen as it contained the highest copy number CNV in the dataset and was therefore an attractive option for qPCR, which can struggle with quantifying changes of a twofold difference (e.g. between single and double copy regions). The qPCR resulted in a prediction of 4.4 +/- 0.4, verifying the predicted copy number of 4.

3.4.1.6. A failure to resolve complex CNVs

In the comparison of predicted CNVs to known CNVs, complex CNVs could not be predicted correctly as the predicted regions did not satisfy the >80% reciprocal overlap rule. Complex CNVs are likely to be present in the large cohort of isolates and therefore there is likely a diversity in the accuracy of the predictions: some are inaccurate. As complex CNVs were judged to be ‘not-ideal’ for CNVnator to predict, the two isolates with complex CNVs were analysed separately in order to investigate how the pipeline interprets this data. The CNV in the B199 strain was composed of nested CNVs arranged in an extremely complex way whilst in F701 there was a CNV disrupted by a small genomic inversion compared to the reference B1917 sequence.

When CNVnator was used to predict these CNVs, none of the results shared an >=80% reciprocal overlap with the true CNVs and thus they were not predicted correctly. However, it is notable that the region had been predicted as a CNV in B199 although it did not satisfy the overlap rule. An additional problem was that the isolate F701 had been predicted as two separates but adjacent CNVs which indicates over-specificity (see below). However, even merging the CNVs into a single unit for F701 did not enable the predictions to pass the strict reciprocal overlap threshold. Strain F701 from the manually resolved dataset had a small inversion and it is likely that some isolates in large cohort dataset would have been disrupted by bigger inversions and therefore be predicted as two CNVs that are distant on the B1917 genome but may be contiguous on the CNV isolates. This effect may have inflated the number of CNVs predicted and the basis of future work could be to resolve many of these isolates using long read technologies. Current platforms and analysis appear be unable to generate consensus sequences in isolates with large CNVs (178), however. These results highlight how challenging it is to predict complex CNVs using read depth alone. It is likely that in the large cohort dataset that a number of CNVs would be complex and therefore have reduced breakpoint accuracy.

3.4.1.7. Pipeline conclusions

Taking all the results of the comparison into consideration, the most conservative statistic is that 74% (20/27) of CNVs of adequate quality were correctly predicted with excellent breakpoint accuracy. Less conservatively, including the 2 extra CNVs which were predicted as a correct (satisfying overlap rules) major fragment but also generated false positive minor fragments, the accuracy was 85% (23/27). This is similar to what is reported for CNVnator and what is generally considered good for read-depth based CNV prediction tools.

3.4.2. Analysis of CNVs in a large cohort of *B. pertussis*

3.4.2.1. Theoretical hurdles in analysing hotspot loci

It was readily apparent from the heatmap (Figure 3.11) that CNVs occurred at specific loci. Without further analysing this data it was not possible to gain further insight into the structure of hotspot loci. This was a considerable theoretical hurdle to overcome and I eventually found that network graphs were an attractive solution to this problem. I considered how novel describing hotspot loci using networks was, considering that hotspot loci have been found in multiple species. It was apparent that network analysis of CNVs is almost non-existent in the literature. Retrospectively searching for publications which utilise network graphs to analyse the spatial organisation of CNVs returned only one methods paper (detailing the bioinformatics tool HD-CNV) with a modest number of citations (15 as of 27/01/2019) (231). Many of these studies utilised HD-CNV to simplify their CNV calls in eukaryotes to a non-redundant set of regions that experienced frequent CNVs, but went no further in analysing the composition of these networks (232,233).

In the present work I exploit the properties of networks to describe how much CNVs overlap, how frequently they overlap and which genes are contained in the majority of CNVs in a network. The analysis shown in this chapter extends networks beyond just a way to group overlapping CNVs and is a novel utilization the inherently flexible and generic properties of networks to analyse the poorly described phenomena of CNV hotspots. The generic nature of networks is particularly important given CNVs are a ubiquitous feature in all kingdoms of life.

An advantage of using networks to describe hotspot loci was the ability to semantically categorise CNVs to unite the findings of many studies and contextualise them with new data. Previously, using limited data, a ‘hotspot-like’ effect had been demonstrated by resolving four CNVs with subtle gene content variations at the same loci corresponding here to Network 1, at which other CNVs had also been reported(105,132). This is contextualised both by the dataset of 24 high quality manually resolved CNVs (Figure 3.6) and the results from the *B. pertussis* cohort which provided another 90 CNVs at this location. I propose that the naming convention proposed here be used in future analysis to categorize CNVs in *B. pertussis* in order to generate new insights.

3.4.2.2. The relationship between network analysis and prediction accuracy

It was demonstrated that the pipeline was oversensitive, thus leading to some CNVs predicted as multiple fragments in addition to a number of predictions to stop short of the true CNV end. This had the potential to impact how the data were interpreted. It is possible that this fragmentation causes an inflated number of CNVs to be predicted. I examined how this might have a knock-on effect to how CNVs were analysed as networks.

Fragmentation of CNV calls was unlikely to significantly affect the network density statistic. Split-CNV calls were observed in two cases (8%) of the ‘high suitability’ dataset and 27 (10%) of the 272 CNVs in the large cohort were found to be directly adjacent (<3 genes). This is also under the assumption that splits do not fall directly on regions which are CNV start/end points in other CNVs, which appeared true (Figure 3.12). Whilst each extra fragment would increase the total number of possible connections in the network, under these assumptions each of the new fragments would inherit all of the overlaps of the true CNV by the same amount and therefore the network density would be unaffected.

The core network statistic was generated by taking the genes that were contained in 90% of the genes in CNVs within a network. As it appears that predicted CNVs do not get split into multiple predictions at the same genes, CNV fragmentation would not decrease the size of the network core for large networks but may have a minor impact on smaller networks.

3.4.3. Potential impacts of CNVs

3.4.3.1. Potential phenotypes of CNVs

One of the most frequent hotspots observed in this study included flagella biosynthesis genes which have been shown to provide motility to *B. pertussis* (Table 3.2 & Table 3.3). These genes may be linked with colonisation and/or virulence as they were found to be expressed during murine challenge(45). The exact role these genes play in these processes is unclear, although at least motility and biofilm formation are clear roles for these genes in the genus, with other roles also possible. *B. pertussis* has been shown to be motile and produce flagellum on their cell surface (225) in addition to flagella being vital to biofilm development in *B. bronchiseptica*

(234). Flagella encoding genes w ere upregulated in the initial stages of B. bronchiseptica biofilm formation but downregulated during the mature biofilm stage, after 48h (234).

Tandem duplication of the genes encoding the flagella apparatus has not been documented before, but the literature suggests that overexpression of these genes results in increased flagellum on the cell surface and can lead to higher motility (235,236). It is therefore possible that increased gene dosage of the flagella encoding genes would lead to increased motility. However, Yang et al found that overexpression of the flagella apparatus effected the integrity of the membrane, making cells vulnerable to osmotic pressure and macrophage killing (237). It is likely that any increase in flagellum therefore comes with additional costs to the cell which may be tolerated by *B. pertussis* if enhanced motility was under positive selection.

Motility is a phenotype that has been described recently in *B. pertussis*. Hoffman et al found that some isolates were never motile but others were motile in only a proportion of experiments (225). It is yet to be determined if the observations of Hoffman et al reflect the true behaviour of B. pertussis or are artefacts of the assay. It appears that flagella genes have not previously been subjected to tandem duplication in the manner described here, although Dalet et al observed that the unstable phenotype of high haemolytic activity (185) was caused by CNVs. The instability of CNVs may be contributing to the instability of the motility phenotype as described by Hoffman et al or may be an added layer of cell-to-cell heterogeneity, which has been described previously by different mechanisms in other species (238–240).

The core of network 3 contained the *fim3* gene which codes for the major subunit of the serotype 3 fimbriae (most commonly annotated as *fimA*)- a potent virulence factor that allows adhesion to the host cells (63,241,242). The *B. pertussis* fimbriae fall into the type 1 category. Amplification is likely to increase expression and translation of this gene and as such, whilst amplification of fimbriae has not been previously documented, this mutation is comparable to overexpression of fim. Overexpression of the major subunit of the fimbriae has been shown to cause elongation of the structure in *Porphyromonas gingivalis* which increased adhesive properties of the cells(243,244). Similarly, Vizcarra et al, who overexpressed the whole *fim* system, *fimA-fimH*, found increased adhesion, macrophage survival and intracellular bacterial load of *Escherichia coli* (245). Considering the whole *fim* operon was overexpressed, it was likely that overexpression of *fimA* alone was not enough to cause such phenotypes. It is therefore unclear the exact impact that amplifying only fim3 would have on *B. pertussis* but given the importance of fimbriae to virulence for many pathogens, warrants further investigation.

3.4.4. Is the observed distribution of CNVs driven by selection?

Selection acts on phenotypes which are caused by specific genotypes and as such, how CNVs may affect the phenotype of an isolate are important to investigating the selection pressure on CNVs in *B. pertussis*. Selection also directly leaves a signature on the frequency and distribution of CNVs in bacteria, however, and this is discussed here (125). The large number of copies of IS481 throughout the *B. pertussis* genome suggests that a very large number of different genome rearrangements and CNVs are possible (23). Despite the vast diversity of possible CNVs, however, 94% of observed CNVs appeared at just 11 hotspot loci. Similarly, Weigand et al demonstrated that there were a number of conserved gene orders (128, 129).

I suggest that this discrepancy between the potential and observed distribution of CNVs is an indication that strong purifying selection acts on CNVs in *B. pertussis*. This is supported by studies that show that in other species genome-wide genesis of structural variants can be shown (166,246,247), but other studies show CNVs only rise to noticeable frequencies under positive selection (125,163). Whilst this hypothesis is supported by experiments on other bacteria, how this impacts *B. pertussis* is unclear. It is also possible that this is attributable (at least in part) to certain loci having a higher likelihood of mutation (166,246).

A seminal study by Anderson and Roth (246) found that in *Salmonella*, certain regions of the genome experienced higher rates of duplication than others, based on their proximity to rRNA operons. This was carried out by ‘trapping’ CNVs as they form by transduction assays (165, 234, 236). However, it is unclear how this translates to *B.pertussis*, which is replete with repetitive DNA. Further analysis of the work presented here could investigate if the hotspot loci have a denser than average cluster of repeats at the median start and end points.

In order to quantify selection pressures acting on DNA sequences, the most used tool is the dN/dS ratio (249), the ratio of non-synonymous (dN) to synonymous mutations (dS) which in this case, however, is not applicable. The unstable nature of CNVs in bacteria, and in particular in combination with the low rate of SNPs in *B. pertussis*, mean that the lifespan of a CNV is much too fleeting to acquire even a single SNP- synonymous or not. This short lifespan was described in the ancestral state reconstruction analysis- hypothesised inheritance of a CNV occurred over a timescale that corresponded to just a few SNPs across the whole genome, in the most pronounced example.

3.4.5. Applicability of the work outside of *B. pertussis*

Whilst the repetitive nature of the *B. pertussis* genome is unusual, it is not unique, as its abundance of IS elements ranks in the top 30 in a study of 1000's of bacterial isolates (250). It is likely that CNVs are a common mutation type amongst genomes with repetitive genomes (224,251). The unusually high number of insertion sequences within the *B. pertussis* genome, and their relatively even distribution, likely facilitates the genome-wide distribution of structural variants. Indeed, genomes of related species *B. parapertussis* and *B. holmesii* each harbour fewer IS elements and thus exhibit fewer rearrangements 50 and very rare CNVs (M.R. Weigand, unpublished). Future work could seek to replicate these analyses for many species and indeed at least one study (PanX) has sought to do this (252), although with its own strengths and weaknesses.

4. Chapter 4- The i*n vitro* genome plasticity of *B. pertussis*

4.1. Introduction

The instability of the *B. pertussis* genome is likely to have given rise to the documented deletions, inversions and CNVs that have been described in the global population of *B. pertussis* . However few studies have documented the dynamics of the *B. pertussis* genome *in vitro*. This is highly important as the genome dynamics in small populations over very short timescale of hours and days underpins the behaviour of large populations over long timescales, such as those documented in Chapter 3. Given that CNVs are known to be highly unstable, I hypothesised that heterogenous populations of cells containing different copy number states of the CNV would be rapidly generated *in vitro*. In this chapter I aimed to investigate how populations of B. pertussis cells changed CNV copy number over short timescales. Establishing methodologies to do this will allow future research to achieve results on a large scale which in turn would allow a viable comparison between the spectrum of mutations *in vitro* vs the spectrum of mutations in sequenced samples, themselves likely a mixture of mutations forms by *in vitro* and *in vivo* pressures.

A variety of methods can be used to study copy number changes in single cells within a population, but most involve synthetic constructs in genetically malleable organisms. Few methods are able to study naturally occurring CNV dynamics in bacterial populations (124). Small repeats (10-100bp in length) can be captured by small reads, such as those generated on the Illumina platform. The CNVs presented in this thesis, however, range in size from 10-350kb and therefore to capture them in the tandem configuration, would require reads up to approximately 700kb long. To study these CNVs fully, therefore, I aimed to study CNV dynamics by generating ultra-long read Nanopore data, analysing it for CNVs and exploring how to assemble this data to resolve CNVs.

4.2. Methods

4.2.1. Nanopore sequencing

B. pertussis strains were stored at -80^{o}C in PBS/20% glycerol at the University of Bath. Bacteria were grown for 72 hours at 37^{o}C on charcoal agar (Oxoid) plates. Harvested cells were resuspended in 10 ml SS broth to an OD_{600} of 0.1 and grown overnight. At approximately OD_{600} 1.0, cultures were diluted in 50 ml SS broth to an OD_{600} of 0.1 and grown to OD_{600} 1.0. Bacteria were centrifuged at 13 000xg for 5 minutes and processed for gDNA extraction using the protocol available from dx.doi.org/10.17504/protocols.io.mrxc57n. The rapid adaptor (SQK-RAD004) Nanopore library preparation steps were included, adapted for sequencing of very long gDNA molecules.

DNA was sequenced for 48 hours on GridION or MinION sequencers using R9.4 flow cells. Base-calling was performed with Guppy (V2.1.3 or V3.2.1) using the “fast” Flip-flop model. Reads spanning the CNV locus were identified using Blastn alignment with a minimum query length coverage of 90% for the 16kb CNV locus and 10% for the single copy flanking regions (~1kb). DNA preparation and Nanopore sequencing was undertaken by Natalie Ring.

4.2.2. Identifying SVs in single long reads

Each read was compared to the consensus sequence using a BLAST search. This was technically very challenging as this process was frustrated by the high repeat content of B. pertussis. Each time a repeat sequence was found, it was a technical hurdle to decide if there was a sequence following the repeat that would be expected or if there was an unexpected sequence (indicating an SV). To overcome this efficiently, it was necessary to remove the repeat gene content from the B. pertussis genome sequence. The sequence was split up into 1kb windows at a step of 200bp and any 1kb section that appeared more than once with adequate homology in at least 50% of the length of the window was deleted. For UK54 this removed 600kb of sequence giving a consensus sequence length of 3.5Mb. Whilst this was highly conservative it aided efficient and accurate analysis of reads.

4.2.3. Generating random sequences and comparing homology

In order to demonstrate how challenging homology matching a read to a Nanopore adapter is, a simulation of this task was undertaken. An adapter sequence was compared to 2000 simulated true negatives (random sequences with GC content of 67%- the same as B. pertussis) and 2000 true positives, the first 100bp of each read. This was undertaken using BLAST.

4.3. Results

4.3.1. Indication from short read data of mixed populations

Intermediate copy number estimates were most recognisable in the dataset of 25 manually resolved genomes (Figure 4.1). The discrepancy between the estimated copy number and the resolved copy number was evident as 52% (13/25) strains had a copy number that differed by at least 0.3. Whilst both sequencing and genome mapping data supported the final genome sequence assembly in these isolates, the data was still a consensus, e.g. an average sequence for the sequenced bacterial population. This means there may have been cells in the population with a different genome sequence, potentially with 3 copies of the locus, for example. A similar trend was observed in the 272 CNVs predicted in the large cohort dataset, in which 71% (193/272) of CNVs were predicted to be have non-integer copy numbers (+/- >=0.3 of an integer); although the true copy numbers of these isolates were not known.

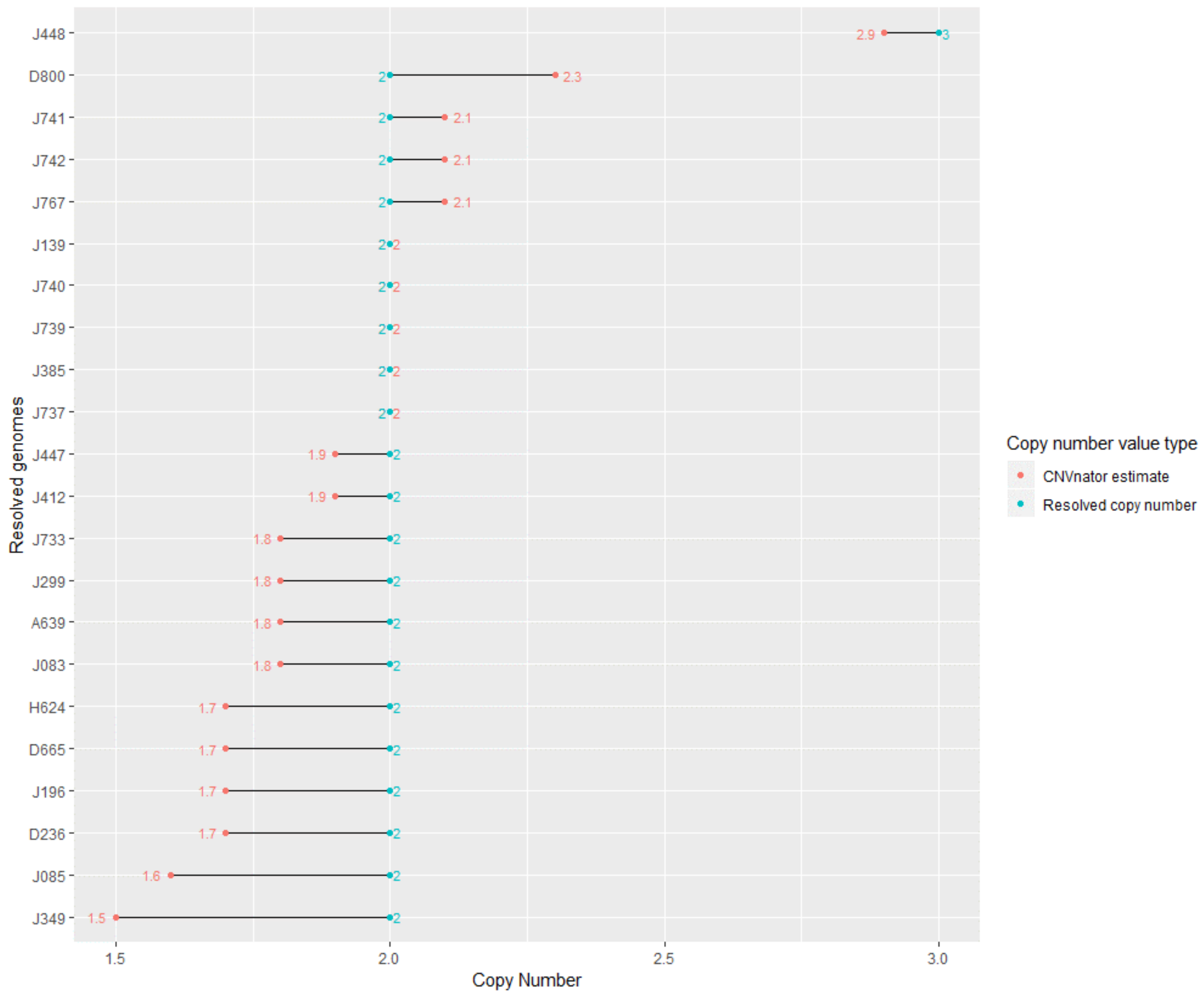


Figure 4.1: The true (orange) and estimated (blue) copy number (X axis) of CNVs was plotted for the manually resolved cohort (Y axis). Large discrepancies between the estimate and the true CNV copy number state can be seen for the majority of isolates.(Reproduced from Chapter 3)

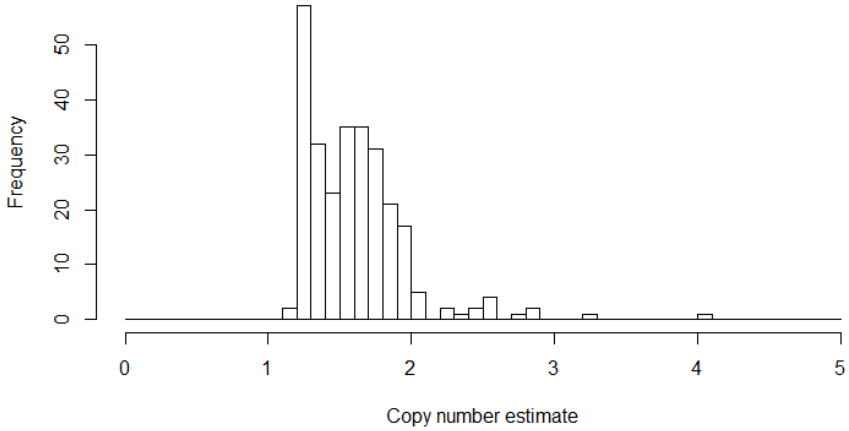


Figure 4.2: Copy number estimates (X) and their frequency (Y) in the full dataset of 273 CNVs described in Chapter 3. It can be seen that most copy number estimates are not integers nor are they clustered closely at integers but are instead a relatively even distribution between integers. This may have been artefactual or a sign of mixed populations of cells with varying copy numbers being present in the populations sequenced.

A technical artefact or combination of artefacts could have been the source of such a pattern, for example: uncorrected sequencing bias (such as poor coverage of certain motifs (253); bioinformatic analysis (e.g. not strict enough coverage cutoffs or not enough normalisation) or sample preparation (254). Alternatively, this pattern could be produced, at least in part, by a biological source. It is hoped, in most sequencing applications, that all cells in the population are clonally derived and have identical genomes. This leads to an easy to analyse genome sequence during the assembly stage: all reads (derived from a number of different cells in the population) will contain the same DNA sequence differing only by sequencing errors. Whilst ideal, this scenario is not often achieved as some mutation types occur so frequently that within just a few generations, either within the host or in-vitro after isolation, variation between clonally derived cells is generated. Most of these mutations are small and inaccuracy in determining their exact composition is well tolerated for most applications. The best-known example is slip-strand mispairing mutations in homopolymeric tracts (255). In model organisms with few repeats, copy number variants are also a mutation type that occurs very frequently, far more frequently than SNPs (161,166,167,246,256). This may be exaggerated in *B. pertussis*, however, which appears to have a low SNP rate and a high number of repeats (42). It was therefore possible that the intermediate copy numbers observed in the dataset were products of cell to cell variation in copy number. These fluctuations may have occurred before (in-vivo) (167) and/or after (in-vitro) isolation of the sample.

4.3.2. Nanopore sequencing confirms mixed populations

4.3.2.1. Resolving CNVs to find cell to cell differences

In order to understand if mixed populations of cells were driving intermediate copy number estimates in *B. pertussis* it was necessary to resolve CNVs on single DNA molecules using long read sequencing. It was hypothesised that if reads could be sequenced that were longer than a CNV in its tandem array then cell to cell differences could be studied. For example, if two reads were found to span the entire CNV locus including the single copy flanking regions, but each read contained a different copy number of the locus it could therefore be deduced there was at least two genotypes in that population. The most parsimonious explanation of such a result is that the two reads came from different cells, each with a different copy number. In this way, long read sequencing can be used to investigate cell-cell differences in copy number.

Whilst long read sequencing on the PacBio platform has been undertaken previously to resolve CNVs in *B. pertussis*, this was achieved only in combination with genome maps (which produce DNA fragments 50-750 kb long) as a guide (105,132). The short-read length of Illumina data or the limited length of PacBio reads (on average 10kb long), prohibits the study of CNVs as the reads are shorter than the CNV length. The Nanopore sequencing platform was therefore very attractive as its unique architecture allows long reads (reads 1Mb-2.2Mb have been observed previously (257)) to capture a whole CNV in its tandem array (predicted to be 50kb to >600kb). Previously, *B. pertussis* isolates had been sequenced on the Nanopore platform using standard length reads and whilst this produced closed genomes, there was evidence that large CNVs remain unresolved (178).I therefore investigated using ultra-long Nanopore reads in resolving long CNVs in *B. pertussis* and in turn, investigating mixed populations of cells.

4.3.2.2. Intermediate copy numbers in short read data linked to mixed populations in UK54

To confirm my predictions from short-read sequencing data (Figure 4.2 and 3) and investigate the basis for non-integer copy numbers, we exploited the tractable size of one relatively small CNV. The genome of UK54 (SAMEA1920853) was predicted to have a 16 kb CNV at a copy number of 4.1; short enough to observe the entire CNV locus in a single sequence read on the Nanopore platform, assuming that each copy occurred in tandem as observed in the manually resolved dataset in Chapter 3.

Ultra-long DNA was prepared according to the Quick protocol (258). Whole genome sequencing on the Nanopore platform yielded 85k reads. This sample contained a median and mean read length of only 1kb and 9.1kb respectively but produced over 3000 reads with a length exceeding 50kb (Figure 4.3). Sequence reads that contained both flanking regions of the CNV locus and the CNV locus itself were identified (n = 9) and contained the CNV at different copy numbers (Figure 4.4). This demonstrated that a laboratory culture of UK54 comprised a mixture of copy numbers at this locus and explains the non-integer copy numbers predicted by the short-read prediction pipeline. Genomic DNA for sequencing is derived from laboratory populations of bacteria and if these harbour CNVs at different copy numbers, subsequent read-depth based predictions will represent the average read depth of all of the bacteria sequenced.

It was not known if the original culture of UK54 involved isolation of a single colony or collection of multiple clones from the diagnostic plate growth. Thus, it was unclear whether the observed variation in copy number resulted from collection of a mixed population or emerged during laboratory growth prior to sequencing. To investigate this, single colonies of UK54 were isolated.

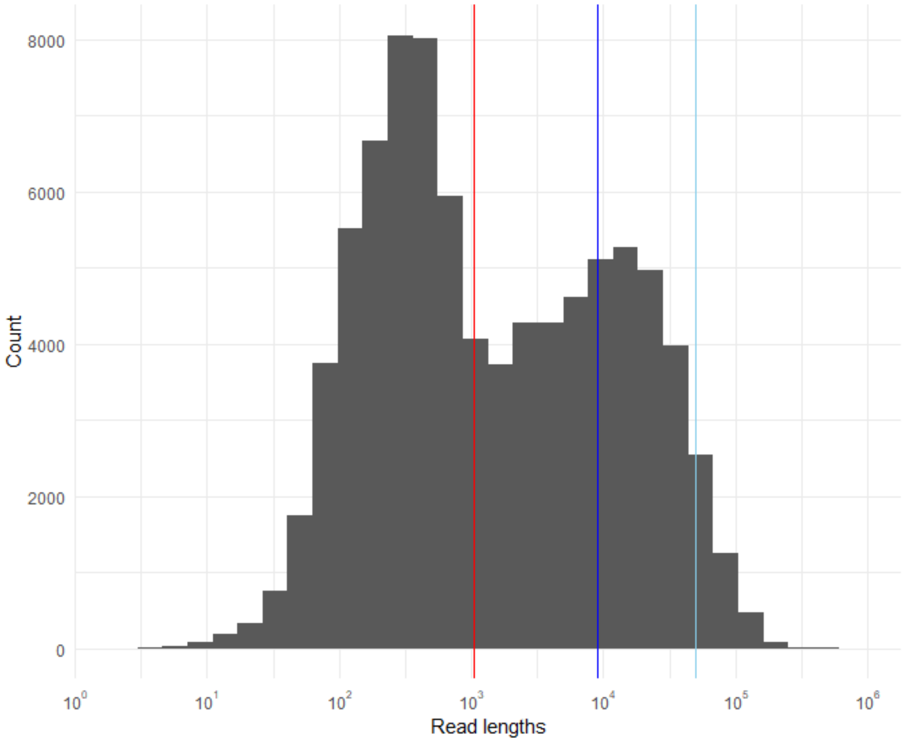


Figure 4.3: Read length (X axis) histogram for UK54. A median of 1kb (red line) and a mean of 9kb (blue line) were observed in addition to over 3000 reads exceeding 50kb in length (light blue line).

4.3.2.3. Mixed populations from pure cultures

To investigate mixed populations, eight single colonies of UK54 were passaged once by growth on agar and then once by growth in SS broth. Each of these clonal populations were theoretically derived from a single bacterium. The copy number of the CNV locus in each of the resulting clones was estimated using qPCR (Figure 4.5) and ranged from 2.2 (clone 6) to 51.2 (clone 8).

The results were not conclusive in regards to what timescale this variation was generated however and the question still remained: were CNVs unstable over very short timescales (<30 generations)?

Seven of the 8 clones had expected copy numbers of between 1 and 4, but clone 8 had an unexpectedly high copy number. Whilst a copy number of 51 may seem improbable from a population which has an average copy number of 4, it is known that tandem duplications are highly unstable and prone to further amplification (155,161,167,208,211). This clone was further analysed using Nanopore sequencing.

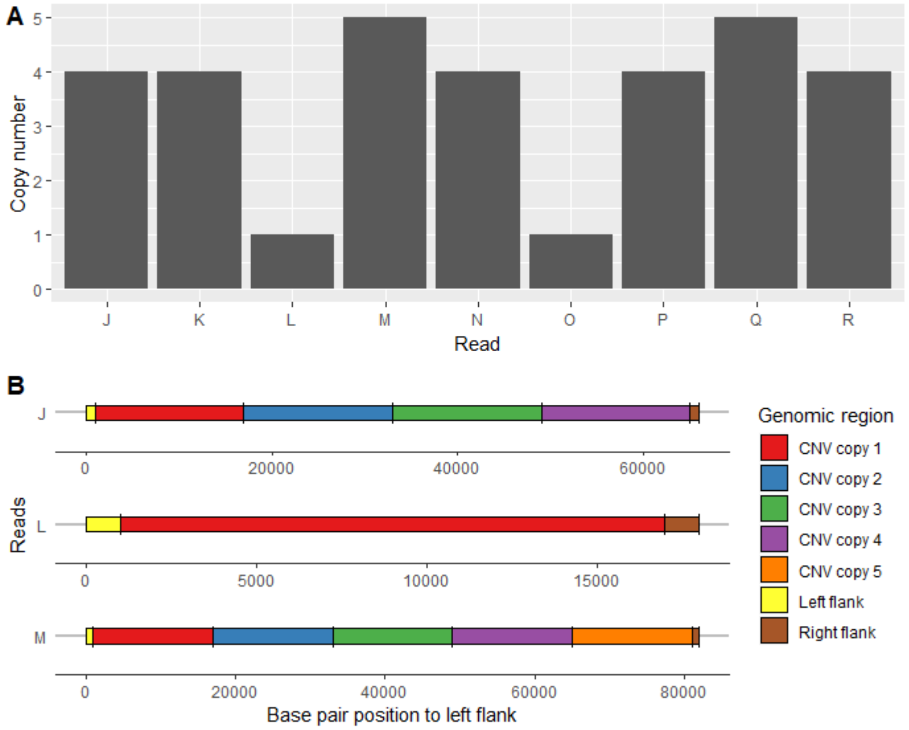


Figure 4.4: Ultra-long read sequencing of UK54 revealed the presence of different copy number CNV loci within a single culture. Individual sequence reads that spanned the CNV loci were identified using BLASTn, labelled J to R. (Panel A). The data shows each read (x-axis) containing 1, 4 or 5 copies of the locus (y- axis) and therefore, as each read appears to be integrated into the chromosome, there were cells present in the population with 1, 4 or 5 copies of the locus. The arrangement of the relevant section of three reads (J, L and M) is illustrated in panel B.

The genome dynamics of CNVs in *B. pertussis* were investigated by ultra-long nanopore sequencing two of these UK54 clones, a low copy (clone 4, copy number 4.3) and high copy (clone 8, copy number 51.2) clone. It was hypothesised that as these samples were clonally derived each sample should, if copy number was not plastic over short time scales (the null hypothesis), contain reads containing the CNV locus at the same copy number. The same BLAST search method was applied as previously.

For clone 4 sequence reads were observed with copy numbers 2 and 5 (Figure 4.8). These data strongly suggested that CNV copy number was plastic, with copy number variants arising during in vitro growth from a single bacterium to the culture from which the gDNA was extracted. Together, these results (Figure 4.5 & Figure 4.6) suggest that copy number change of CNVs is a dynamic, fluid and continual process in *B. pertussis*.

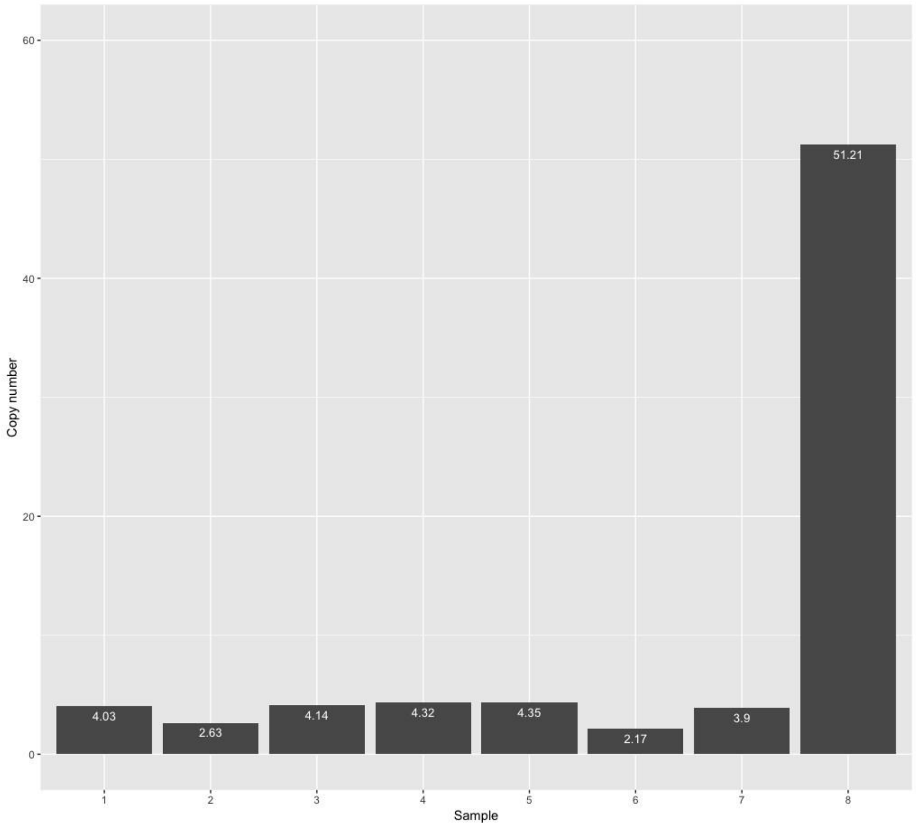


Figure 4.5: Quantification of CNV copy number (Y axis) of 8 clones of UK54 (X axis) by qPCR demonstrated a range of copy numbers from 2.17 to 51.21.

For clone 8, no reads spanning the entire CNV locus (i.e. the CNV locus with single copy flanking DNA on each side) were produced, presumably due to its extreme length (predicted length, 816kb) (Figure 4.6). However, reads containing up to 7 copies of the locus, without flanking regions, were identified. Relaxing the BLASTn alignment parameters from a 90% minimum query length of the CNV locus to 50% identified a maximum of 9 copies of the locus present on a single read with incompletely sequenced copies at each end. Consistent with the copy number prediction from qPCR, the read depth at this locus for UK54 clone 8 from the Nanopore data was approximately 60x higher than the genome average, strongly supporting the very high copy number estimate for this locus in this clone (Figure 4.7).

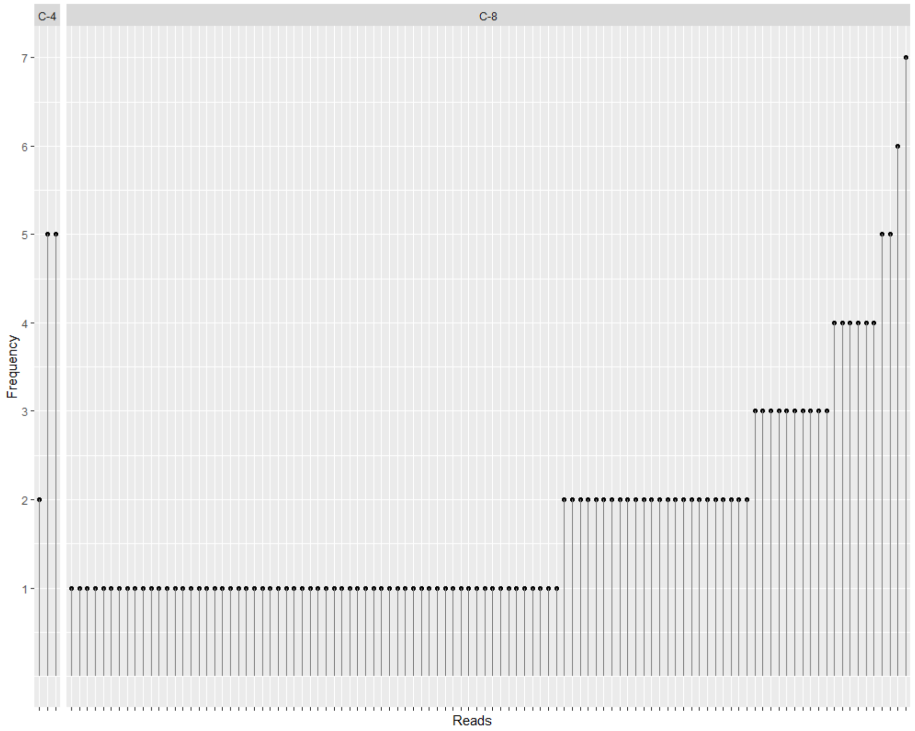


Figure 4.6: Ultra-long read sequencing of UK54 clones 4 and 8 (C-4 and C-8) revealed the presence of different copy number CNV within a single culture. Individual sequence reads that spanned the CNV loci were identified using BLASTn. Such reads were identified in clone 4 whilst clone 8 had no reads spanning the full locus. Therefore, partial reads were plotted for clone 8. The data shows each read (X axis) contained between one and seven copies of the locus (Y axis).

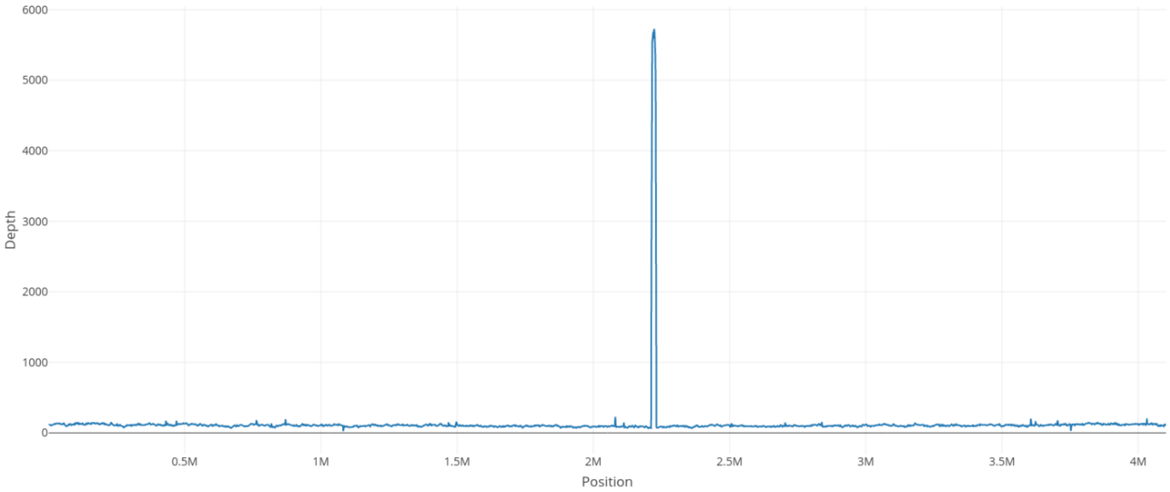


Figure 4.7: 10% of the ultra-long Nanopore data generated from the UK54 clone 8 sample was mapped back to the UK54 consensus sequence. A clear spike in coverage can be seen at the CNV locus which corresponds to approximately 60x higher coverage (baseline coverage ~100 and peak coverage of ~6000).

4.3.2.4. Gene expression is linked to copy number

A preliminary investigation of the effect of CNV formation on phenotype was undertaken. It was reasoned that an increase in gene copy number by CNV formation would increase the relative level of gene expression for that gene, compared to genes outside of the CNV. To investigate this, the relative expression level of one gene within the CNV locus was compared to a non-amplified gene, in three UK54 clones was measured. We selected clones 2, 4, and 8, with originally screened copy numbers of 2.63, 4.32, and 51.21, respectively. RNA expression of gene B1917_RS10525 (CNV gene) was normalised to the single copy gene recA, often used as a stably-expressed, housekeeping, control gene in RT-qPCR experiments (259).

As it was demonstrated that each culture comprises a heterogenous mixture of cells with varied CNV copy number, the locus copy number for each clone was re-assayed using the same laboratory culture from which RNA was extracted. Upon regrowing these clones for RNA extraction, the average copy number in each changed (statistically non-significantly) to 4.1, 6.5, and 53.1 in clones 2, 4, and 8, respectively.

The relative expression of the gene B1917_RS10525 correlated with the copy number (Figure 4.8); normalising the transcript level in clone 2 to a value of 1, it was 16.8-fold higher (P<0.0001) in clone 8. It was also higher, but not significantly, in clone 4 (P=0.76). However, broadly, using the data as a whole, there was correlation between DNA copy number and transcript abundance. This strongly suggests that the CNV had a gene dosage effect.

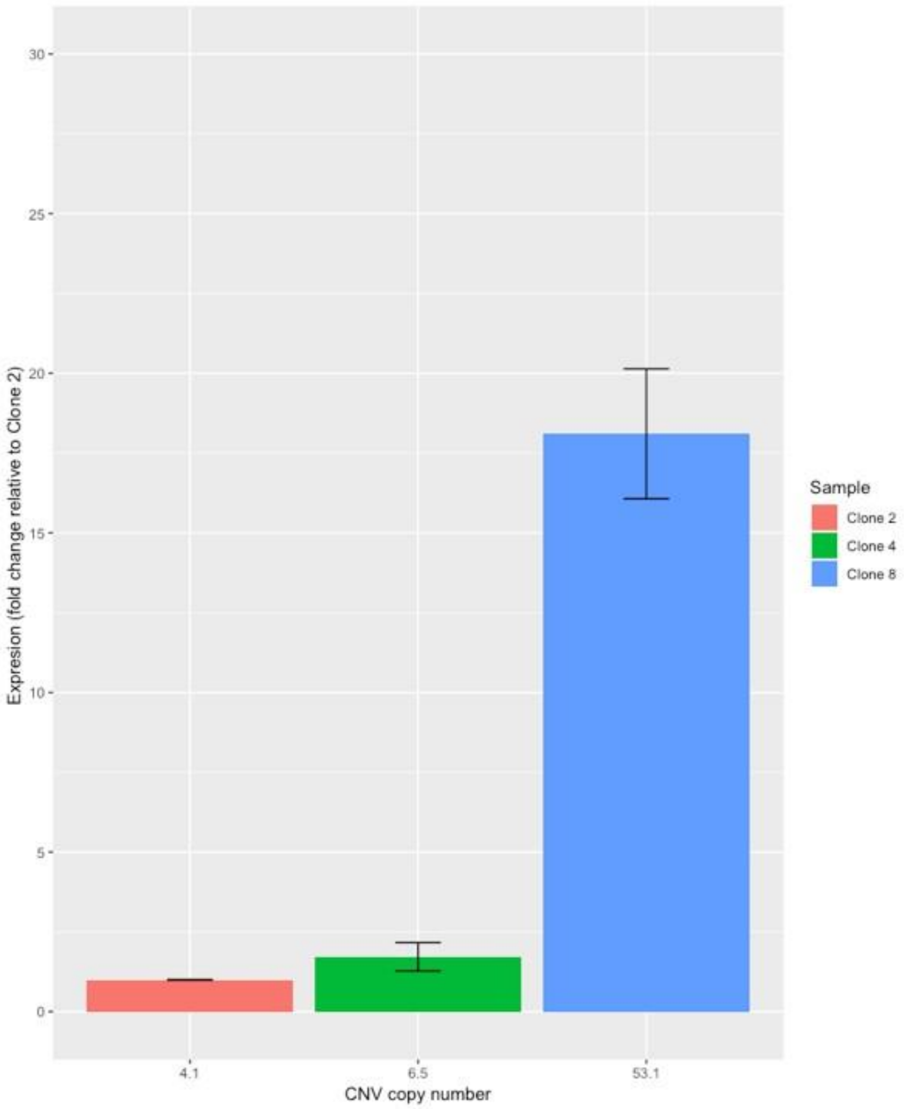


Figure 4.8: Copy numbers of clones 2, 4 and 8 were quantified using qPCR (X axis) and expression of a gene within the CNVs was quantified by RT-qPCR. Expression is shown as a relative fold change to Clone 2 (Y axis). Error bars represent standard deviation of relative gene expression. The results show that copy number corresponds to the level of expression.

4.3.3. Genome wide structural variants in Nanopore reads

Analysing the Nanopore data from clonally derived populations strongly suggested that the CNV locus in UK54 was plastic, leading clonal populations to quickly diversify. These analyses and experiments, however, studied only one specific locus- the locus that was predicted to change. I hypothesised that other genomic loci were also undergoing structural variation. I tested this hypothesis in order to generate further insights into genome dynamics of *B. pertussis*.

4.3.3.1. Naive identification of SV events in single reads

To investigate if other loci underwent SV during in vitro culture, reads that contained sequences which were proximal on the read but distant on the consensus genome sequence were identified. These reads were derived from putative structural variants. After reads were BLASTed against the consensus genome, any read which satisfied the SV criteria (see methods) was studied further. This resulted in 29 reads in this analysis.

It is naive, however, to assume that all these reads are true SVs. It is possible that they are sequencing errors known as chimeras. These errors are named as such because they read derived from two or more distant parts of the genome. They can arise either physically or through data analysis (expanded on below) and have a characteristic appearance (260,261). In order to verify these results, the hypothesis that at least some of these reads are chimeras was thoroughly investigated.

4.3.3.2. How do sequence errors appear?

It is possible that adapter sequences can be attached to two DNA fragments on either side, effectively joining two random parts of the genome and thus, when sequenced, appearing as a structural variation. It is also possible that two separate reads can pass through the pore in quick succession, leading to them being classed as a single read. Both situations produce chimeric reads. Some studies have estimated that chimera formation happens at a rate of up to 2% (260,261) during Nanopore sequencing, although this may depend on whether a ligase enzyme is used in the library preparation, which was not the case here.

There are three traits of SVs that allow them to be distinguished from chimeric reads: long length, association with repeats and gapless junctions. This means these characteristic hallmarks in combination are rare to appear by chance in sequence data. This is in contrast to SNPs for which, due to their comparatively simple nature, it is impossible to tell the difference between a true SNP and a sequencing error- there is simply not enough information. It is therefore theoretically possible to distinguish between an SV and a sequencing error in a single read.

There is a type of sequencing artefact that can be mistaken as a SV. Adapter sequences are synthetic DNA that is added to the sequencing reaction and does not occur in the sequenced genome. When two reads pass through the nanopore in very quick succession, or when two read are accidently ligated together, a chimeric read composed of two parts of the sequenced genome is produced. In order to distinguish between chimeric reads and SV reads I hypothesised that when each read was BLASTed against the consensus sequence the chimeric reads will have a gap in sequence homology to the reference and that within this gap an adapter sequence may be found.

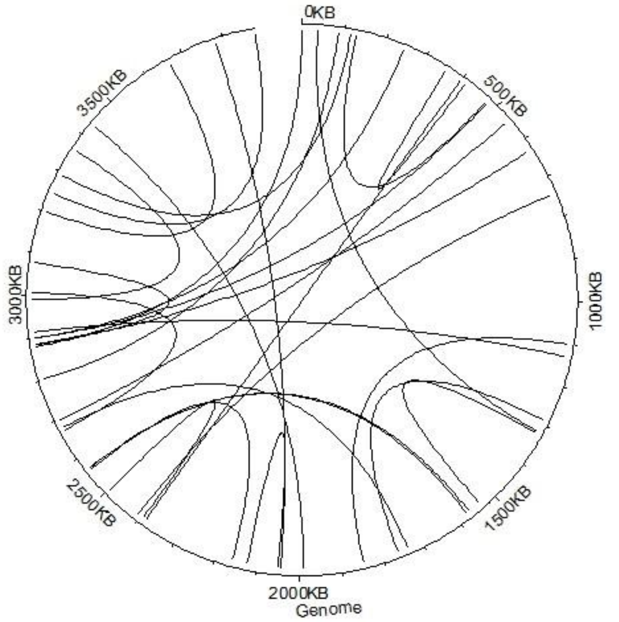


Figure 4.9: A Circos plot with all the 29 reads found to have juxtaposed genome positions. Lines join the two regions of the genome sequence that were proximal on the read but distant on the genome.

4.3.3.3. Gaps in alignments indicate chimeric DNA

Analysing the BLAST results, three main types of reads were found: reads with junctions in close proximity (but not immediately adjacent) to repeats (Figure 4.10); reads with gaps immediately adjacent to a repeat element (Figure 4.11) and gapless reads (Figure 4.12). The Circos plot was modified to reflect these categories (Figure 4.13).

Nine reads were found to contain junctions (with or without gaps) in close proximity to, but not directly adjacent to, repeats. Reads with junctions in close proximity are likely identified in this experiment because I selected reads in which the junction occurred within 1kb up or downstream of an insertion sequence. This was done to study the appearance of chimeras in this analysis as I expected chimeras to have junctions outside of repeat sequences. These reads may indicate CNVs arising from alternative processes which make SVs (such as non-homologous end joining) chimeric reads. They were analysed for adapter sequences and then excluded from further analysis. Future work will include limiting the junction to occurring only within 50bp of the repeat to only capture likely SVs.

Reads with gaps immediately adjacent to a repeat sequence are likely to be chimeras, based on the hypothesis that an adapter sequence would cause a gap in an alignment. Several reads appeared with no gap in the sequence alignment. I hypothesise that these reads are true SVs that have occurred during brief culturing. Whilst gaps indicate that a read might be a chimera, the presence of an adapter sequence in this gap is unequivocal evidence of this. Adapter sequence searching, however, is complicated by the inherent raw error rate of Nanopore reads which, for the chemistry used here, results in approximately 75-90% accurate reads. It was therefore necessary to identify reads not only containing the perfect adapter sequence, but sequences that were similar to the adapter sequence- a fraught process.

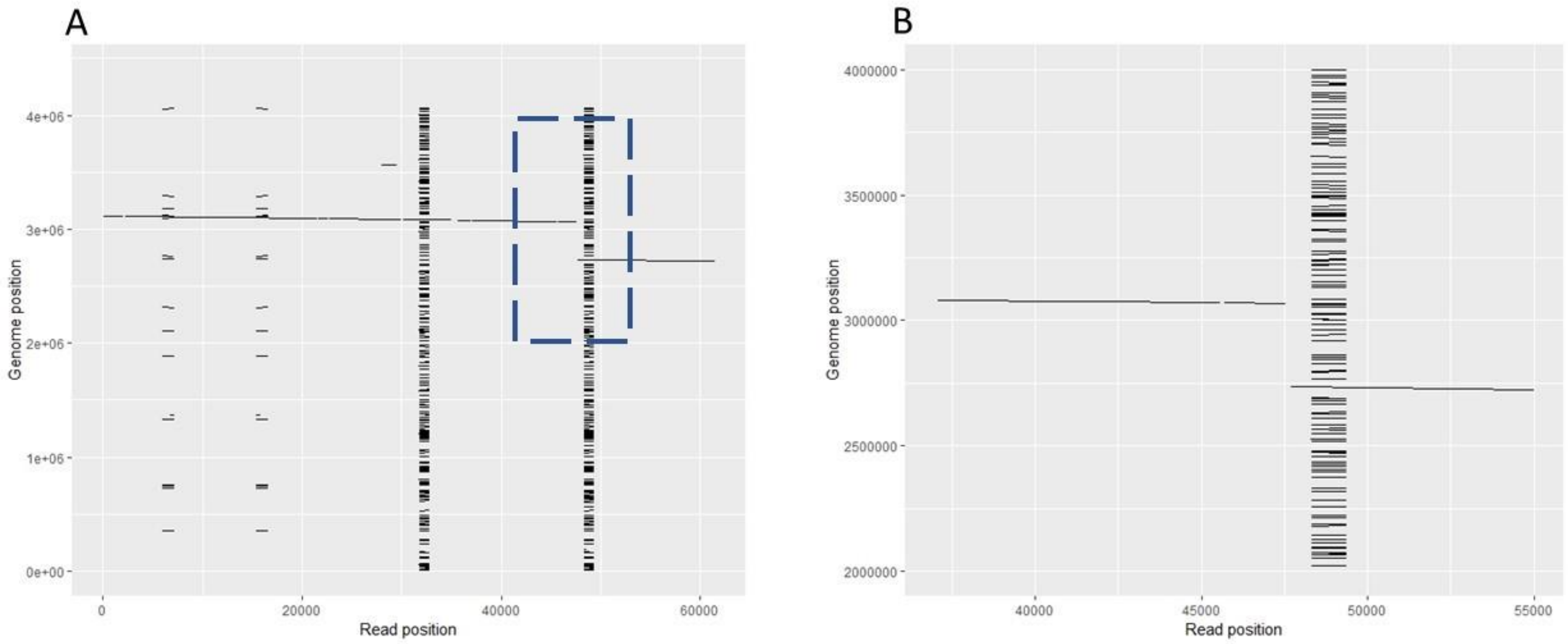


Figure 4.10: A read which contained a gap near, but not adjacent to a junction sequence. The full read is shown in A and shown in closer detail in B. This is likely a chimera. Short areas of the read (X axis) which are present multiple times in the consensus genome

(Y axis) appear as vertical ‘columns’ of hits whereas long regions of the read which map to single areas of the consensus form horizontal lines.

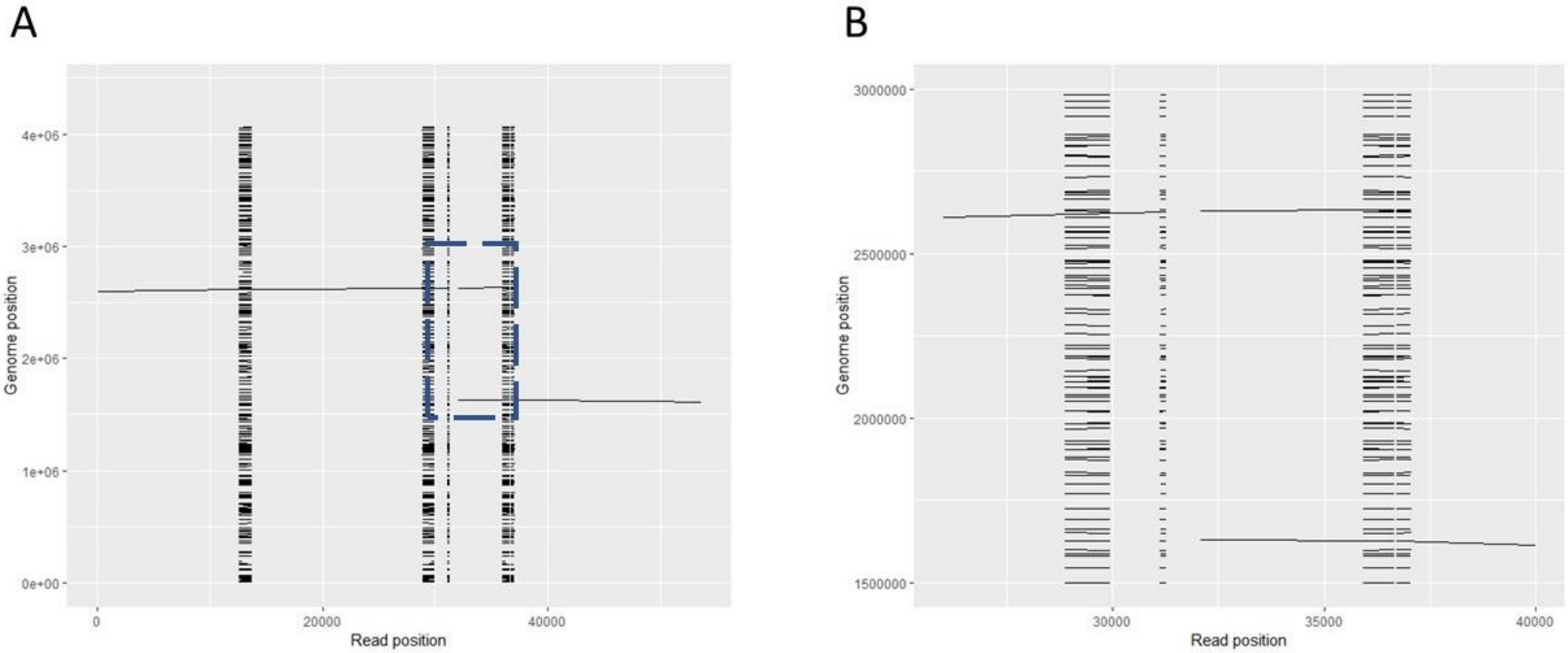


Figure 4.11: A read that contained a gap following a repeat element, in this case a 3kb sequence which has two copies in distant parts of the genome. The full read is shown in A and shown in closer detail in B. This read is likely to be a chimera. Short areas of the read (X axis) which are present multiple times in the consensus genome (Y axis) appear as vertical ‘columns’ of hits whereas long regions of the read which map to single areas of the consensus form horizontal lines.

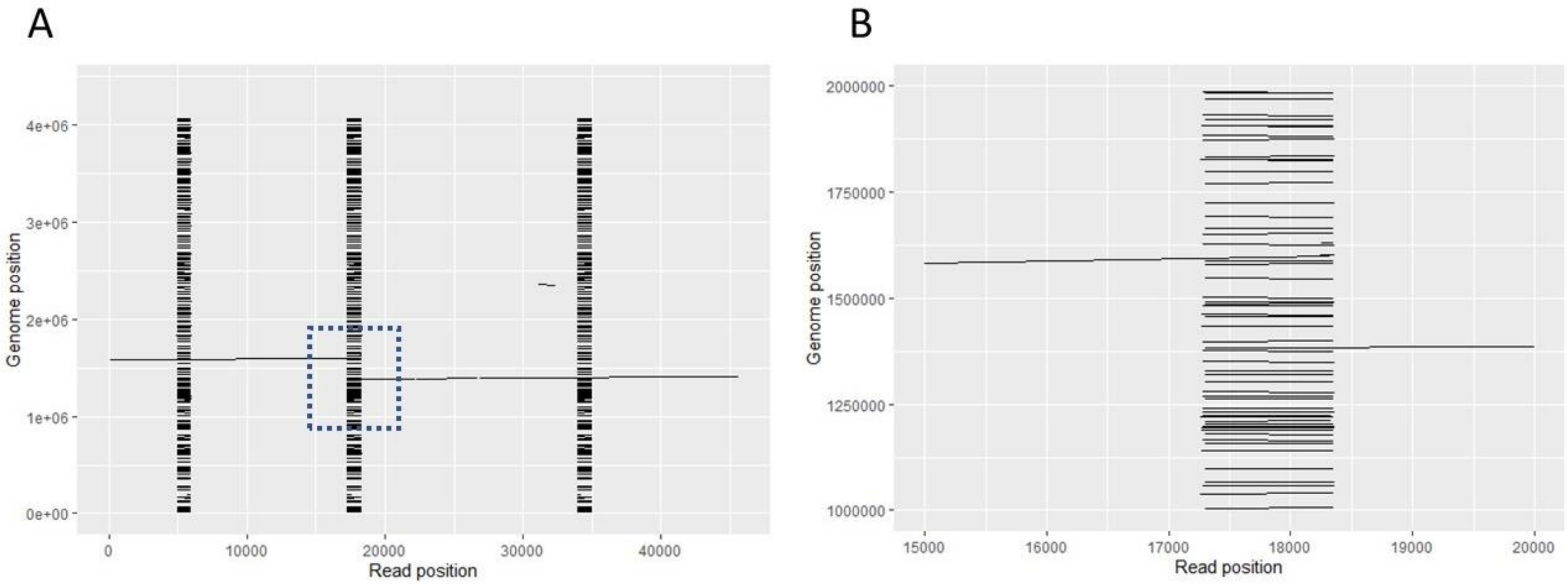


Figure 4.12: An exemplary read which did not contain any gaps in the read and consensus sequence alignment. The full read is shown in A and shown in closer detail in B. This read is likely a true SV. Short areas of the read (X axis) which are present multiple times in the consensus genome (Y axis) appear as vertical ‘columns’ of hits whereas long regions of the read which map to single areas of the consensus form horizontal lines.

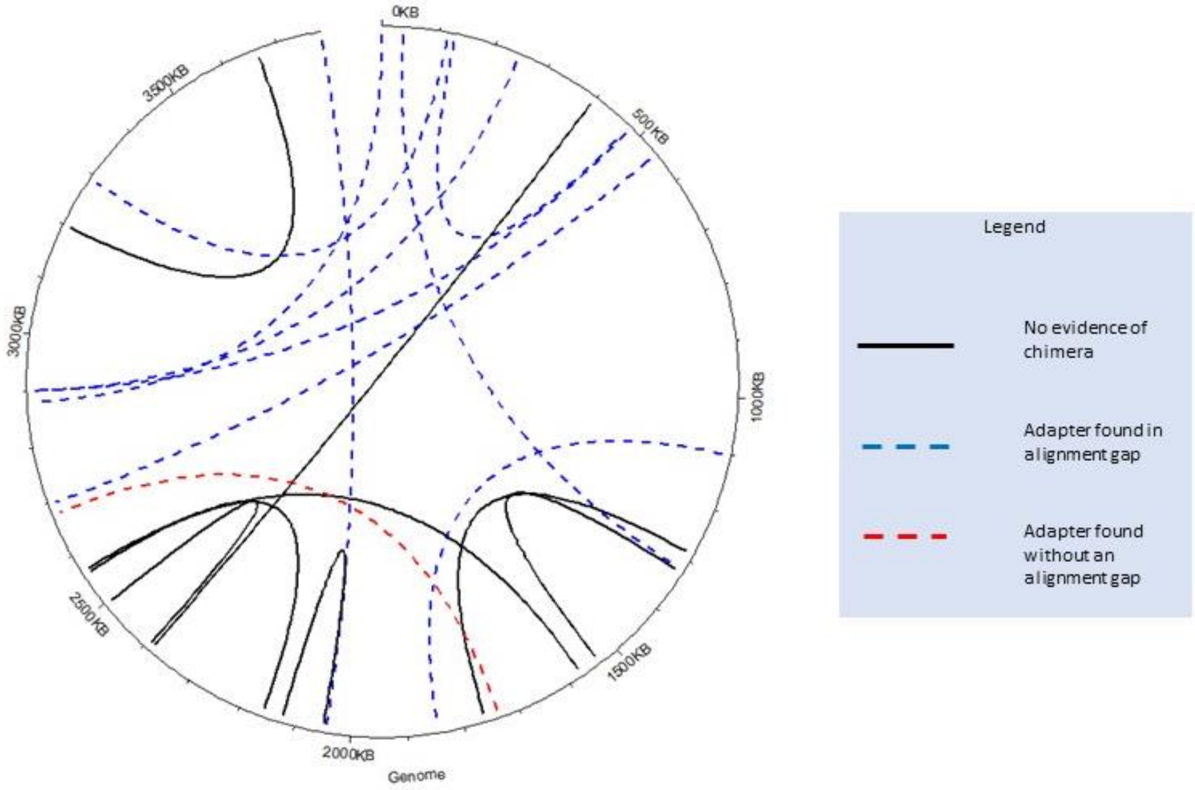


Figure 4.13: A Circos plot of the three main types of reads identified in this analysis.

4.3.3.4. Inexact string matching on error prone data is difficult.

As DNA is composed of a 4-letter alphabet, two random sequences can match with approximately 50% homology (Figure 4.14). Therefore, identifying sections of DNA that are approximately 60-80% similar to their true sequence is a technical challenge.

To demonstrate how random sequences may share high homology, simulated and true positive DNA segments were aligned to the adapter (see methods) and the percentage homology plotted (Figure 4.14) . The results show that in order to recognise 99% of adapter sequences in this dataset it would be necessary to allow homology matching down to 52% homology. This has a true positive rate of 99.1%, a false positive rate of 21.6% and a false negative rate of just 1%. However, if only 90% of the chimeras were to be excluded then adapter homology threshold could be set to 78%, leading to a true positive rate of 91%, a false positive rate of 0 and a false negative rate of 8.7%. The impact of this on the UK54 dataset is worked through in Table 4.1, with an estimation that chimeras are formed at a rate of 2% (261).

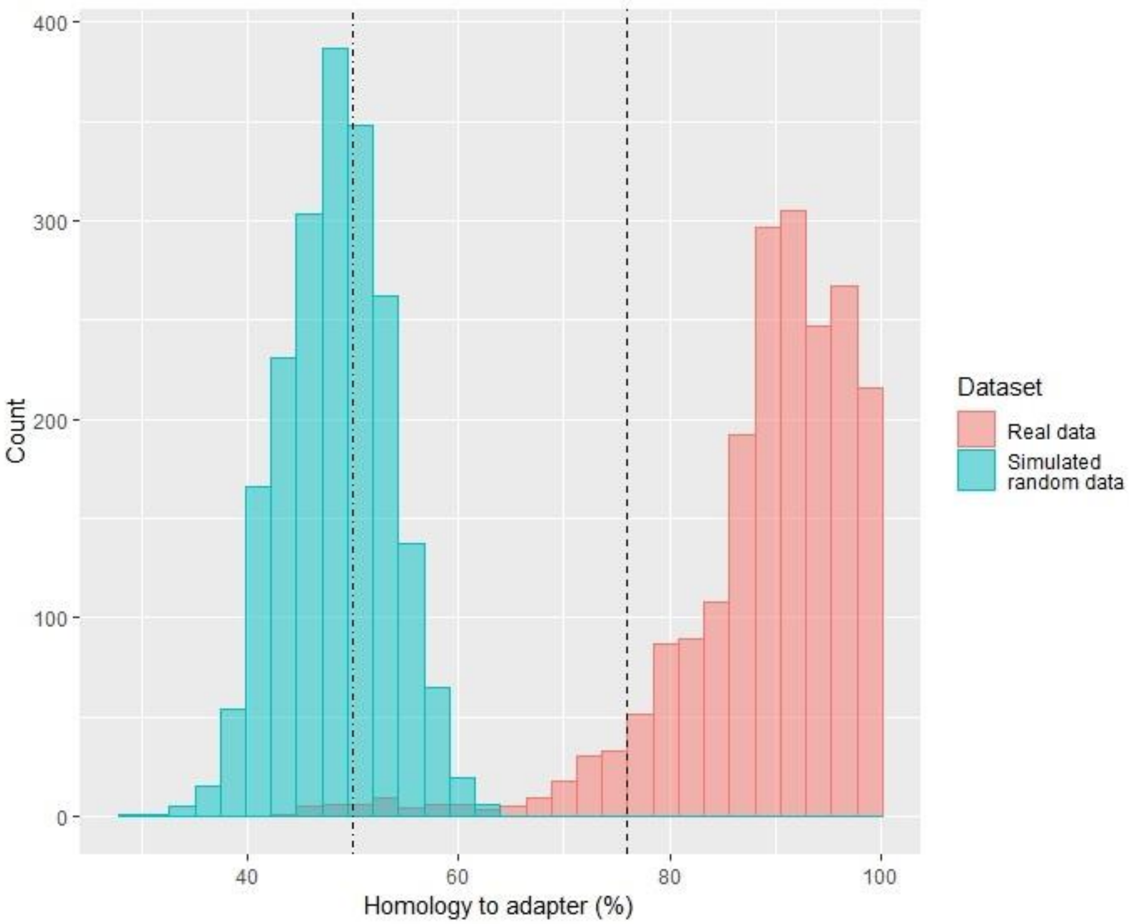


Figure 4.14: A histogram of adapter homology search results (X axis) in real data containing adapters and simulated data which did not contain adapters. The threshold that would remove 90% of chimeras (76% homology) is indicated with a dashed line and the threshold that would remove 99% of chimeras is indicated with a dash-dot line.

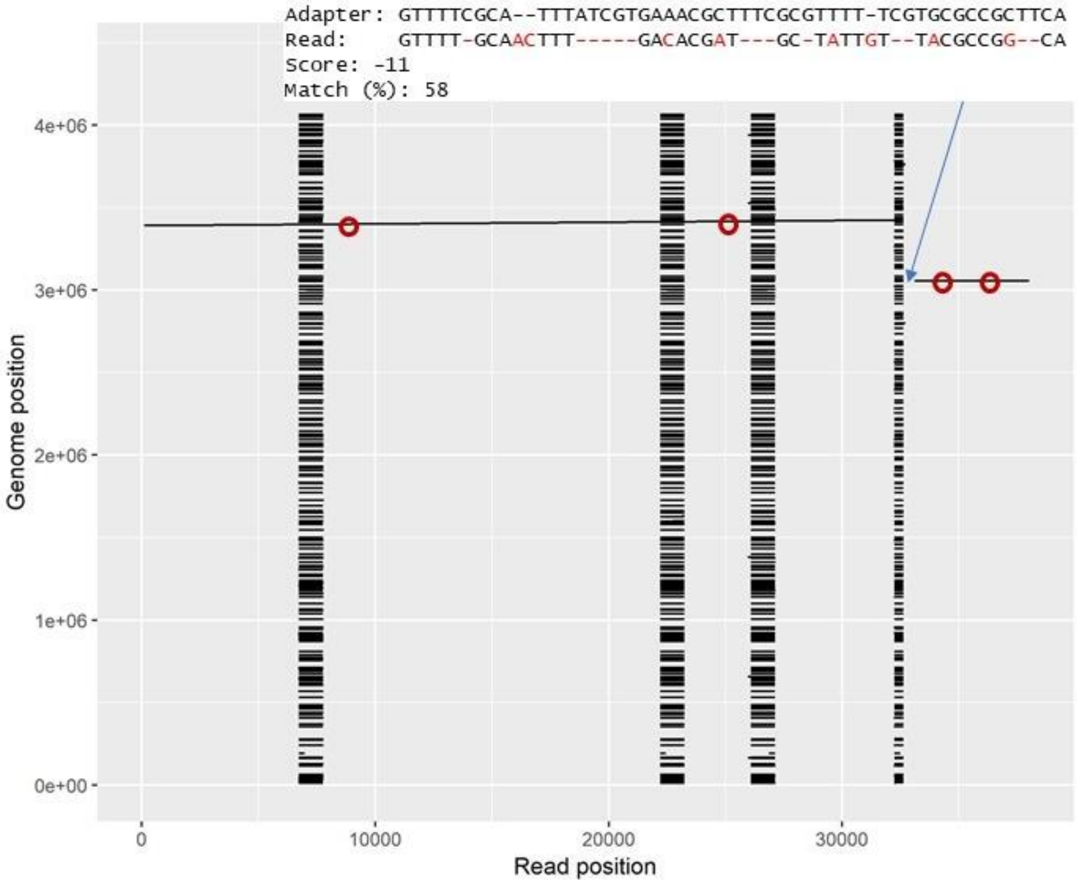
Table 4.1: A table of worked examples of chimera matching based on the UK54 clone 4 datasets.

Genome	Reads generated	Example chimera rate	Adapter exclusion threshold	Adapter homology threshold	True positive	False positives	False negatives
UK54 clone 4	700,000	2% (14,000 chimeric reads)	99%	56%	13874	3031	210
			90%	75%	12803	0	1225

Conducting a simulated analysis on adapter sequence matching demonstrated how easily false positives are generated when inexact string matching is undertaken on error prone reads. It was found that even removing chimeras with a permissive homology match would result in 200 chimeric reads being produced in the UK54 dataset. I sought to find reads with adjacent DNA sequences that mapped to distant parts of the consensus sequence, a process that would also highlight many chimeras. Therefore, in order to increase the confidence in my dataset, I needed a more stringent method to remove chimeric reads.

4.3.3.5. Searching for adapters within alignment gaps effectively identifies chimeras

There were 9 reads that were selected by the Circos analysis as having junction sequences within 1.5kb of a repeat element. A broad search was conducted to find reads close to repeat elements but only the reads which have junctions fully adjacent to repeats are likely true SVs, thus these 9 reads were excluded. It was hypothesised that these reads were chimeras. In support of this, 7 out of these 9 reads contained an adapter sequence at the sequence gap, one of which is shown in Figure 4.15. This indicated they were chimeras rather than true SVs caused by non-homologous recombination or homologous recombination between small repeats.



```
Adapter: GTTTTCGCA--TTTATCGTGAAACGCTTTCGCGTTTT-TCGTGCGCCGCTTCA
Read:    GTTTT-GCAACTTT-----GACACGAT---GC-TATTGT--TACGCCGG--CA
Score: -11
Match (%): 58
```

Figure 4.15: A read with a gap in sequence homology between the read(X axis) and the consensus sequence (Y axis). An adapter was found in the gap with 58% homology to the true adapter sequence. Additional matches to the adapter sequence are marked with a red circle. The alignment of the adapter sequence to the found sequence is shown and differences are highlighted red and gaps with a dash.

Interestingly, despite the adapter being only 50bp long, the gaps were often much larger. This may indicate how the chimeras were formed. It may be the chimeras formed by two DNA sequences quickly passing through the pore cause large gaps in sequence (Figure 4.11) as the pause between reads gets interpreted as random sequence. Small gaps in sequence may be caused by two reads physically joined together in the adapter attachment step (Figures 4.10 and 4.15). Excluding these 9 reads from the 29 reads left 20 reads which had junctions directly adjacent to repeat elements.

Table 4.2: The categories of reads representing putative SVs that were analysed for chimeras.

Read category	Number of reads	Number with detectable adapter	Number without adapter
Junction not adjacent	9	7	2
Adjacent with gap	11	9	2
Adjacent without gap	9	1	8

As expected, gaps in sequence alignment were often due to adapter sequences being present (Table 4.2). This provided evidence that in order to exclude chimeras with the highest stringency, an additional search rule should be added: reads which contained gaps directly adjacent to junctions in their alignment to the consensus sequence should be excluded.

In summary these are the rules which were adhered to in order to highlight reads as true SVs:Contains DNA sequences which are proximal on the read but at least >=30kb apart on the consensus sequence, forming a ‘junction sequence’; The junction sequence must be directly in or adjacent to a repeat region; There must be no gaps in homology between the read and the consensus sequence adjacent to the junction sequence. Following these search parameters left 8 reads which passed all tests and had no evidence that they were chimeras. I hypothesise that these are structural variants.

4.3.3.6. High confidence reads and SV type

The final step in this analysis was to determine which SV type gave rise to the high confidence reads. A rearrangement can be distinguished from a deletion or CNV by analysing the orientation of the pre- and post-junction sequences. If the two sequences are in an opposing orientation then an inversion has been detected, whilst if the two sequences are in the same orientation then a deletion or CNV has occurred (Figure 4.16). This was exploited to label each SV as either a deletion/CNV or rearrangement (Figure 4.17).

Using this method, it could be demonstrated that putative SVs in single reads that were not present in the consensus sequence could be identified. In this sample there were only 8 such reads. Whilst it appears that there was a slight bias of reads towards the terminus, that was not repeated in the following analysis of UK76. What this data does demonstrate, however, is that the *B. pertussis* genome is highly plastic and this can be observed in as little as two passages using Nanopore reads.

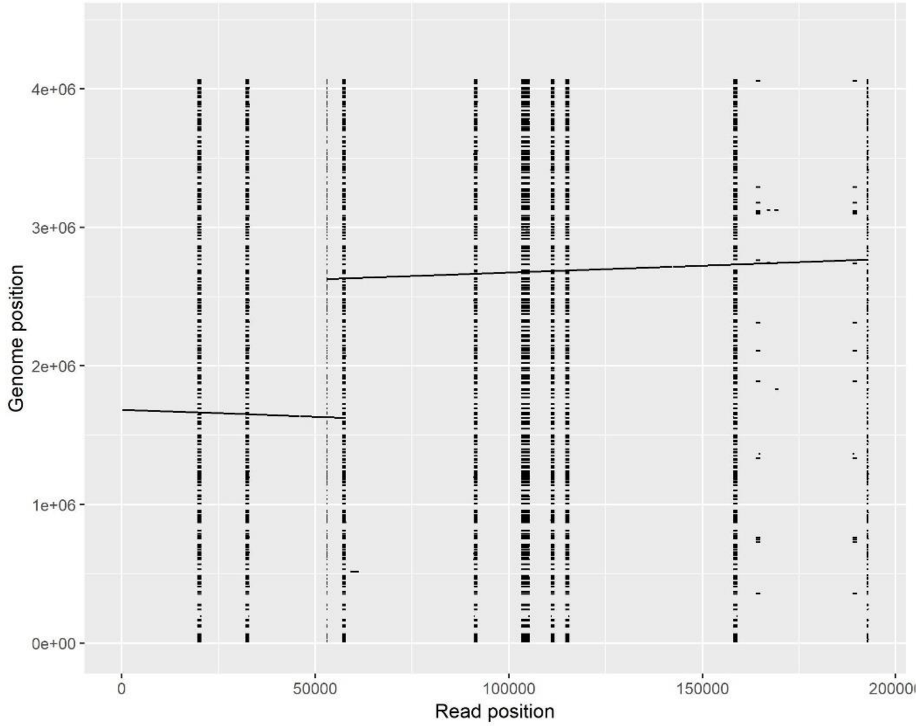


Figure 4.16: A read which had no evidence of being a chimera and appears to be a genome rearrangement. The pre junction DNA from 0-53kb on the read (X axis) is in the inverse orientation to the genome (Y axis) as it matches in a negative orientation (appearing as a negatively sloped line)- starting at 1.63Mb and ending at 1.62Mb in the consensus sequence. The DNA after the junction, however, is positively orientated with the consensus sequence (appearing as a positively sloped line) - the read sequence from 53kb-200kb matches from 2.62Mb to 2.76Mb on the consensus sequence.

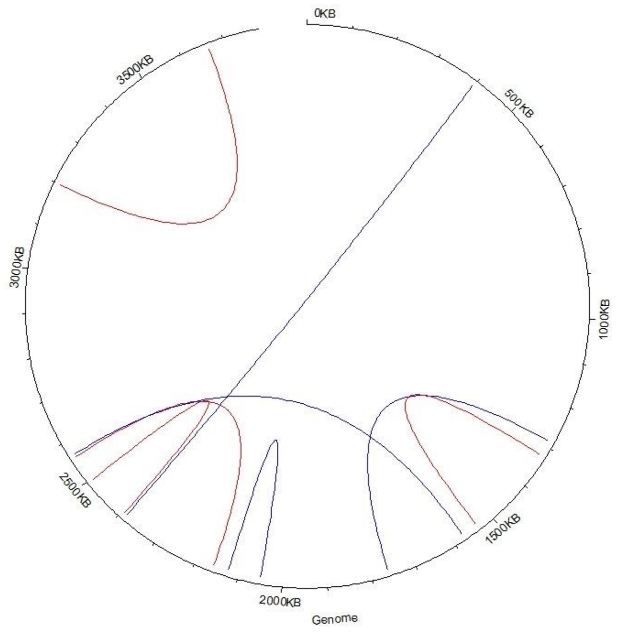


Figure 4.17: A Circos plot showing the 8 reads from UK54 with no evidence of being a chimera. Blue lines indicate reads which arose from CNVs or deletions and red lines indicate reads which arose from rearrangements.

4.3.3.7. UK76 contains at least 6 unique junction sequences

Having established that structural variations could be detected in single Nanopore reads in the UK54 dataset; it was important to establish this phenomenon in another sequence dataset. The strain UK76 presented one of the largest CNVs at >300kb (with a tandem length of >600kb) and had a predicted copy number of this locus of 1.3. This strain was sequenced as part of a generic investigation into CNVs in *B. pertussis*. When ultra-long nanopore sequencing was undertaken it was unsurprising that no reads were found to span the whole tandem repeat region.

Undertaking the same analysis on this sample provided insight into CNV formation in *B. pertussis*. Of the 400k reads generated, 136 reads were found to contain DNA sequences that were proximal on the read but distant on the consensus sequence of which 95 were found to have junction sequences in repeat elements. Of these 95, 67 had no adapter sequence or homology gaps at the junction -remarkably higher than 8 reads passing the same tests in the UK54 sample. This different abundance of these reads was not proportional to how many total reads there were in each sample, given that both UK54 clone 4 and the original UK54 sample both contained approximately 700k reads.

It was found that UK76 had multiple different CNV junctions present in the sample (Figure 4.18). Whilst previous experiments demonstrated that heterogeneity of copy number was possible, these results indicated that heterogeneity of the sequence composition of CNVs was possible within a single sample. Comparing this spectrum of start/end positions to those found within the global cohort of isolates in Chapter 3 resulted in considerable overlap. For example, 20 reads had the junction of 1.315Mb and 1.709Mb (Figure 4.19) matched the predicted CNV for UK76 in Chapter 3, an estimate based on read depth of Illumina data. There were many reads which had subtle variations of this, however, with junctions differing by only a few thousand bases. The second most popular configuration was 1.385Mb-1.655Mb where at least 6 reads had the breakpoint, again with subtle variations (Figure 4.20). There was a small number of reads that appeared to have their own unique junctions. Whilst it was not feasible to manually check each break point, every new break point configuration that was manually analysed indicated that the Circos analysis was true (Figure 4.18). There were at least 6 junctions present in this sample: 2 popular (Figures 4.19 and 4.20) and 2 rare/unique (Figures 4.21 and 4.22).

Of these 6 studied junctions, 5 were found previously in the large-scale analysis in Chapter 3. This may mean that certain breakpoints are particularly prone to SV and therefore occur repeatedly both on long timescales, such as in the large dataset in Chapter 3 and on short scales, such as described here. Whilst the gene sets of each of the 6 CNV identified here do differ, and this includes genes with functions involved in homeostasis (pdxK,BP1320) and glycogen biogenesis (*glgB*, *glgX*), such subtle differences are overshadowed by the size of the CNVs, which involve approximately 200 genes.

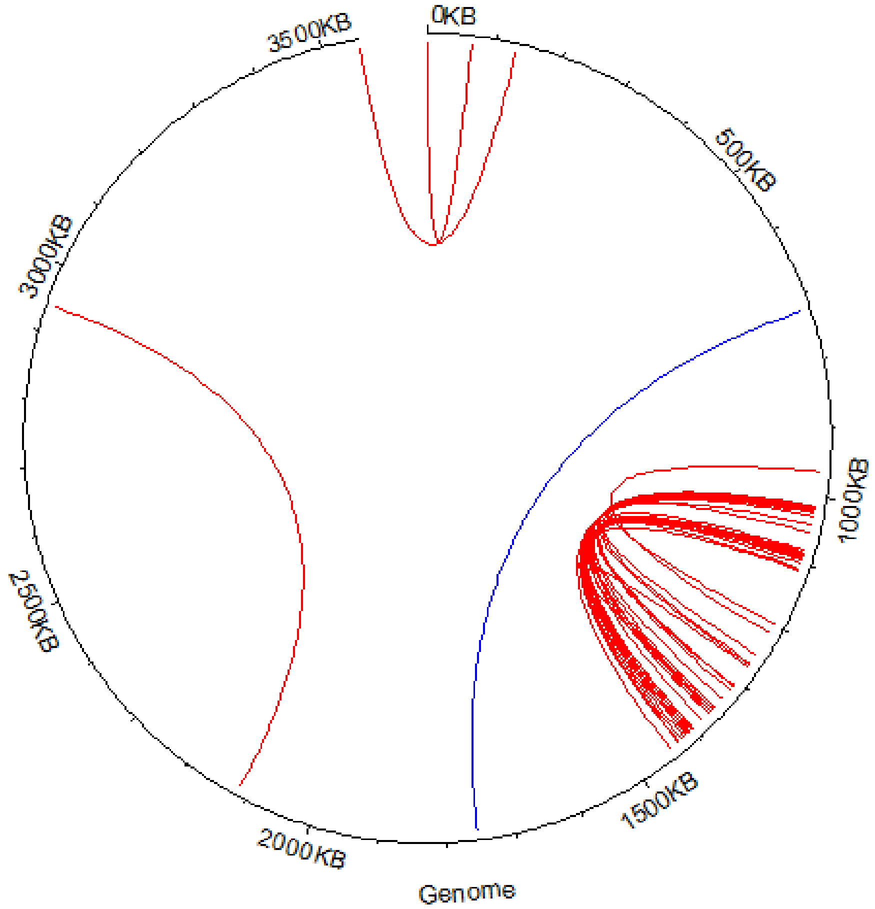


Figure 4.18: A Circos plot showing the 67 reads with no evidence of being a chimera in UK76. Red lines indicate reads which arose from CNVs or deletions and blue lines indicate reads which arose rearrangements. Many reads with slight variation in junctions were found between 0.9Mb and 1.5Mb.

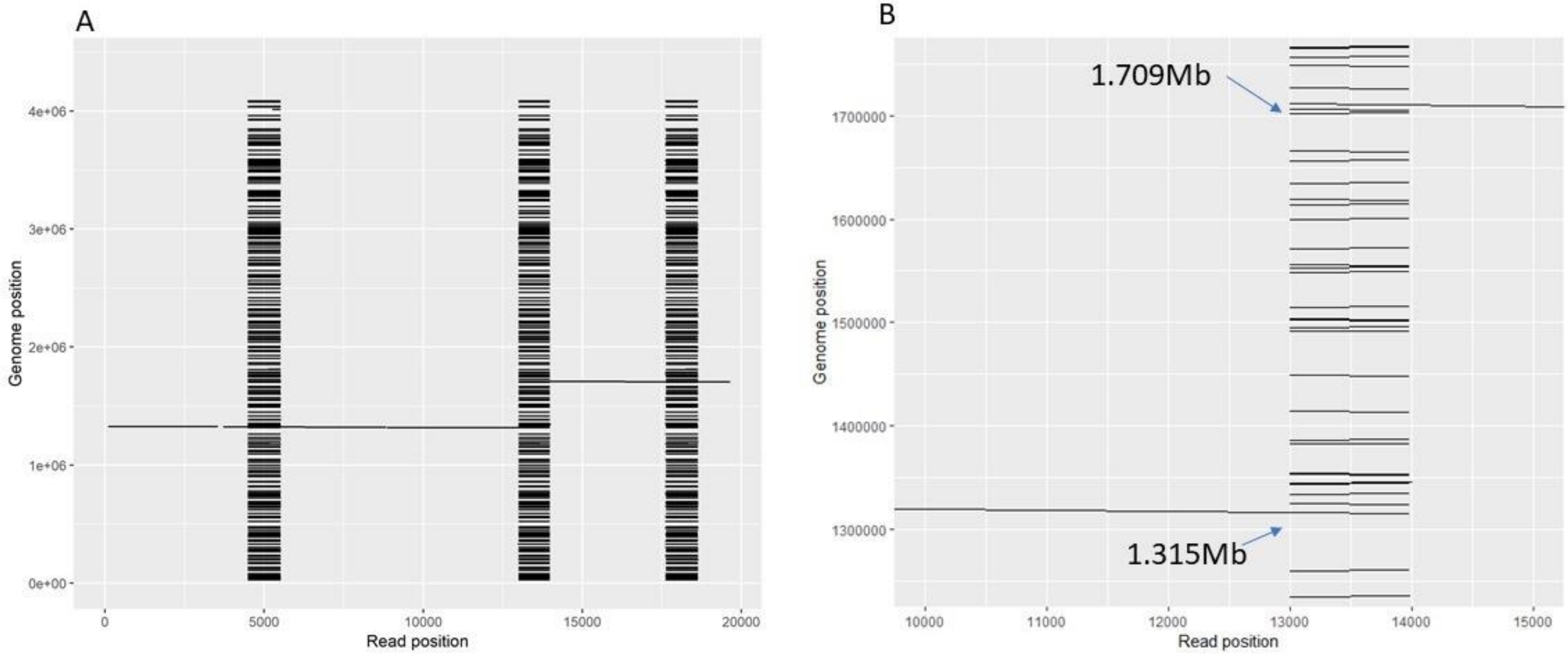


Figure 4.19: An alignment of a read (X axis) to the consensus genome (Y axis) with the most frequent junction (A). The junction is enlarged in (B) and corresponds to a CNV between 1.315Mb-1.709Mb.

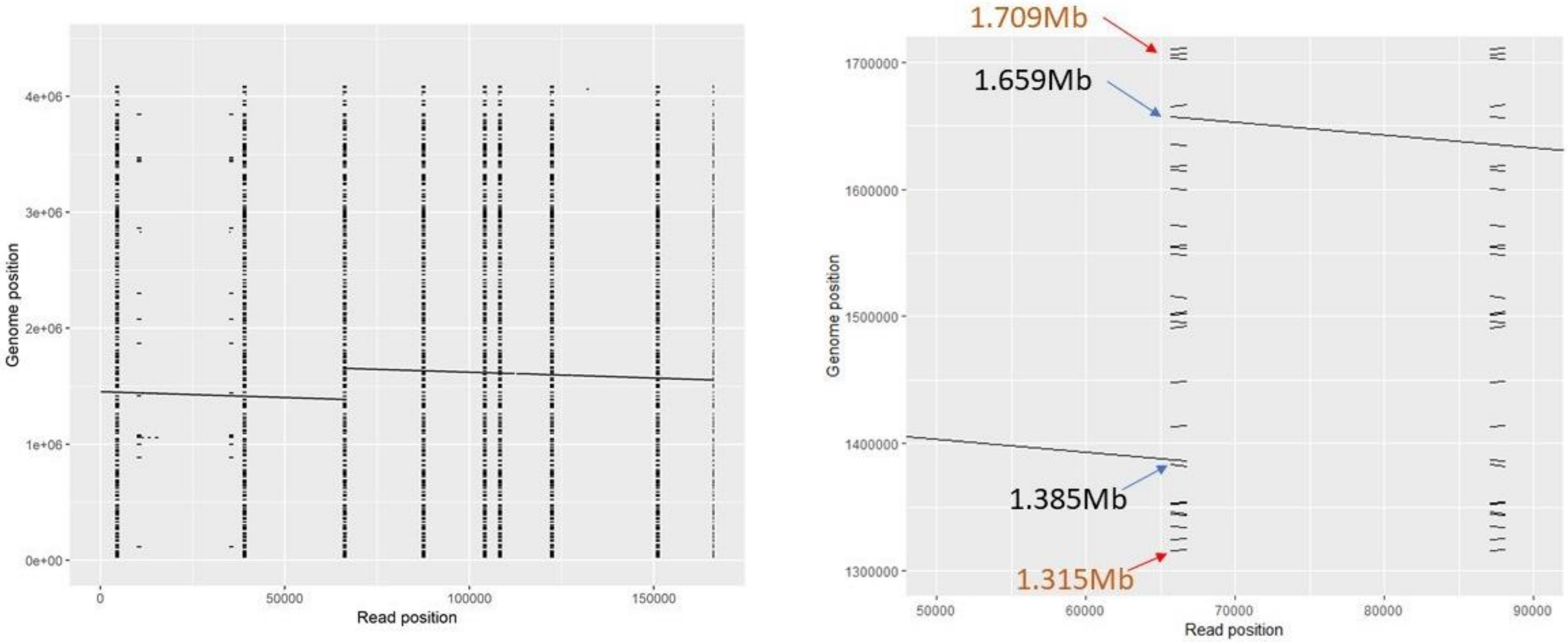


Figure 4.20: An alignment of a read (X axis) to the consensus genome (Y axis) with an alternative breakpoint (A). The junction is enlarged in (B). The breakpoint was 1.385Mb-1.655Mb. In red are the breakpoints displayed in (Figure 4.19).

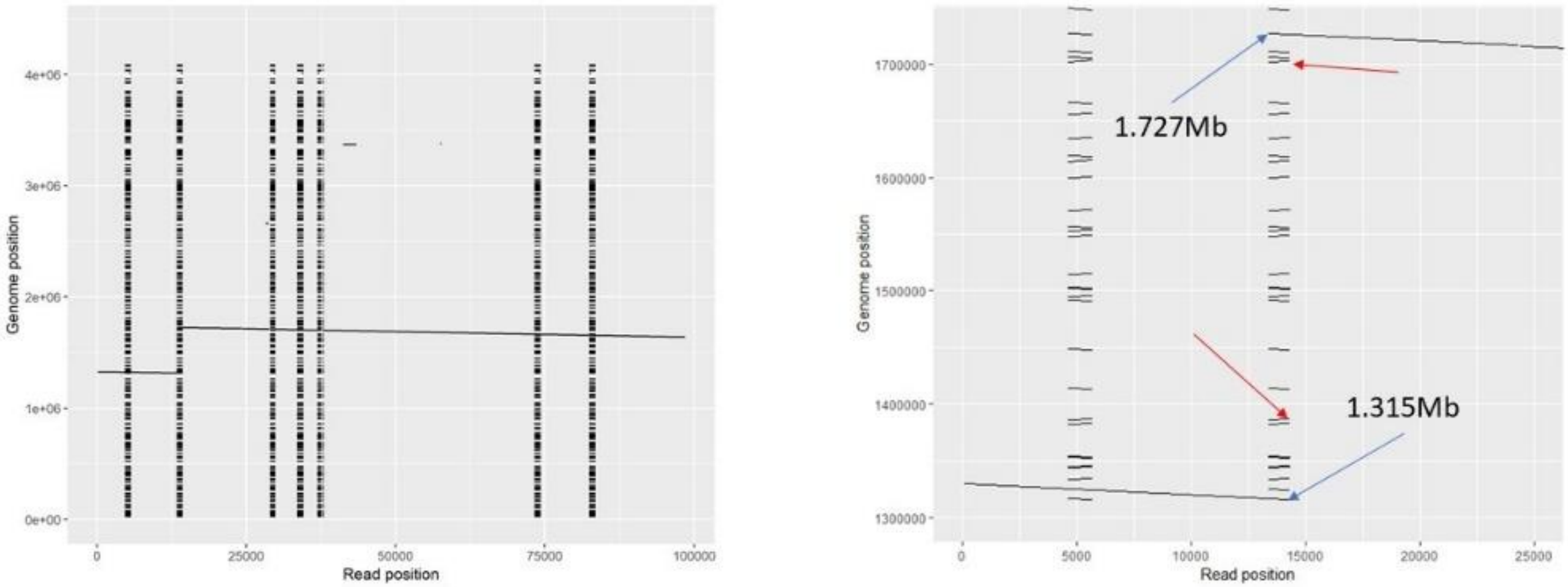


Figure 4.21: An alignment between a read (X axis) and the consensus genome (Y axis) with an alternative breakpoint (A). The junction is enlarged in (B). The breakpoint was 1.315Mb-1.172Mb. In red are the breakpoints found in previous analyses but not in this read (Figure 4.19 & Figure 4.20).

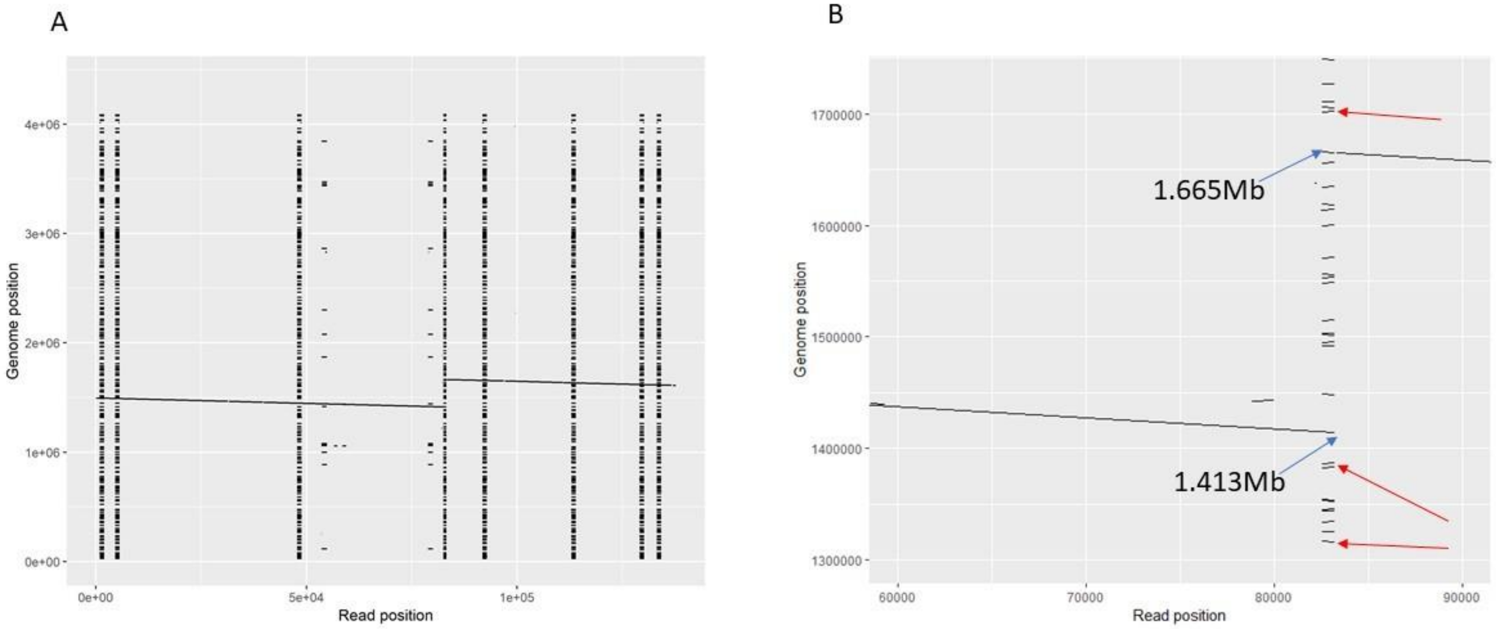


Figure 4.22: An alignment between a read (X axis) and the consensus genome (Y axis) with an alternative breakpoint (A). The junction is enlarged in (B). The breakpoint was 1.413Mb-1.665Mb. In red is the breakpoints found in previous analyses but not in this read (Figures 4.18-4.20).

This analysis of UK76 therefore was a valuable experience. It was clear that in this sample a high amount of genome plasticity had been observed. I had, however, hypothesised that there would be frequent de-novo SVs all around the genome. This was not true in either sample, although in UK54 not many reads with de-novo CNVs were observed. Like other *B. pertussis* isolates from the 2012 epidemic, it was unclear if these samples had been isolated from a single colony on a plate or the observed diversity was generated in vivo and had been captured by taking a sweep of the plate. It cannot therefore be concluded whether this observed diversity of breakpoint occurred during in vitro passage or in vivo.

4.3.3.8. Analysis of CNV containing reads

I detected 8 reads in UK54 and 4 reads in UK76 that contained strong evidence of CNVs occurring in loci unrelated to the known CNV in each genome. Of these, 7 were CNVs/deletions and 5 were inversions. The majority of these reads bore no resemblance to each other and appeared to be ‘randomly’ distributed throughout the genome- not at the 11 hotspot loci previously established. One read did overlap Network 1, although this is not enough evidence to suggest that this locus is a mutational hotspot, rather than a hotspot generated by selection. There was slightly elevated frequency of CNVs between 2Mb and 3Mb in both genomes, however, although the dataset is too small to definitively suggest this. The general lack of overlap with the hotspots found in highly interesting considering that 93% of the CNVs found in the dataset in Chapter 3 were found at hotspots. This supports the argument that the CNVs found in that dataset were not products of the populations adapting to invitro environments, but adaptations to other environments, given the disparity in gene sets that were observed. These SVs were too long to have their genes and putative functions listed; however, the functional categories were described in Figure 4.23 and were expressed in terms of fold enrichment compared to the whole genome.

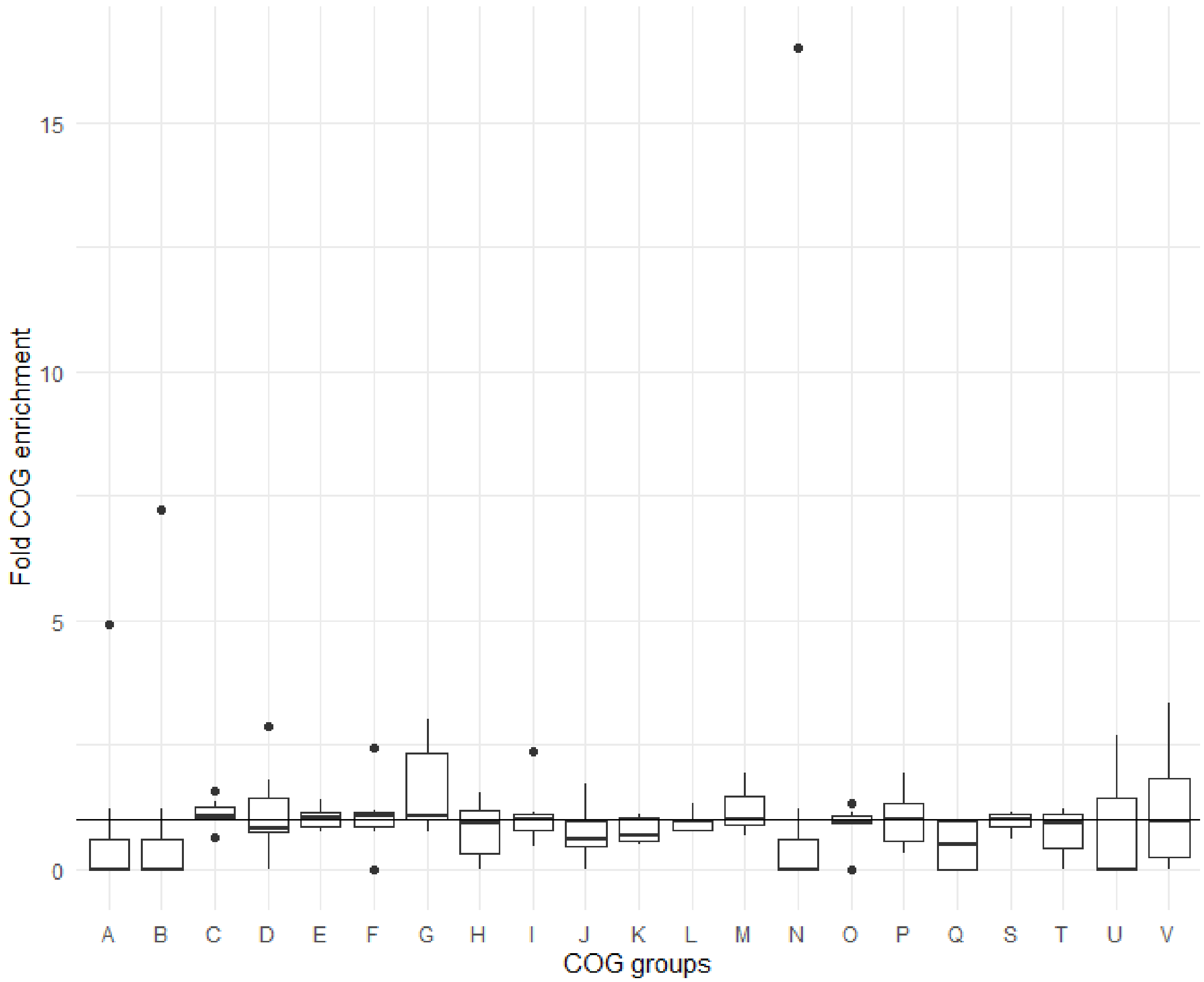


Figure 4.23 Seven putative deletions/CNVs were analysed for enrichment of genes belonging to the 21 COGS (X axis) in comparison to the genome as a whole. This was expressed as a fold enrichment (Y axis). It can be seen that most boxplots have a median of 1(noted as a horizontal line), indicating no functional enrichment on average. Values above 1 indicate enrichment and values below 1 indicate depletion.

Table 4.3 COG categories and their full designation.

A	RNA processing and modification
B	Chromatin Structure and dynamics
C	Energy production and conversion
D	Cell cycle control and mitosis
E	Amino Acid metabolism and transport
F	Nucleotide metabolism and transport
G	Carbohydrate metabolism and transport
H	Coenzyme metabolism
I	Lipid metabolism
J	Translation
K	Transcription
L	Replication and repair

M	Cell wall/membrane/envelop biogenesis
N	Cell motility
O	Post-translational modification, protein turnover, chaperone functions
P	Inorganic ion transport and metabolism
Q	Secondary Structure
T	Signal Transduction
U	Intracellular trafficking and secretion
Y	Nuclear structure
Z	Cytoskeleton
R	General Functional Prediction only
S	Function Unknown

Analysing the six reads using COGs revealed that on average these reads were not enriched or depleted for most functions. Categories A(RNA processing and modification) ,B(Chromatin Structure and dynamics) ,N (Cell motility) and U (Intracellular trafficking and secretion) were highly depleted, having median fold enrichments of 0. Whilst both COGs A and B were depleted on average, each had one CNV which was highly enriched for these categories. This may have been because the two that were enriched were CNVs and those which were depleted were deletions, given that these categories of genes are likely to be essential for growth. The putative CNV with the highest enrichment was the one CNV that overlapped with network 1 and was highly enriched for motility genes. Category G (Carbohydrate metabolism and transport) was on average not enriched, but had an upper quartile significantly above 1. This may have indicated that CNVs were enriched for these genes whilst deletions were not. Such an arrangement could be due to in vitro selection for carbohydrate metabolism. Future research should aim to establish if this spectrum of enrichment/depletion is consistent between samples.

4.3.4. Mining assembly graphs for supporting data

It has been previously reported that long read assemblers failed to assemble *B. pertussis* genomes predicted to contain long tandem CNVs (105, 130, 178). Assembly is not a single stage process, however, and there exists many intermediate steps that generate data. I investigated several intermediary graphs in the assembly process of Unicycler to investigate how this data corroborates my Circos analysis and to generally explore the assembly process for any potential utility.

Unicycler relies on the Miniasm (262) assembler which uses de-Bruijn graphs constructed with nodes and edges which are made of reads and overlaps, respectively. The process of de-novo genome assembly using de-Bruijn graphs underpins most assembly tools and at its very core, is as simple as finding overlaps between reads (or K-mers in many other tools) and then determining which overlaps have the best support. During the assembly process the graph gets more refined as low-support relationships and redundant reads/relationships are pruned until only the most high-confidence nodes and edges remain. Unicycler, almost uniquely, allows viewing of these intermediate assembly graphs and I used these to understand how junction sequences were being interpreted throughout the assembly process.

As Ring et al noted in their assemblies, use of Nanopore reads with Illumina reads in a hybrid assembly resulted in a single contig but with the probable CNVs collapsed (178). Conversely, assembling just long read data also gives a poor assembly although with different flaws. The 'nanopore-only' assemblies of the Ring et al study were replicated here, but only in Unicycler (263) and with ultra-long nanopore data rather than standard length. In this experiment, not only were the final assemblies studied, but also intermediate graphs, which may be informative on why the genome could not be closed.

The genome of UK48 had a predicted CNV size of 170kb and this strain had been sequenced with ultra-long Nanopore sequencing in an attempt to resolve the CNV. Like other samples, however, it was not possible to resolve the CNV. Assembling long read data for UK48 with Unicycler gave a fragmented assembly of 12 contigs and a total assembly size of 4.4Mb- 300kb longer than would be expected if the CNV was not resolved (Figure 4.24). BLASTing the putative 170kb duplicated locus, in addition to 90kb of flanking sequence either side of it (300kb total), against this graph revealed that the CNV locus had not been resolved as a tandem array nor was it present in two full copies (264). The data showed the locus was spread over multiple contigs- one of which contained 40% of the CNV locus and was noted in the genome assembly graph as having a copy number of 1.52. In total the CNV locus was present at 130% of its original length. It therefore appeared that the CNV was partly resolved, although it is unclear which parts had been collapsed to a single copy. It also was not clear why the assembly was larger than it should be, even accounting for the CNV being partly resolved.

Using the same 'BLAST-to-graph' methodology (264), a contig was identified which had a juxtaposed colour scheme of green next to red (Figure 4.23). This indicated that this contig contained an order of DNA that was different to the reference genome. On further examination it was clear that this was a junction sequence between the tandem duplication copies - direct evidence a CNV is present in this sequence. The ability to resolve junction sequences from sequence data is a clear advantage of long read Nanopore data.

In order to further understand the assembly process and data which gave rise to such an assembly, the UK76 string graph was examined. In this graph, the >500k reads have been reduced to a less redundant set, although with much redundancy still present. The duplicated region was shown as a complex structure with reads that were clearly junctions and a ‘bubble’ that contained the duplicated sequence. Examining this graph revealed the complexities of long read assembly for *B. pertussis*. Such junction sequences can be used, in combination with other sources of evidence, to automate the construction of hypothesised CNVs in the future. This could provide a genome sequence which is of a quality intermediate between fully resolved and partially resolved. This task, however, was outside of the scope of the current work and this process would struggle to resolve very complex CNVs such as those outlined in Chapter 3.

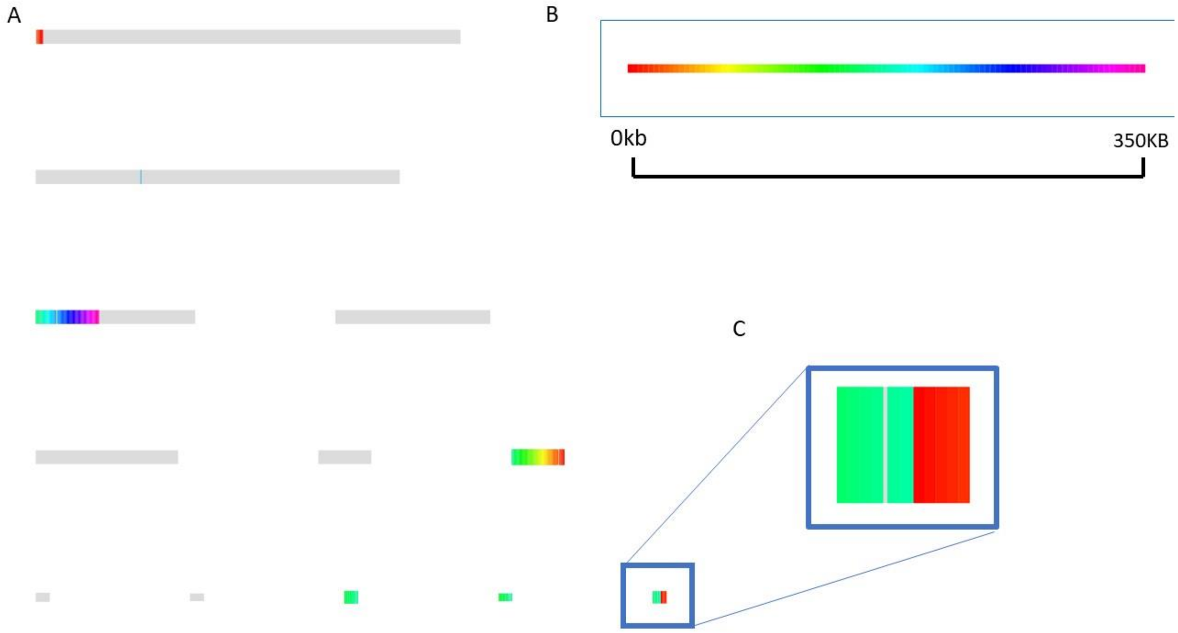


Figure 4.24: UK48 was assembled into 12 contigs (A) and the putative CNV sequence, with flanking regions, was searched against the assembly (B). The rainbow colour spectrum (red, orange, yellow, green, blue, indigo, violet) is used to show matching sequences in the subject (the assembly) to the query (the CNV). Panel C shows a closer examination of a contig: green sequences are adjacent to red sequences, despite these sequences being separated by 170kb on the query sequence (B). The contig therefore appears to be a junction sequence as these locations correspond to increased coverage in the UK48 short read dataset.

I then investigated how unicycler handled the UK76 assembly which had many different junction sequences. Assembling UK76 with Unicycler resulted in a single contig which did not have the CNV resolved, something also observed by Ring et al (178). I searched for the duplicated locus from UK76 (with single copy flanks) in an intermediate graph. I could identify many sequences that were SV junctions, visible because of their juxtaposed colour scheme (Figure 4.24). Further investigation into these reads corroborated the work from the Circos plot: there were reads showing different breakpoints thus indicating multiple duplication events co-occurring within the population. As the process of assembly aims to reduce the readset to increasingly less redundant state and that this was a graph mid-way through this process, it can be presumed that these nodes had multiple reads which covered them, meaning they were unlikely to be chimeric sequences. Unicycler was discarding these reads, likely because of their divergent gene content. I could therefore show in two independent ways, analysing the same dataset, this novel feature of the UK76 sample.

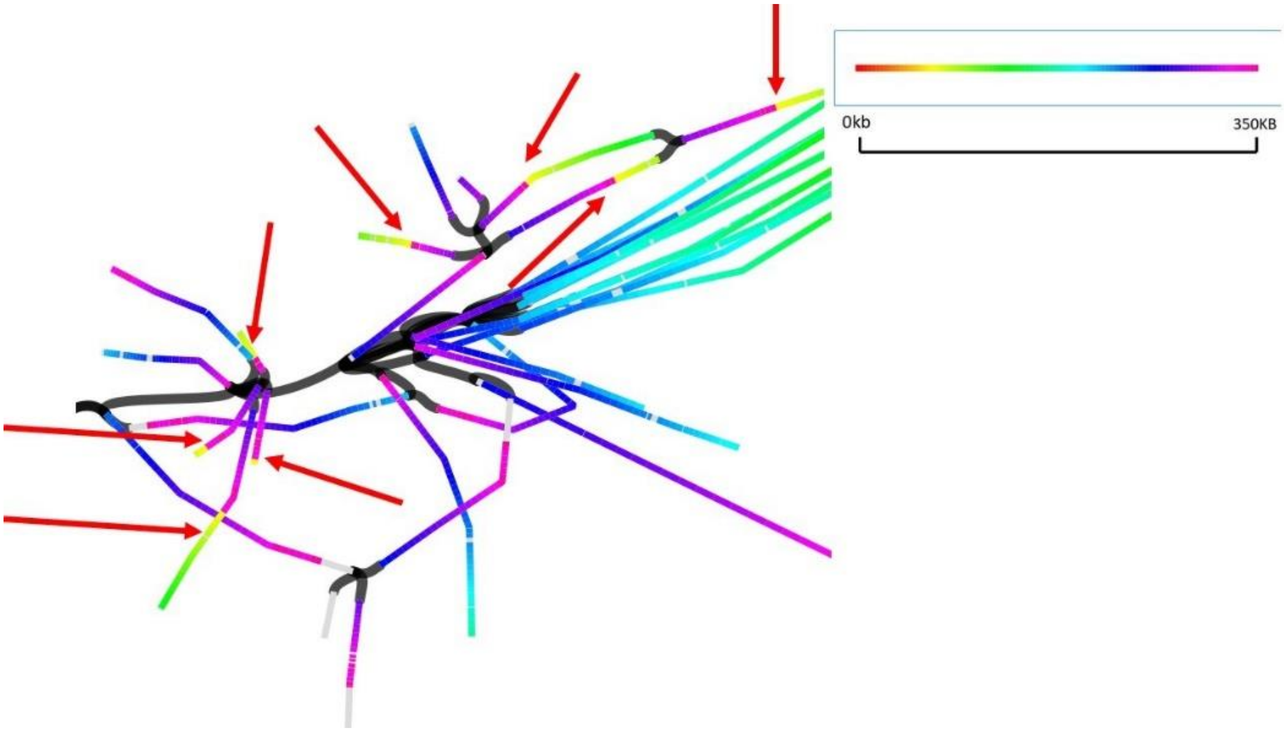


Figure 4.25: An early stage of the assembly was examined further. Reads (coloured lines) with homology are connected by a blank linker region. The CNV region with single copy flanks (coloured bar, top right) was blasted against this graph. The rainbow colour spectrum (red, orange, yellow, green, blue, indigo, violet) is used to show matching sequences in the subject (the assembly) to the query (the CNV). Reads with an expected order of DNA have the colour spectrum in the same order as the query (blue-purple-green-yellow) whilst lines with juxtaposed colours (yellow to blue) indicate junction sequences and are indicated with arrows.

4.4. Discussion

4.4.1. Copy number estimates using short read data predicts mixed populations

The large cohort dataset was reanalysed to find that many of the copy number estimates were not integers. This was in addition to the copy number discrepancies described in Chapter 3 in the manually resolved dataset. I hypothesised that this is because cells in the population have different copy numbers of the locus and when an average is taken (as is the case for read depth-based estimates), the value is intermediate between the true copy numbers.

These non-integer estimates could be, at least partly, due to inaccuracies of CNVnator estimating the copy number via read depth. The distributions of copy number estimates in both the manually resolved and full cohort (Figure 4.2 and 4.4, respectively) datasets were skewed towards lower copy numbers. For example, in the manually resolved CNV dataset, of the 25 CNVs that were approximately copy number 2, 5 were exactly copy number 2, only one was considerably higher (copy number 2.3) and 11 were considerably lower (<=1.7). I hypothesise that such estimates were produced by a biological process as no mention of this systematic bias has been reported. Future work should include the analysis of published data to find if CNVnator systematically under-estimates copy number. Under this hypothesis, eukaryotic datasets, where genomes are generally more stable, should have a relatively even distribution of copy numbers around integers when compared to prokaryotic datasets.

4.4.2. Nanopore sequencing confirms mixed populations in UK54

4.4.2.1. Ultra-long Nanopore reads can resolve tandem arrays

The CNV estimates generated in Chapter 3 were used to screen for a CNV that was tractable using Nanopore sequencing. In this instance, this also highlighted the CNV with the highest copy number in the dataset. The UK54 strain had a CNV locus that was 16kb long. with its full tandem configuration, at a copy number of 4, predicted to be 64kb- short enough to be captured on single, long, Nanopore reads. Regular protocols produce reads with an N50 of 5-10kb which was not adequate to sequence this CNV in its tandem configuration (178). As the Nanopore platform can theoretically sequence unlimited length reads, given careful DNA and library preparation, much longer reads can be generated. To this effect, the ultra-long read protocol was followed.

Sequencing UK54 using ultra-long reads demonstrated that a mixture of copy numbers could be isolated (Figure 4.4). This was a successful experiment in two ways: resolving the CNV and demonstrating that there was a mixed population. Due to the wide distribution of read lengths within a sequencing sample (Figure 4.3) which is skewed towards short reads, it is much less likely to isolate a read with a higher copy number than it is to isolate a read with a lower copy number. There was, therefore, no attempt here to use these single reads to quantify the average copy number of the sample as each copy number, from 1 to 5, would have a different minimum read length needed to observe it and therefore a different chance of being observed, given the spectrum of read lengths generated. It is more reliable to use the read depth as this is derived purely from how many reads covered that position with no stipulation of their length.

It could not be confirmed if the original stock of UK54 had been clonally derived. Public Health England guidance suggests that multiple single colonies from a plate can be used to prepare samples for sequencing and also for identification (265). This strategy is a good way to represent the diversity of bacteria from the sample and will assist in accurate phenotyping for this specific scenario. One such scenario where this strategy would be beneficial, for example, would be if a mixture of antibiotic resistant and sensitive isolates are on the diagnostic plate and picking a single colony would inaccurately describe the antibiotic resistance profile of the sample. As mixed populations were sequenced here, the quantity of *B. pertussis* samples on the SRA which are not clonally derived is unknown. It could not therefore be established if the mixed populations from this initial sample had evolved in vitro or were present in vivo. Despite this, I hypothesised that these CNVs were unstable in vitro under limited passage, as this has been previously documented (167,208,266). I repeated the experiment with clonally derived samples.

4.4.2.2. Mixed populations from clonally derived isolates

Eight clones from the UK54 stock were isolated and their copy number established using qPCR (Figure 4.5). qPCR was used to screen for copy number differences between clones and to select clones to sequence. Clones 4 and 8 were chosen as they captured an average (copy number 4) and extreme copy number (copy number 51) respectively. The results from sequencing clone 4 clearly indicated that mixed copy numbers could still be recovered as reads with both 5 and 2 copies were found (Figure 4.6). This therefore meant that culturing UK54 from both single colonies and the original stock produced a similar result. It could therefore be established that copy number changes were being generated rapidly during limited in vitro passage. Although the 3 reads with two different copy numbers observed in this experiment were sufficient to prove the hypothesis, it was unexpected that many fewer reads were found to span the locus in this sample (3) than in the original sample (9). This may involve an interplay between the generated read lengths and the time taken for tandem arrays to segregate (125,167,208). Further experiments showing how tandem arrays degrade over time and understanding how this process would be expected to be observed, given the spectra of read lengths generated by the Nanopore platform, should be undertaken.

Finding that brief in-vitro passage can lead to intermediate copy numbers supports that the 'inaccuracies' of copy number estimates from Illumina data generated in Chapter 3 were at least partly due to populations of mixed genotypes being sequenced. This evidence indicates that entering clinical isolates into in vitro passage leads to generation of mixed populations of cells. This is highly interesting and one of the key results of the thesis. What remains unsolved, however, is whether or not in vivo instability (211) occurs in *B. pertussis*, as observing in vitro instability of the CNV does not exclude the possibility that the observed genome plasticity of the original mixed sample of UK54 also happened in vivo.

4.4.2.3. 'Extreme' gene amplifications

One of the most surprising results generated here was the isolation of a clone with a copy number of 51. Whilst such an outlier was at first approached with scepticism, read depth coverage on the Nanopore platform agreed with the qPCR results. It is widely reported that tandem arrays are highly unstable and that instability increases with increasing copy number (208,267). Such knowledge comes from experimental systems which are manipulated with extreme in vitro selection pressures such as single carbon sources (161). Whilst mutations are being produced at all loci with an array of mutation types (SNPS, indels and inversions), the frequency at which mutations are observed within a population is determined by both its mutation rate and its fitness impact (234, 256). Therefore, if thousands of UK54 clones had been studied and picked in the same way (which minimised selection), it is expected to see a number of ‘extreme’ copy number clones, yet in the present study only 8 clones were studied.

These ‘extreme’ gene amplifications have been studied in bacteria for decades, although mostly in non-pathogenic experimental systems. A recent landmark study by Nicoloff et al was able to demonstrate that clinical isolates of several species contained copy number variable repeats and the isolates were able to rapidly amplify their native antibiotic resistance cassettes, in some cases up to 100 copies, in response to antibiotic challenge in vitro (208). Whilst these isolates were clinically derived, they underwent extreme amplifications in response to extreme selection pressures. It is therefore puzzling why a clone had been isolated with 50 copies of the CNV locus as it was assumed that the CNV was not being selected for. A key direction for future work is ￼to identify the impact on fitness of these CNVs, although this may help to further identify why this clone was isolated. The most important parallels between the Nicoloff et al study and the work presented here is that both studies are observing amplification of genes using systems native to the isolates(208), rather than relying on genetic constructs and mutants (124,161). This highlights the need to understand more about the UK54 population and why such a high copy number clone had been isolated.

4.4.2.4. Removing adapters via homology search is effective

Given that there was a lack of pre-existing tools and/or data to identify chimeric reads with very high sensitivity and specificity, alternative methods were undertaken. Adapter sequence searches were combined with read-consensus homology searches to indicate if a read is likely to be a chimera. The results showed that gaps in alignment between the read and the consensus sequence were highly correlated with the presence of adapter sequences (Table 4.2). In the cases where adapter sequences were not found but a gap had occurred, it can be assumed, due to the close relationships between gaps and adapters, that this ‘foreign’ DNA is an adapter sequence which was sequenced with drastically lower accuracy. Therefore, any read which contained an adapter or gap directly adjacent to the repeat was excluded (Figures 4.10 and 4.11). This was a highly effective strategy which produced reads which had no evidence of being chimeric.

Due to the results of the simulation study (Figure 4.14 and Table 4.1), it was decided to not rely on only string-matching tools such as Porechop (available: https://github.com/rrwick/Porechop). Porechop is one of the most popular read trimming tools which removes adapters from the start/end of reads and splits reads or deletes them if adapters are found in the middle. It is technically possible to rely on Porechop with a low adapter homology threshold to split chimeras- even though this may produce many false positives. The resulting highly fragmented set of reads would likely still contain the right information (e.g. reads containing DNA from two different parts of the genome adjacent on the read). There are many disadvantages to this strategy however and it was not pursued for the following reasons. The primary concern was that there would be less ‘anchoring’ DNA on either side of the junction leading to a difficult analysis for some reads and given that Porechop may split approximately 20% of the reads (according to the simulations I undertook). A second reason was that as part of the presented method, each read in the raw data was split into an individual file and Porechop would make extra files, the running of the method presented here would become more technically challenging. This research indicates that if it is imperative to find chimeric sequences with minimal false positives then reference guided adapter exclusion performs well. Repeating the work proposed here with Nanopore’s new flowcells (R10), which promise improved errors rates, may make direct chimera detection more reliable. Preliminary research of the R10 sequencing chemistry on the Oxford Nanopore sequencing platform shows that it is possible to reduce the raw error rate of reads to 5-10% (269).

4.4.2.5. Examining Nanopore ‘squiggles’ to find chimeras

An avenue of future research into the identification of chimeras is to investigate other hallmark signals of chimeric reads. During Nanopore sequencing, the final DNA sequence of each read has been interpreted from the raw electronic signal the DNA molecule made as it passed through a nanopore (colloquially known as a ‘squiggle’ signal). This signal is constantly being read and the signal between an empty pore and a pore with a DNA fragment in are quite different. When a read enters and exits the pore, this characteristic change in signal can often be easily interpreted by basecallers. This process sometimes does go wrong, however, leading to the in-silico formation of chimeric reads. Other studies have demonstrated that these chimeras can be manually diagnosed by re-examination of this squiggle (260,261). This appears to be true for both in silico chimeras (two reads passing through the pore in quick succession) and true chimeras (two DNA molecules attached by adapter sequences) where the specific signal of an adapter sequence can be recognised. Whilst this is impractical for hundreds of reads, as the number of reads is so low in these analyses, this method is entirely feasible for 1-100 reads.

Turning electronic signals into DNA base calls is a highly complex task which is achieved with high accuracy and efficiency by many well used bioinformatics tools. It is a rare event when these tools create chimeras from two reads passing through a pore consecutively. A lack of homology between the junction of these two reads to the species being sequenced may a useful way to recognise chimeras. This should be a focus of future work so that an automated `chimera polishing' tool can be established.

4.4.2.6. Computational inefficiency results in inconclusive data

Using BLAST to find reads with juxtaposed DNA sequences is resource intensive and therefore only two sequence datasets were investigated. The most compute intensive task is the BLAST search itself which compares each read to the consensus sequence. This step takes roughly 24 hours to run whilst the rest of the analysis can be carried out in approximately 6 hours. Therefore, there is scope to drastically improve the efficiency of this method.

One method to speed up this process would be to rely on pre-existing mapping pipelines which are well designed for the task of homology searching (270,271). Most of these tools will mark reads which do not fully align to the consensus sequence by either noting a ‘not primary alignment’ or ‘supplementary alignment’ value in the FLAG field of the SAM file format. Any read that does not wholly and uniquely map to the reference sequence would be noted with these warnings. Whilst all the SVs noted here would be highlighted in this way, many false positives would be flagged too. For example, an additional source of supplementary alignments from a mapping pipeline would be Nanopore sequence glitches. Glitches are stretches of DNA that have been sequenced poorly on the Nanopore. It is likely that many reads highlighted as divergent by mapping pipelines would be due to the high repeat content of *B. pertussis*. By using the mapping pipeline to find reads which don’t map uniquely to the consensus sequence, the number of reads which are analysed could be reduced but it could not be assumed that all (or the majority) of reads that map with a supplementary/not primary alignment would be true SVs. This means that these reads must still be analysed using a BLAST-like approach.

Whilst two Circos analyses were presented here, they gave distinctly different results (Figures 4.17 and 4.18). The analysis of UK76 showed that there was considerable heterogeneity of CNV content and low levels of de-novo CNV generation whilst UK54 showed only low-moderate levels of denovo CNV generation. Following an improvement in efficiency of the Circos analysis, more sequencing samples should be undertaken to investigate a number of trends that could exist in the data. For example, it is unclear if all CNVs are as equally unstable as the UK76 CNV.

4.4.2.7. Circos analysis can be undertaken using reads <=7kb

Both samples analysed with Circos analysis (UK54 and UK76) were generated by ultra-long Nanopore sequencing, however, these reads need only to span a repeat element to be used in this way. This means that reads that are 7kb long will span all junction sequences in *B. pertussis,* taking into consideration insertion sequences (~1kb), rRNA operons (6kb) and a two-copy 3 gene CNV that is found in some isolates (3kb) (23). This is far shorter than was generated here on the Nanopore platform and is possible to achieve on the PacBio platform too.

This means the method presented here is generic and can be used to analyse the back catalogue of over 500 PacBio sequenced *B. pertussis* isolates. Future work could include analysing this dataset and merging the results to gain a better understanding of which genomic loci experience de-novo SVs in *B. pertussis*.

4.4.2.8. Have mutations occurring by other mechanisms been captured?

In this data I studied only SVs that were bounded by large (>=1kb) repeats. Whilst this is the type of SV that has been observed previously in *B. pertussis*, recombination between non-homologous regions or recombination between small repeats is possible. There would have been many reads that were highlighted in this analysis that were SVs that occurred by these mechanisms but were discarded. Future work could include analysis of these reads to investigate how alternative forms of SVs form and if these have remained undetected in the analysis of CNVs in the large cohort analysis in Chapter 3.

4.4.3. The enigmatic genome of UK76

Having thoroughly investigated the ‘Circos’ method in UK54 clone 4, a second sample was analysed.UK76 had a large CNV that in its tandem configuration is predicted to be over 600kb in addition to having a copy number estimate of 1.2. As the prediction was an intermediate copy number it indicates that some cells contained the CNV and some did not, although this was not confirmed. Nanopore sequencing this sample and analysing the breakpoints found provided an extraordinary discovery: on top of the predicted variation in CNV copy number there were multiple different variants of the predicted CNV detected in the sample. It was confirmed by plotting a number of the highlighted reads that there were at least 6 different CNV breakpoints confirmed in the dataset (Figures 4.19-4.22 and 4.24). This meant, therefore, that there were likely two different mechanisms by which genetic diversity was generated in this sample: copy number and CNV composition. These results were not investigated in UK54 as the current methodology only analyses reads which have putative SVs longer than 30kb and UK54 had a 16kb CNV.

The data does not, however, shed any light on the series of events that generated this diversity, although it can be speculated on. It is important to note that this sample was likely not clonally derived and therefore it is unclear how many passages this had undergone and if the diversity represented the diversity in the host or had been generated in vitro. There are two scenarios for how this diversity was generated: a single CNV with subsequent recombination or multiple independent duplication events. It is known that CNVs can undergo subsequent SVs as part of the inherent instability generated by tandem homologous sequences. For example, it was found that in the lac system that while clones were amplifying the lac operon, these CNVs they were also undergoing further SVs to reduce the fitness cost of the amplification (124,161). It is possible therefore than the observed CNV diversity in UK76 was observed from a single SV mutation and had subsequently segregated into multiple novel forms through further recombination. Further work could include passage of the UK76 clones to establish the CNV dynamics in-vitro.

4.4.4. Ultra-long Nanopore reads allows analysis that outperforms other strategies

Ultra-long read generation was undertaken here to successfully resolve a CNV in its tandem configuration. This is a highly revolutionary technique that has the potential to replace multiple traditional methodologies to describe and elucidate CNVs. In both Chapter 3 and here in Chapter 4 qPCR was used to amplify DNA inside the hypothesised CNV and compare to a single copy region elsewhere on the genome. Whilst qPCR is a ‘tried and true’ methodology which is relatively cheap to undertake, it was found to consistently agree with depth of coverage estimates generated by the Nanopore platform, qPCR is therefore a good screen for further analysis but only answers a very specific question in comparison to sequencing.

In order to achieve sufficiently high read depth to capture putative denovo CNVs or to achieve ultra-long reads, a whole Minion flowcell was used to sequence a single strain. As these flowcells cost approximately (at a minimum) £500, these sequencing experiments were expensive. If Nanopore sequencing was used to merely find the difference in read depth between two regions of the genome (as was used in Chapter 3 with Illumina data) then runs could be barcoded and 12 samples could be sequenced on the same flow cell. This considerably reduces the cost of Nanopore sequencing and brings it to near parity with qPCR, depending on how many primer/probe concentrations are used. In addition, to use qPCR you must know the sequence of the CNV region, whereas the same is not true for Nanopore sequencing.

A second novel application of long reads was proposed here: analysing single reads for SVs not present in the consensus sequence. For example, both Nanopore and Pacbio platforms can produce reads which can span the junction of an SV. It was demonstrated here that long read sequencing on the Nanopore platform can elucidate sub-populations in a mixed population without prior knowledge of their existence. In UK76 it could be shown that in this single sample there were at least 6 different CNV start and end points detected.

This is a unique type of analysis that is hard to achieve using other methods without considerably more resources. The same results could be achieved by isolating hundreds of clones of UK76 and sequencing them on a long-read platform. This would result in the same junctions being recovered but at a consensus level in each sequenced clone. Alternatively, there are several classic molecular assays that are similarly cumbersome to undertake compared to the proposed Circos analysis here. Transduction or linear transduction assays aim to establish if a CNV is pre-existing in the population. The process involves identifying a metabolic gene within the duplicated locus and inserting a selectable marker, once per cell. If the metabolic gene was present in two copies then it will be both prototropic and resistant as it would have one intact metabolic locus and one resistant locus. This however requires knowledge of the predicted duplication and a known metabolic gene in the CNV. This method is therefore only suitable in a narrow range of applications in well studied organisms.

4.4.5. Evidence that *B. pertussis* is genetically diverse

B. pertussis is widely described as an organism that has low genetic diversity (42,112). This is normally in reference to single nucleotide variants but can often include the loss of genes via deletions, both of which are mutations that are easy to study with short read data (42,120,212). Reviewing the results generated in Chapter 3, Chapter 4 and in a number of studies by Weigand et al., the view that *B. pertussis* is an organism with low genetic diversity is beginning to be revised (130,131). These three lines of evidence can be used in conjunction to show that *B. pertussis* has previously undescribed genetic diversity. Here it was demonstrated that *B. pertussis* can readily generate de-novo genetic diversity by both inversions and deletions/CNVs (Figures 4.17 and 4.18). This is directly related to the results found in Chapter 3 where it was found that in >2700 *B. pertussis* isolates >200 CNVs were found and the work by Weigand et al which found many genomic rearrangements present in the population. It was found that novel SVs are created over short time scales and that CNVs/rearrangements were found in a global cohort of isolates. These three lines of evidence can be used to conclude that *B. pertussis* readily creates genetic diversity in ways previously under-appreciated.

One of the key results in Chapter 5 and the work by Weigand et al was that whilst there is a sizeable quantity of SVs in the global *B. pertussis* population, they were found at common regions (130,131). In the CNV analysis the mutations were found at 11 hotspots and a similar conservative pattern of genome rearrangements was found by Weigand et al. Naively examining these results and the results in Chapter 4 in conjunction presents a confusing scenario: how is it possible that there is promiscuous generation of SVs (as evidenced in Figure 4.17) but conserved patterns detected in the global population? I hypothesise that this is because such patterns are influenced by the forces of selection. The power of selection in comparison to other forces such as mutagenesis chance is unknown, however.

4.4.6. De-novo assembly can indirectly identify CNVs

4.4.6.1. Graph based genome assemblies

The human genome project was a monumental effort to create a representative human haploid genome sequence (272). However, it was not representative of the human population. Similarly (although on a different scale), the research outlined in this chapter highlights that consensus assemblies of bacterial populations are not representative of the true population of bacteria from which they are derived. Whilst understanding global human genetic variation is highly important so too is a similar appreciation of diversity in bacterial and viral populations, both on intra-sample and global scales. For bacterial isolates, a body of work exists that antibiotic resistance heterogeneity is a powerful force in antimicrobial resistance (208) whilst for viruses there have been multiple studies that indicate understanding the composition of viral quasispecies can aid in understanding their epidemiology (273,274). I therefore propose that the methods presented here can aid in the understanding of the makeup of a population of cells (or viruses).

The Circos analysis presented in this chapter represents the structural variations that had been found in the population. The use of this analysis can be greatly improved however, by formally describing these results in a graph-based genome assembly. This is a much more holistic solution to representing diversity in populations as graphs can contain SNPs and indels whereas the Circos analysis cannot. Representing all types of mutations in a single structure can be achieved by using genome graphs which replace genome consensus sequences. The aim of these is to create a graph of the observed DNA diversity but collapse only sequencing errors into a consensus sequence. For example, a graph-based genome representation of UK76 would be composed mainly of the consensus sequence, but at the CNV loci (of which at least 6 were detected) there would be alternative ‘bubbles’ or ‘arcs’ that represented the alternative gene order contained in these reads. This would be a true representation of the sequenced DNA rather than just an average sequence. This methodology, however, does require each variant to be have good coverage, which was achieved in the UK76 Circos analysis but not in UK54.

Most importantly, once a graph-based genome representation has been created, reads from other datasets can be mapped to it in order to find if they contain similar genotypes. Mapping to a graph-based genome would be a highly attractive extension of the current work and answer crucial questions that arose during my research. For example, one of the most exciting questions arising is whether the 11 CNV hotspots described in Chapter 3 arise because of selection or because they are more frequently occurring. This can be answered using graph genomes by comparing the CNVs found in the global cohort of isolates to de novo generated CNVs in all available long-read data. A graph containing the B1917 sequence with the unique junctions found in the 274 CNVs found in the global cohort of isolates in Chapter 3 could be created and all long-read data on the SRA (500 samples generated mostly on the PacBio platform) can be mapped to the graph. This graph can be explored by analysing the reads that map to CNV junctions. If the junctions had relatively high coverage and came from a diverse set of sequencing samples it may indicate that these loci are predisposed to becoming multi-copy regions whereas if these junctions are equally as likely to occur in in-vitro culture it is likely that these CNVs were formed primarily by selective forces.

4.4.7. Linking genotypes to phenotypes: an initial step

The data presented here demonstrates and echoes one of the key findings of Chapter 3: *B. pertussis* is adept at generating genetic diversity via structural variant mutations. Little is known about the functions of many genes in the *B. pertussis* genome outside of the vaccine antigen genes. The full impact of the main CNV studied here, a 16kb CNV in UK54, is unknown. To elucidate more about this CNV an RT-qPCR experiment was conducted to investigate the expression changes associated with copy number changes-a vital first step in elucidating genotype-phenotype links. It was found that additional gene copies led to increased transcript level. This is as expected and has been found in many other studies.

In this experiment, only one gene within the CNV had its expression assayed. It is possible that the other genes will have different relationships between copy number and expression. Further work should investigate these genes which would lead to further descriptions of the dynamics of CNVs in *B. pertussis* and also assist with finding genotype-phenotype links. This may be achieved simply by designing new primer/probe pairs to study the other genes in the CNV or, in a much more sophisticated way, an RNA-seq experiment could be devised.

Studying genome-wide gene expression using RNA-seq would elucidate the impact of CNVs on gene regulatory networks in addition to describing the expression of all genes in the CNV. By understanding the regulatory networks these genes are in, the role of these genes in the cell can be discovered. For example, if increased copy numbers of these genes are linked to increased expression of genes in a particular metabolic pathway, this may mean these genes are also involved in that pathway-in some way. Conducting an RNA-seq experiment would therefore be a logical next step in studying CNVs in *B. pertussis*.

The ‘Holy grail’ of genetics is to find genotype-phenotype links. This often is done by creating knockouts and observing phenotype changes which can be a lengthy process in non-model organisms like *B. pertussis*. CNVs may form a natural experimental system to investigate the role of genes in *B. pertussis* - increased copy numbers would lead to altered phenotypes and systematic disambiguation of exactly which genotype causes which phenotype can be conducted using a GWAS. This is the question that is explored in Chapter 3: Can structural variants be used as a genotype in GWAS?

5. Chapter 5- A preliminary investigation into Genome Wide Association Studies for *B. pertussis*

5.1. Introduction

The DNA sequence of a cell ultimately determines its phenotypes. Establishing the link between genotype and phenotype, however, is highly challenging. Genome Wide Association Studies (GWAS) provide a statistical method to associate the known genotypes of a population to a phenotype. One of the biggest hurdles GWAS aims to overcome is that the DNA sequences causing a specific phenotype are inherited with sequences which do not contribute to the phenotype. This is known as linkage. Because bacteria reproduce asexually, linkage is preserved through the replication process which leads to lineages having exceptionally similar genomes. This can be controlled for, to an extent, by employing different models which account for the underlying population structure that is being studied. The common sources of linkage disequilibrium (the disruption of linkage) are inter-molecular recombination, normally between two different cells, and the same mutation arising independently in two or more lineages.

The job of statistically untangling a causal DNA sequence from a non-causal sequence cannot be achieved when causal and non-casual genotypes are in complete linkage. Because *B. pertussis* is highly clonal, does not undergo horizontal gene transfer and has a low SNP-rate, it has a genome with a very high degree of linkage. If many mutations are in complete linkage (e.g. always occur together) then this makes establishing specific genotype-phenotype links impossible. In this chapter I investigated how suitable CNVs and deletions would be for this kind of analysis, given the highly clonal nature of the species and having previously shown CNVs to be highly homoplasic in Chapter 3.

5.1.1. The need for genotype-phenotype links in *B. pertussis*

There is a lack of knowledge of the impact on phenotype of all classes of mutation in *B. pertussis,* although of SNPs and deletions, the two most commonly studied mutation types, SNPs are better studied than deletions. This is despite deletions in *B. pertussis* being highly important to the genetic diversity of the species as *B. pertussis* does not participate in any ongoing horizontal gene transfer. Deletions are the only way the accessory genome of approximately 10% is generated.

Four studies have previously looked at deletions in large (>50 isolates) cohorts of isolates and found some level of homoplasy for a limited number of deletions. It is likely that the level of homoplasy is under appreciated, however, as deletions have not been studied using WGS for large strains collections, with previous studies using DNA hybridisation arrays and PFGE profiles to establish phylogenetic relationships. These pre-WGS methods only provide coarse phylogenetic relationships and are not capable of detailing the finer grained relationships with confidence. In addition, they were undertaken on small datasets containing less than 200 isolates each. Only one of the four pangenome studies in *B. pertussis* used a SNP based phylogeny to find phylogenetic relationships between isolates with the same deletions and was therefore capable of determining fine grained phylogenetic relationships. This study, however, only used 16 isolates.

In order to inform on the viability of a future GWAS to establish genotype-phenotype links for deletions, I aimed to provide an updated WGS based description of deletions and analyse how homoplasic they were. I hypothesised that with updated methods, which can resolve phylogenetic relationships with high precision and with a larger set of isolates, that finer-grain phylogenetic analysis would reveal gene deletions to be more homoplasic than previously thought, leading to deletions having a high linkage disequilibrium and therefore being good candidates for GWAS in the species

In Chapter 3 I detailed the 273 CNVs of a cohort of 2431 *B. pertussis* isolates. This was a large dataset of CNVs contained in the known population of *B. pertussis*. Mapping these isolates onto a phylogenetic tree showed that CNVs were highly homoplasic and occurred in multiple genetic backgrounds. These qualities mean that CNVs are excellent candidates to phenotype using

GWAS. GWAS has yet to be adapted to study CNVs as a mutation type in bacteria. This is likely in part because CNVs are not often described but it is also because the current GWAS methods cannot detect them.

Current GWAS tools fail to account for CNVs as they either use K-mer presence/absence or SNPs and gene presence/absence as genotypes. State of the art GWAS analyse DNA in sequences of ‘k’ length: K-mers (e.g. 12-mers, 14-mers or 21-mers) (70) in order to simplify the calling of genotypes. The presence or absence of K-mers not only is a proxy for SNPs in the isolate, but also gene presence/absence (275). K-mer based tools for bacterial GWAS only interpret K-mer abundance as presence or absence and do not acknowledge copy number states beyond this, such as duplications or triplications. Tools which rely on a user-defined list of SNPs or presence/absence of genes also are not designed to use CNVs as genotypes as of yet.

GWAS may not be effective for studying all mutation types in *B. pertussis* and in this chapter, I aimed to describe how different mutations could be used as input to GWAS. I aimed to define a method to represent CNVs suitably for this analysis (as CNVs were shown to be highly homoplasic in Chapter 3) and the degree to which deletions are homoplasic.

5.2. Methods

5.2.1. Representing CNVs with K-mers

K-mers were generated using fsm-lite (available: https://github.com/nvalimak/fsm-lite) on the default settings to make all K-mers of length 21-100 of the input sequences. To test the suitability of K-mers to find deletions and CNVs, three isolates with closed genomes had K-mer counts generated from their assemblies: one with a CNV and two with the same locus at single copy. Only K-mers that occurred once in the genome without a CNV were used. If a K-mer was found once in the control genome but two or more times in the CNV genomes, its abundance was set as ‘1’ and it was found once it was set as ‘0’.

5.2.2. Deletions

The same dataset was used for deletions as CNVs in Chapter 3. Deletions were counted if the predicted copy number was <=0.1 as copy numbers between 0.1 and 1 are more likely to be formed from one of the weaknesses of read mapping: reads that map to more than one place get assigned to just one of their mapped regions. Multi-mapped reads become distributed in this way to avoid repeat regions having greatly enhanced coverage. If reads that mapped to multiple places were counted multiple times then repeat regions would have coverage proportional to their copy number.

5.2.3. Phylogenetics

Phylogenies were produced as previously using the core-genome SNPs dataset from Chapter 3. One clade of the tree was plotted here in order for the tips of the tree to be visible. This allows the scale of the tree to be apparent and the degree of homoplasy to be visible.

5.3. Results

5.3.1. Using K-mers derived from closed genomes as a signal of CNVs

As the CNV predictions in Chapter 4 were based on the read depth of short read data, the same signal (increased coverage of CNV loci) can be exploited again to produce K-mer abundances instead of read-depth. Copy number states beyond presence and absence, such as duplication or triplication, are not considered in most GWAS. Therefore, this means that some isolates having twice the frequency of some K-mers is an invisible signal to the analysis and gets analysed merely as K-mer presence rather than a possible K-mer duplication. I modified K-mer abundance data to describe copy number states above 1. The dataset consisted of closed genomes sequences: one genome with a CNV and two genome sequences with the locus at single copy (contained in network 1 of the analysis in Chapter 3). K-mers that were uniquely mapping in the two genomes with single copies were kept and their abundance in the isolate with a CNV was plotted. To simplify visualisation, the median K-mer abundance was plotted in 10kb windows (Figure 5.1).

This experiment showed that a modified K-mer abundance highlighted only the CNV locus with no false positives or false negative K-mers being highlighted. This analysis is limited in its potential at the time of writing, however, as there are only 10 isolates with CNVs resolved at this locus (see Chapter 3) and other loci have even fewer CNVs resolved. Whilst it is possible a GWAS on this locus would be successful, the analysis would remain inflexible and limited in scope by the lack of resolved genome sequences- only 1 out of the 11 hotspots could effectively be studied.

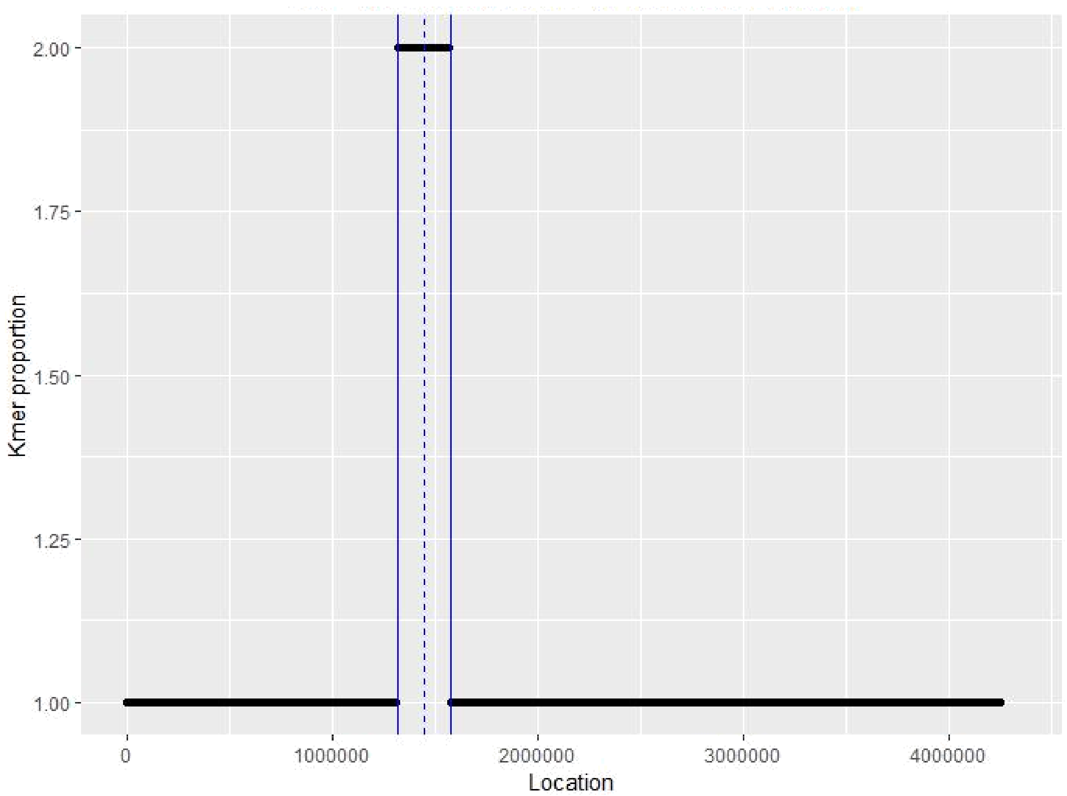


Figure 5.1 K-mers were generated for two closed genomes containing an CNV in its tandem array and compared to a genome without a CNV. The differences in the abundance of these K-mers between the CNV and non-CNV group was expressed as a proportion (Y axis). K-mer location (X) was plotted against this proportion in 10kb windows. The entire tandem CNV locus (blue solid lines), including the junction sequence (blue dashed line) is noted and it can be seen

that the median K-mer proportion in this area is 2 and therefore K-mers can be used in this described way as a proxy for copy number variants.

5.3.2. Using K-mers derived from sequence reads as a signal of CNVs

Sequence read abundance was used to generate the CNV predictions in Chapter 3 and this can also be analysed using K-mers. This would circumvent having to fully resolve CNVs using long reads and genome maps, thus increasing the scope of this work. There are two major problems with this strategy. Firstly, read-depth based tools such as CNVnator (198) (in addition to my own modifications) have many complex normalisation and read-depth smoothing algorithms to process the read-depth data into predictions and any K-mer based approach would have to recreate these steps. Secondly, analysing raw reads may be computationally demanding- to the point of being prohibitive. Due to the high redundancy of sequence data (often being 100-1000x the size of the final assembly), analysing K-mer abundance in such data is challenging.

I trialled an analysis of Pyseer on raw sequence data of 400 isolates. This analysis was not adapted for CNVs but instead the computationally simpler task of detecting SNPS and gene presence/absence. I found that whilst generating the K-mers was possible, Pyseer was not able to analyse the data with the 60Gb of RAM available on the server used. Analysing K-mers in raw sequence reads therefore is likely possible but is beyond the scope of the current study.

5.3.3. Deletions are more homoplasic than previously thought

There have been limited study of deletions in *B. pertussis*, with all studies relying on outdated methods such as SNP typing (to produce phylogenies) and/or comparative genome hybridisation (to establish gene presence/absence (120,181,184,187,212). The two studies which looked specifically for homoplasies found that a limited number of deleted loci were homoplasic (184,212). The dataset of 720 isolates with high quality Illumina data used in Chapter 3 was therefore a significant advancement over previous datasets and reflected recent technological advancements in sequencing. I therefore analysed these data to find homoplasic deletions.

Analysing the large cohort dataset in Chapter 3 identified 474 deletions which occurred in 50 networks of which 22 had 3 or more nodes (deletions). The 10 most numerous networks are described in closer detail (Table 5.2). Most of these networks had a core set of genes that were highly representative of the mean length of the network which indicates the networks did not have considerably variable gene content. In addition, most networks had high network density indicating that most of the deletions overlapped with each other. The deletion networks were less variable than the CNV networks found in Chapter 3. This matches previous research which found (with limited tools) high conservation of deletion start/end locations (120,181,184,187,212).

Table 5.2. The top 10 most frequently found deletion networks. All loci were found to have been deleted in at least one strain in the literature ('previously known' column). Only three networks had been found to contain homoplasies in the literature ('Previously homo') yet in the 7 networks that homoplasies were searched for, 6 networks had at least one homoplasy.

Network	Frequency	Mean length	Median start	Median end	Previously known	Demonstrated homoplasic	Previously homoplasic	Homoplasy reference
1	196	20	B1917_RS13515	B1917_RS13610	Y	Y	Known homoplasy	Van gent et al 2012
2	75	9	B1917_RS03580	B1917_RS03625	Y			
3	49	20	B1917_RS07960	B1917_RS08060	Y	Y	Known homoplasy	Lam et al 2014
4	34	17	B1917_RS03115	B1917_RS03235	Y	Y		
5	32	15	B1917_RS15835	B1917_RS15905	Y	Y		
6	28	31	B1917_RS17200	B1917_RS17365	Y	-		
7	27	11	B1917_RS19625	B1917_RS19455	Y	N		
8	27	7	B1917_RS13865	B1917_RS13895	Y	-		
9	27	10	B1917_RS02505	B1917_RS02555	Y	Y	Known homoplasy	Lam et al 2014
10	26	12	B1917_RS15305	B1917_RS15365	Y	-		

I found that many studies had reported on deletions occurring at loci equivalent to 9 out of these 10 deletion networks found here ('Previously known' in Table 5.2). This corroborated that the present analysis was of high quality as the same results had been found in other isolates. A direct comparison between the deletions found here and previous results was not possible, however, due to different sets of isolates used in addition to a different reference genome (B1917 was used here and Tohama used previously). This meant that the false positive and false negative rates were not known.

Previous research had shown that a number of deletions were homoplasic as they had occurred independently multiple times on a phylogenetic tree. I therefore wanted to investigate homoplasies in my dataset, given that it was larger than previous ones. A deletion is much more heritable than a CNV since a deleted gene cannot go back to single copy (as there is no appreciable level of horizontal gene transfer occurring in *B. pertussis*). This therefore meant that fine scale relationships between isolates would be important to accurately resolve or otherwise spurious homoplasy events may be predicted. I therefore sought to produce a bootstrapped tree which had high support. Bootstrapping provides a level of confidence in the phylogenetic placement of each isolate by removing one tree tip at random, remaking the tree and assessing if the topology changes.

I ran HomoplasyFinder (207) which uses consistency index to find homoplasic traits or mutations (277,278). The consistency index aims to score patterns of traits (or nucleotides in a core-genome alignment) observed in the tips of the tree against their phylogenetic relationship. In my application here, the consistency index was used to find how consistent deletions were with the phylogenetic tree. Each trait is scored between 0 and 1, with 0 being a pattern which is completely inconsistent with the phylogenetic tree and 1 being completely consistent with the phylogenetic tree.

Running homoplasyFinder resulted in 6 of the 7 studied deletion networks being predicted to contain at least 2 state switch events. In this example, a state switch from both 1 to 0 (a deletion) and 0 to 1 (a gene gain) were counted. Because a switch from 0 to 1 is improbable in *B. pertussis*, these may indicate poor quality parts of the tree and may inflate the results (Figure 5.7). Previous research had identified homoplasies in only three of these networks. The most comprehensive previous analysis of homoplasy (212) used SNP-typing to create a phylogenetic tree (which used 60 SNPs) and studied only 42 isolates with unique SNP profiles. These limitations may have caused the difference between those results the results presented here.

Most genes that were part of deletions that were not previously shown to be homoplasic were annotated as hypothetical and little was known about their function. There were some exceptions, however. Deletion network 5 was described as homoplasic by this analysis and also composed almost entirely of a prophage, D3. It is possible that this deletion is actually driven by excision of this phage, potentially under stress conditions. This would also give a plausible mechanism for the deletion to be complemented- the phage could jump back into the same position. Deletion network 6 contained multiple genes coding for pyruvate dehydrogenase (BP0628+BP0629) and other metabolic processes. It is possible that this deletion was under positive selection as these genes were superfluous to the metabolism of B. pertussis.

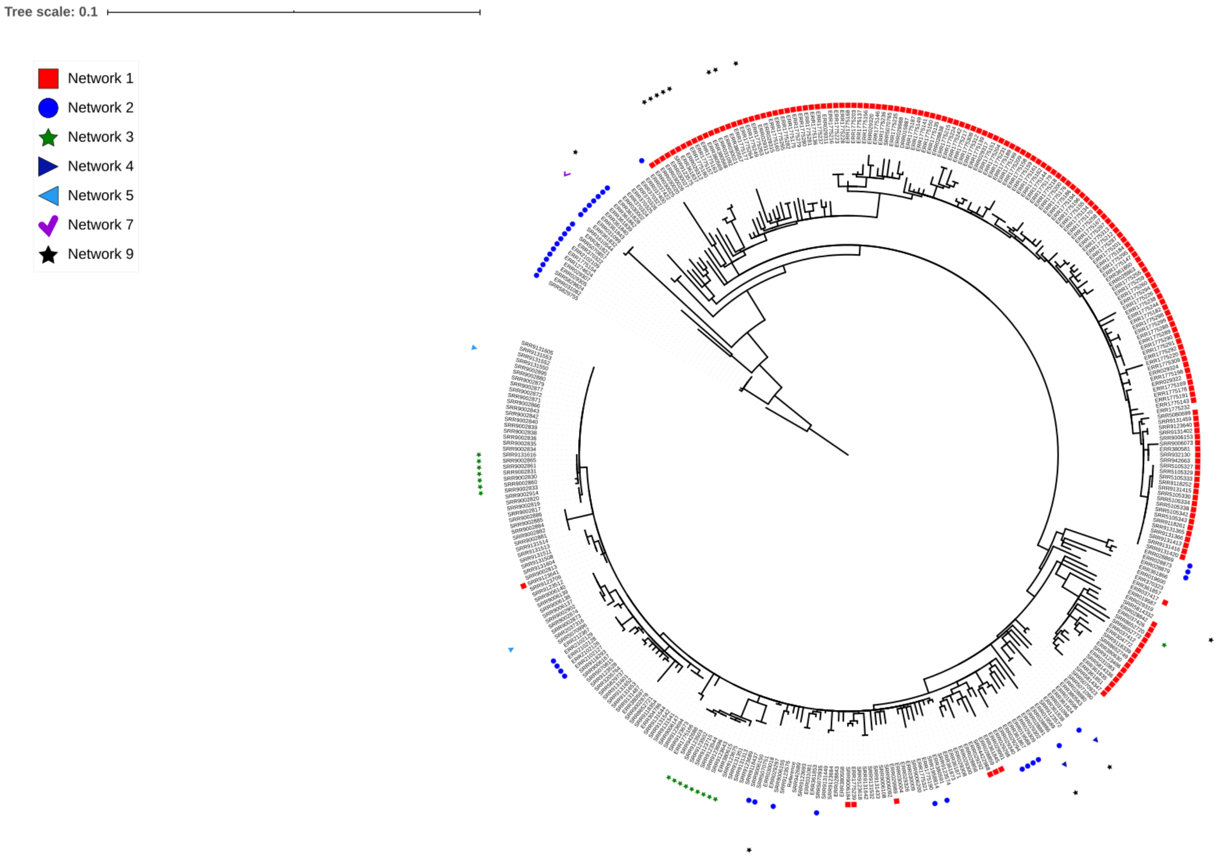


Figure 5.6: A portion of the full tree was plotted. Deletion networks 1,2,3,4,5,7 and 9 were annotated. Most deletion networks can be found in disparate clades of the tree. This shows that deletions are highly homoplasic, more than was previously known. Deletion networks 6 and 8 were not present in these isolates and were not annotated on the tree.

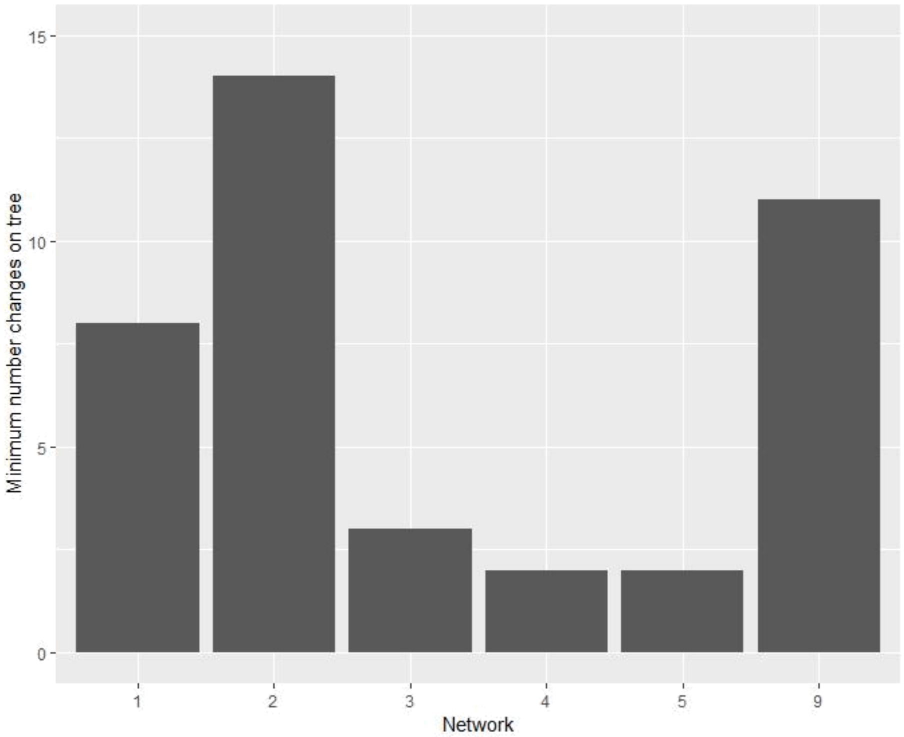


Figure 5.7: Minimum number of copy number changes at internal nodes needed to explain the observed pattern of deletions at the tips of the phylogenetic tree (displayed in Figure 5.06).

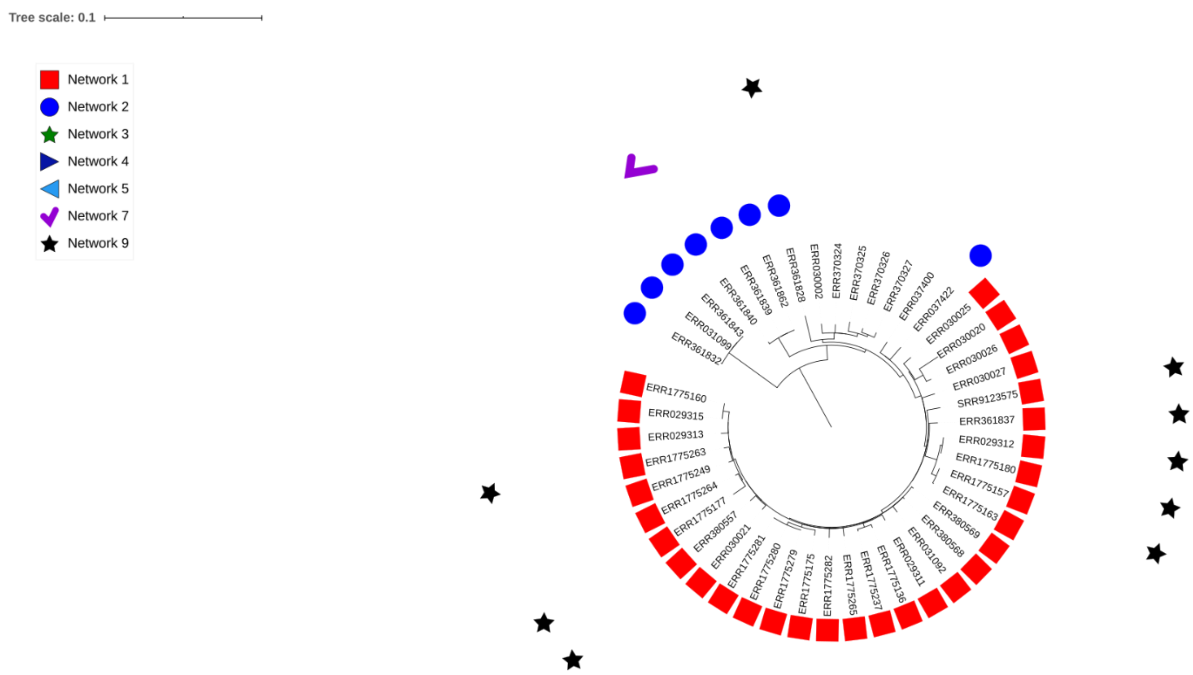


Figure 5.10: A subtree from Figure 5.09 of 45 isolates shows a fine-grained view of homoplasic deletions in *B. pertussis*. Networks 9 and 2 can be seen to be homoplasic, having occurred in multiple independent clades.

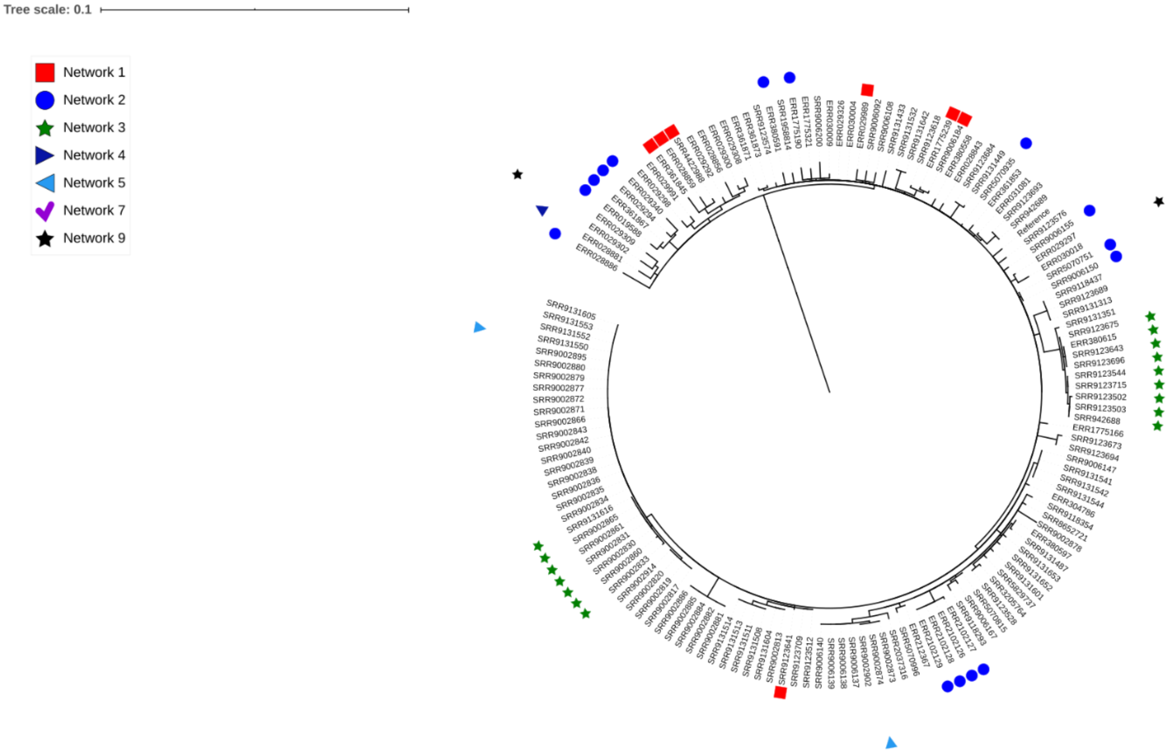


Figure 5.11: A subtree from Figure 5.10 of 146 isolates shows a fine-grained view of homoplasic deletions in *B. pertussis*. Networks 1,2,3 and 5 can be seen to be homoplasic, having occurred in multiple separate clades.

5.4. Discussion

5.4.1. Structural variations as an effective genotype for GWAS

The results showed that K-mers were challenging to use at scale. It is therefore likely that using CNVs coded as genomic intervals, as was used in Chapter 3, was the most reliabile way to use these mutations in GWAS. This may have been prone to technical artefacts, however. Technical artefacts that split a predicted CNV into two fragments (discussed in Chapter 3) would have little effect on genotype-phenotype links, under the assumption that the single copy genes in between the two fragments (often just 1 gene) were not crucial to the phenotype. As was discussed in both Chapter 3 and 4, further verification of the predictions will increase the confidence in their exact coordinates and also will improve the GWAS analysis.

5.4.1.1. Genome plasticity as both a hurdle and feature of GWAS

It was shown in Chapter 4 that clonally derived populations with CNVs are unstable during minimum passage in vitro which has also been shown previously (185). It was hypothesised that in the absence of selection and with sufficient amount of passages that the copy number would reduce, potentially back to single copy (208). This may be a considerable hurdle to defining the CNV genotype for GWAS as the copy number of a population will change during an assay. The impact of quickly segregating populations can be somewhat alleviated by using only single-use aliquots for assays. In particular, assays which do not involve direct measurements of growth (such as agglutination or complement killing) would benefit from such measures. Assays which directly measure growth (such as growth curve under various conditions or agar plate growth) would still benefit from using aliquots, but stochastic events may alter the spectrum of mutations between assays. The degree to which this will impact GWAS, however, is unclear and will need to be investigated in future work.

Genome plasticity is not just a hurdle, the change in the average copy number of the sample during an assay could also be a measurable phenotype in itself and thus be a feature of this GWAS. Nicoloff et al have previously shown that some specific isolates of certain species have antibiotic resistance cassettes that will increase in copy number under selection, sometimes up to 100 copies (208). Such an increase is a clear sign that the function of those genes is under selection and as such could be utilised as a binary or quantitative phenotype or as a confirmation that indeed the CNV is the causative mutation of the phenotype.

5.4.2. Considerations for future work

Using K-mers to analyse CNVs in *B. pertussis* was not pursued further although it remains attractive for further exploration. Advancements in long-read data generation and analytics or greater adoption of enzyme mapping (e.g Nabsys) may mean complex CNVs can be automatically assembled in the future (rather than manually resolved as in Chapter 3). In turn, this would mean that CNV-resolved isolates could be used with a K-mer based analysis in the future. Alternatively, there may be ways to reduce the computational load of analysing K-mers in raw sequence data. As these problems caused a K-mer based approach to be difficult, however, this strategy was not pursued further.

5.4.3. Homoplasious deletions

It was surprising that the deletions found here were described as homoplasic for the first time, given that gene deletion is a well-known path of genetic variation in *B. pertussis*. It is likely that limited diversity and quantity of strains in previous analysis contributed to this. This may have been caused by the deletion threshold (<=0.1 predicted copy number) causing spurious homoplasies to be predicted. This may be because some isolates with predicted deletions may have just had low read coverage at that region (false positive) or that a real deletion was present but for some reason had reads mapped to it (false negative). This is unlikely to fully explain all state-switch predictions as many of the nodes which were predicted to be homoplasic were deep in the tree and therefore any inaccuracies in finding deletions would have to extend to hundreds of isolates.

Mapping reads to a reference genome is not the most effective way to find gene deletions. This is because deletions containing repeat sequences can cause a systematic bias to occur due to the way reads mapping to repeats are distributed. This is another reason that deletions were only counted here if the region had a predicted copy number <=0.1. A more effective alternative is de-novo assembly as if there is a region with 0 coverage, it would not be assembled. Deletions were studied using mapping, however, as this offers a unified methodology to predict both deletions and CNVs. In order to definitively show that deletions are homoplasic in *B. pertussis*, denovo assembly, pangenome analysis and PCR verification of these deletions is required.

In this chapter I aimed to evaluate K-mers as a way to represent CNVs in GWAS and to investigate the degree of homoplasy found in deletions in *B. pertussis* . Evaluating these aims in relation to the results, I investigated K-mers as a way to represent CNVs but found that they could be used successfully only for assemblies but not easily for sequence reads and were therefore of limited use in the scope of this context. I then successfully expanded the known repertoire of deletions in *B. pertussis* and found that they were more homoplasic than previously thought. In conclusion, therefore, I believe that more work must be undertaken to represent CNVs using K-mers and that future GWAS will have the statistical power to associate deletions with phenotypes, given their homoplasic nature.

6. Conclusions

6.1. *B. pertussis* has an underappreciated repertoire of CNVs

The genetic diversity of bacteria is often described solely using base pair changes and gene presence/absence. This is despite a wide variety of other mutation types likely being major contributors. In Chapter 3, I sought to investigate the prevalence of CNVs in the population given the repeat-rich *B. pertussis* genome and that 28 CNVs had been found previously.

I established a customised CNVnator based pipeline which included steps to normalise noise between samples. This pipeline had between 75% and 85% accuracy and could scale to analyse sequence samples of varying quality. Due to the inherent quality of mapped short reads, however, it could not resolve complex CNVs (CNVs made of multiple SVs) well.
Using this pipeline, the number of CNVs identified in *B. pertussis* was expanded 10-fold, although future work should include verifying these predictions. Crucially, 93% of these CNVs were found at just 11 loci. This is in contrast to these mutations being randomly distributed around the genome, given that experiments in other species show that CNV frequency is related to proximity to repeat sequences and that *B.pertussis* has a high amount of relatively evenly spaced repeats (246). This data was analysed using network graphs which concluded that there was a diversity of different network topologies but generally most networks had a conserved core of genes around which they varied. These genes were implicated in motility, haemolysis and metabolism and likely affect these processes, given that it could be demonstrated (in Chapter 4) that the level of gene expression correlated with gene copy number for one CNV.

It is highly likely that other bacteria contain CNVs, our absence of knowledge of this stems from a lack of systematic study of bacterial CNVs in general. Therefore, the pipelines and frameworks established in this thesis are a blueprint to apply to other bacterial species and discover similar undiscovered genetic diversity.

6.2. The genome of *B. pertussis* is highly unstable

Whilst Chapter 3 focussed on observing CNVs in a global cohort of isolates, the work of Chapter 4 details the genome diversity within single laboratory cultures of *B. pertussis*.
It was demonstrated that UK54, which had a predicted CNV of copy number 4 (confirmed using qPCR) was composed of cells with a mixture of copy numbers. Ultra-long Nanopore reads were generated and each individual read captured a tandem array of the locus with reads ranging in copy number between 1 and 5. This bypassed the assembly process, which failed to resolve the CNV. Whilst Illumina sequencing can resolve cell-to-cell differences for small genomic loci (such as rapid changes within homopolymeric tracts) this application of Nanopore sequencing resolved such differences in very large genomic loci, up to 80kb long.

Extending this work, I showed that sub-cultures of UK54 could give rise to novel copy numbers of the locus, including an isolate with a copy number of approximately 50. This was evidence that tandem arrays were unstable and demonstrates that the *B. pertussis* genome can rapidly change. I widened the search for SVs to the whole genome and found other sites around the genome were undergoing copy number variation. Furthermore, it could be seen that within one sample of UK76 there was >100 reads that indicated there were at least 6 versions of the same structural variant present, indicating either multiple CNV events or extensive remodelling of a single CNV.

These results make it clear that a consensus sequence, even for a clonally derived population, does not adequately describe the true genetic diversity of the sample. This has far reaching consequences for the study of the bacterial kingdom: it is possible that heterogenous populations of bacterial pathogens exist within the host and that this has an impact on the disease. To investigate this, there must be a greater appreciation of within-sample diversity rather than generating a single consensus sequence for the sample. This can be achieved using genome graphs. A consensus can be thought of as just one of the possible genotypes present in the population whereas the construction of genome graphs takes into consideration all observed genotypes and their frequency within the population and includes the study of polymeric tracts.

6.3. Homoplasic structural variants as genotypes for GWAS

Deletions, like CNVs, were found to also follow a network structure and had less variable gene content with denser connections. In Chapter 5, analysis of 2709 *B. pertussis* isolates for deletions showed that there were 474 deletions found at 50 loci represented by 50 network graphs. It was found that the majority of deletions were contained in 22 networks which contained 3 or more deletions.

Mapping both CNVs and deletions to the phylogenetic tree suggested that these mutations were homoplasic. Deletions occasionally appeared in disparate parts of the tree whilst CNVs almost always did, reflecting their very limited heritability. This was confirmed using ancestral state reconstruction, which confirmed that deletions occurred multiple times.

Which regions of the genome are deleted in *B. pertussis* is thought to be driven by both selection and genetic drift. The homoplasies described in Chapter 5, however, show that at least for the most frequent deletions, selection is an important driver of this process, given that isolates showed convergent evolution-a hallmark of selection. Further work to verify the balance of infrequent deletion events to homoplasic deletions will shed light on the selection pressures facing *B. pertussis* as high purifying selection would favour homoplasies and low purifying selection would favour infrequent mutation events.

The impact that specific deletions or CNVs have on *B. pertussis* is unclear and as such, a systematic framework to investigate this is needed. GWAS is a promising framework to find genotype-phenotype links, but mainstream tools do not natively support this genotype and as such I trialled using K-mers to represent CNVs. I found that K-mer analysis was limited but that it could be performed easily for CNV resolved assemblies, which are rare at the time of writing. I found it was computationally infeasible to analyse K-mer abundance in raw reads at a usable scale. GWAS works best when the causal mutation of a phenotype is present in many genetic backgrounds but as *B. pertussis* does not undergo frequent HGT and is highly clonal, this may have posed a significant hurdle. The work outlined throughout this thesis, however, shows that deletions and CNVs are homoplasic and occur in multiple backgrounds. I therefore have shown that these would be good material for GWAS and future work should include a GWAS to phenotype deletions and CNVs in *B. pertussis.*

7. References

1. Pittet LF, Emonet S, Schrenzel J, Siegrist CA, Posfay-Barbe KM. Bordetella holmesii: An under-recognised Bordetella species [Internet]. Vol. 14, The Lancet Infectious Diseases. Lancet Publishing Group; 2014. p. 510–9. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1473309914700210
2. Kilgore PE, Salim AM, Zervos MJ, Schmitt HJ. Pertussis : Microbiology , Disease , Treatment , and Prevention. Clin Microbiol Rev. 2016 Jul 1;29(3):449–86.
3. Mattoo et al. Clinical Manifestations of Respiratory Infections Due to Bordetella pertussis and Other Bordetella Subspecies Molecular Pathogenesis , Epidemiology , and Clinical Manifestations of Respiratory Infections Due to Bordetella pertussis and Other Bordetella Su. Clin Microbiol Rev. 2005;18(2):326–82.
4. Nieves DJ, Singh J, Ashouri N, McGuire T, Adler-Shohet FC, Arrieta AC. Clinical and laboratory features of pertussis in infants at the onset of a California epidemic. J Pediatr. 2011 Dec;159(6):1044–6.
5. Heininger U, Stehr K, Cherry JD. Serious pertussis overlooked in infants. Eur J Pediatr. 1992 May;151(5):342–3.
6. Immunisation Department P. Guidelines for the Public Health Management of Pertussis in England [Internet]. 2018. Available from: www.facebook.com/PublicHealthEngland
7. Yui M, Chow K, Khandaker G, Mcintyre P. Global Childhood Deaths From Pertussis: A Historical Review. 2016;
8. Christie CDC, Baltimore RS. Pertussis in Neonates. Am J Dis Child. 1989 Oct;143(10):1199–202.
9. Mattoo S, Cherry JD. Molecular Pathogenesis, Epidemiology, and Clinical Manifestations of Respiratory Infections Due to Bordetella pertussis and Other Bordetella Subspecies. Clin Microbiol Rev. 2005;18(2):326–82.
10. Agrawal A, Singh S, Kolhapure S, Kandeil W, Pai R, Singhal T. Neonatal Pertussis, an Under-Recognized Health Burden and Rationale for Maternal Immunization: A Systematic Review of South and South-East Asian Countries. Vol. 8, Infectious Diseases and Therapy. Springer Healthcare; 2019. p. 139–53.
11. Whooping cough - NICE CKS [Internet]. [cited 2020 Mar 17]. Available from:

https://cks.nice.org.uk/whooping-cough#!topicSummary

12. He Q, Mertsola J, Soini H, Skurnik M, Ruuskanen O, Viljanen MK. Comparison of polymerase chain reaction with culture and enzyme immunoassay for diagnosis of pertussis. J Clin Microbiol. 1993 Mar;31(3):642–5.

13. Dragsted DM, Dohn B, Madsen J, Jensen JS. Comparison of culture and PCR for detection of Bordetella pertussis and Bordetella parapertussis under routine laboratory conditions. J Med Microbiol. 2004 Aug;53(8):749–54.

14. Heininger U, Cherry JD, Eckhardt T, Lorenz C, Christenson P, Stehr K. Clinical and laboratory diagnosis of pertussis in the regions of a large vaccine efficacy trial in Germany. Pediatr Infect Dis J. 1993 Jun;12(6):504–9.

15. Xu Z, Wang Z, Luan Y, Li Y, Liu X, Peng X, et al. Genomic epidemiology of erythromycin-resistant Bordetella pertussis in China. Emerg Microbes Infect. 2019 Jan 1;8(1):461–70.

16. Soumana IH, Linz B, Harvill ET. Environmental origin of the genus Bordetella. Front Microbiol. 2017;8(JAN):1–10.

17. Taylor-Mulneix DL, Soumana IH, Linz B, Harvill ET. Evolution of bordetellae from environmental microbes to human respiratory pathogens: Amoebae as a missing link. Front Cell Infect Microbiol. 2017 Dec 11;7(DEC).

18. Diavatopoulos DA, Cummings CA, Schouls LM, Brinig MM, Relman DA, Mooi FR. Bordetella pertussis, the Causative Agent of Whooping Cough, Evolved from a Distinct, Human-Associated Lineage of B. bronchiseptica. PLoS Pathog. 2005 Dec;1(4):e45.

19. Woolfreyt BF, Moody JA, Woolfrey BF, Moody JA. Human infections associated with Bordetella bronchiseptica. Clin Microbiol Rev. 1991 Jul 1;4(3):243–55.

20. Register KB, Ivanov Y V., Harvill ET, Davison N, Foster G. Novel, host-restricted genotypes of Bordetella bronchiseptica associated with phocine respiratory tract isolates. Microbiol (United Kingdom). 2015 Mar 1;161(3):580–92.

21. Register KB, Ivanov Y V., Jacobs N, Meyer JA, Goodfield LL, Muse SJ, et al. Draft Genome Sequences of 53 Genetically Distinct Isolates of *Bordetella bronchiseptica* Representing 11 Terrestrial and Aquatic Hosts: TABLE 1 . Genome Announc. 2015;3(2):e00152-15.

22. Ahuja U, Liu M, Tomida S, Park J, Souda P, Whitelegge J, et al. Phenotypic and genomic analysis of hypervirulent human-associated bordetella bronchiseptica. BMC Microbiol. 2012;12:167.

23. Parkhill J, Sebaihia M, Preston A, Murphy LD, Thomson N, Harris DE, et al. Comparative analysis of the genome sequences of Bordetella pertussis, Bordetella parapertussis and Bordetella bronchiseptica. Nat Genet. 2003 Sep 1;35(1):32–40.

24. Linz B, Ivanov Y V., Preston A, Brinkac L, Parkhill J, Kim M, et al. Acquisition and loss of virulence-associated factors during genome evolution and speciation in three clades of Bordetella species. BMC Genomics. 2016 Sep 30;17(1):767.

25. Tizolova A, Guiso N, Guillot S. Insertion sequences shared by Bordetella species and implications for the biological diagnosis of pertussis syndrome. Eur J Clin Microbiol Infect Dis. 2013 Jan;32(1):89–96.

26. Diop A, Raoult D, Fournier PE. Rickettsial genomics and the paradigm of genome reduction associated with increased virulence. Microbes Infect. 2018 Aug 1;20(7–8):401–9.

27. Aslanabadi A, Ghabili K, Shad K, Khalili M, Sajadi MM. Emergence of whooping cough: Notes from three early epidemics in Persia [Internet]. Vol. 15, The Lancet Infectious Diseases. Lancet Publishing Group; 2015. p. 1480–4. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26298206

28. Dunelmensi RM. Libellus de vita et miraculis S, Godrici heremitae de Finchale.

29. Moulton T. Myrroir or Glasse of Helth.

30. Van Lieburg MJ. De Geschiedenis van de Kindergeneeskunde in Nederland.

31. Willis T. Dr. Willis's Practice.

32. Weston R. Whooping Cough : A Brief History to the 19th Century. 2012;29:329–49.

33. MORSE D, BROTHWELL DR, UCKO PJ. TUBERCULOSIS IN ANCIENT EGYPT. Am Rev Respir Dis. 1964 Oct;90:524–41.

34. Zimmerman MR. Pulmonary and osseous tuberculosis in an Egyptian mummy. Bull New York Acad Med J Urban Heal. 1979;55(6):604–8.

35. Cave A, Tuberculosis AD-BJ of, 1939 undefined. The evidence for the incidence of tuberculosis in ancient Egypt. Elsevier.

36. Daniel VS, Daniel TM. Old Testament Biblical References to Tuberculosis. Clin Infect Dis. 1999 Dec 1;29(6):1557–8.

37. Wagner DM, Klunk J, Harbeck M, Devault A, Waglechner N, Sahl JW, et al. Yersinia pestis and the Plague of Justinian 541 – 543 AD : a genomic analysis. Lancet Infect Dis. 14(4):319–26.

38. Plague and the End of Antiquity: The Pandemic of 541-750 - Google Books [Internet]. [cited 2020 Mar 17]. Available from:

https://books.google.co.uk/books?hl=en&lr=&id=DKhLOd6gGlAC&oi=fnd&pg=PA3&dq=Plague+and+the+end+of+antiquity:+the+pandemic+of+541750&ots=oHrNbTJwDn&sig=NJi2VLohx_CQAFguEEYp6SIzcJE&redir_esc=y#v=onepage&q=Plague and the end of antiquity%3A the pandemic of 541750&f=false

39. Haensch S, Bianucci R, Signoli M, Rajerison M, Schultz M, Kacki S, et al. Distinct clones of Yersinia pestis caused the black death. PLoS Pathog. 2010;6(10).

40. Creighton. History of Epidemics, 2.

41. Human Mortality Database [Internet]. [cited 2020 Mar 17]. Available from: https://www.mortality.org/mp/auth.pl

42. Bart MMJ, Harris SSR, Advanid A, Arakawae Y, Botterof D, V B, et al. Global Population Structure and Evolution of Bordetella pertussis and Their Relationship with Vaccination. MBio. 2014;5(2):1–13.

43. Pepperell CS, Casto AM, Kitchen A, Granka JM, Cornejo OE, Holmes EC, et al. The Role of Selection in Shaping Diversity of Natural M. tuberculosis Populations. Sassetti CM, editor. PLoS Pathog. 2013 Aug 15;9(8):e1003543.

44. Cui Y, Yu C, Yan Y, Li D, Li YY, Jombart T, et al. Historical variations in mutation rate in an epidemic pathogen, Yersinia pestis. Proc Natl Acad Sci U S A. 2013 Jan 8;110(2):577–82.

45. Wong TY, Hall JM, Nowak ES, Boehm DT, Gonyar LA, Hewlett EL, et al. Analysis of the In Vivo Transcriptome of Bordetella pertussis during Infection of Mice. mSphere. 2019;4(2).

46. Amman F, D’Halluin A, Antoine R, Huot L, Bibova I, Keidel K, et al. Primary transcriptome analysis reveals importance of IS elements for the shaping of the transcriptional landscape of Bordetella pertussis. RNA Biol. 2018 Jul 3;6286(7):967–75.

47. Ford Doolittle W, Sapienza C. Selfish genes, the phenotype paradigm and genome evolution. Vol. 284, Nature. Nature Publishing Group; 1980.

48. Moran NA, Plague GR. Genomic changes following host restriction in bacteria. Curr Opin Genet Dev. 2004 Dec;14(6):627–33.

49. Siguier P, Gourbeyre E, Chandler M. Bacterial insertion sequences: Their genomic impact and diversity. FEMS Microbiol Rev. 2014 Sep 1;38(5):865–91.

50. Bentley SD, Parkhill J. Genomic perspectives on the evolution and spread of bacterial pathogens. Vol. 282, Proceedings of the Royal Society B: Biological Sciences. Royal Society of London; 2015.

51. Cummings CA, Bootsma HJ, Relman DA, Miller JF. Species- and Strain-Specific Control of a Complex , Flexible Regulon by Bordetella BvgAS †. 2006;188(5):1775–85.

52. Merkel TJ, Barros C, Stibitz S. Characterization of the bvgR locus of Bordetella pertussis. J Bacteriol. 1998;180(7):1682–90.

53. Merkel TJ, Boucher PE, Stibitz S, Grippe VK. Analysis of bvgR Expression in Bordetella pertussis. J Bacteriol. 2003 Dec;185(23):6902–12.

54. Cotter PA, Jones AM. Phosphorelay control of virulence gene expression in Bordetella. Vol. 11, Trends in Microbiology. Elsevier Ltd; 2003. p. 367–73.

55. van Beek LF, de Gouw D, Eleveld MJ, Bootsma HJ, de Jonge MI, Mooi FR, et al. Adaptation of Bordetella pertussis to the Respiratory Tract. J Infect Dis. 2018 May 25;217(12):1987–96.

56. Coutte L, Huot L, Antoine R, Slupek S, Merkel TJ, Chen Q, et al. The multifaceted RisA regulon of Bordetella pertussis. Sci Rep. 2016 Sep 13;6:32774.

57. Warfel JM, Zimmerman LI, Merkel TJ. Acellular pertussis vaccines protect against disease butfail to prevent infection and transmission ina nonhuman primate model. Proc Natl Acad Sci U S A. 2014 Jan 14;111(2):787–92.

58. Warfel JM, Beren J, Kelly VK, Lee G, Merkel TJ. Nonhuman Primate Model of Pertussis. Infect Immun. 2012 Apr 1;80(4):1530–6.

59. Ratledge C, Dover LG. Iron Metabolism in Pathogenic Bacteria. Annu Rev Microbiol. 2000 Oct;54(1):881–941.

60. Brickman TJ, Armstrong SK. Temporal signaling and differential expression of Bordetella iron transport systems: The role of ferrimones and positive regulators. In: BioMetals. NIH Public Access; 2009. p. 33–41.

61. Leininger E, Roberts M, Kenimer JG, Charles IG, Fairweather N, Novotny P, et al. Pertactin, an Arg-Gly-Asp-containing Bordetella pertussis surface protein that promotes adherence of mammalian cells. Proc Natl Acad Sci U S A. 1991;88(2):345–9.

62. Hovingh ES, Mariman R, Solans L, Hijdra D, Hamstra H-J, Jongerius I, et al. Bordetella pertussis pertactin knock-out strains reveal immunomodulatory properties of this virulence factor. 2018;7(1).

63. Scheller E V., Cotter PA. Bordetella filamentous hemagglutinin and fimbriae: critical adhesins with unrealized vaccine potential. Pathog Dis. 2015;73(8).

64. Zipfel PF, Hallström T, Riesbeck K. Human complement control and complement

evasion by pathogenic microbes - Tipping the balance [Internet]. Vol. 56, Molecular Immunology. 2013. p. 152–60. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0161589013003908

65. Wolfe DN, Goebel EM, Bjornstad ON, Restif O, Harvill ET. The O antigen enables Bordetella parapertussis to avoid Bordetella pertussis-induced immunity. Infect Immun. 2007 Oct;75(10):4972–9.

66. Jongerius I, Schuijt TJ, Mooi FR, Pinelli E. Complement evasion by Bordetella pertussis: implications for improving current vaccines. Vol. 93, Journal of Molecular Medicine. Springer Verlag; 2015. p. 395–402.

67. Marr N, Luu RA, Fernandez RC. Bordetella pertussis Binds Human C1 Esterase Inhibitor during the Virulent Phase, to Evade Complement-Mediated Killing . J Infect Dis. 2007 Feb 15;195(4):585–8.

68. Carbonetti NH. Pertussis toxin and adenylate cyclase toxin: Key virulence factors of Bordetella pertussis and cell biology tools. Vol. 5, Future Microbiology. NIH Public Access; 2010. p. 455–69.

69. Cherry JD. Epidemic Pertussis and Acellular Pertussis Vaccine Failure in the 21st Century. 2015;

70. Chen PE, Shapiro BJ. The advent of genome-wide association studies for bacteria. Vol. 25, Current Opinion in Microbiology. Elsevier Ltd; 2015. p. 17–24.

71. van Gent M, Bart MJ, van der Heide HGJ, Heuvelman KJ, Mooi FR. Small Mutations in Bordetella pertussis Are Associated with Selective Sweeps. Hozbor DF, editor. PLoS One. 2012 Sep 28;7(9):e46407.

72. Prevention C for DC and. Pertussis epidemic - Washington, 2012. Morb Mortal Wkly Rep. 2012;61(28):517–22.

73. Dauer CC. Reported Whooping Cough Morbidity and Mortality in the United States. Public Heal Reports. 1943;58(17):661.

74. Cherry JD, Brunell PA, Golden GS, Karzon DT. Report of the Task Force on Pertussis and Pertussis Immunization—1988. Pediatrics. 1988;81(6).

75. Fine PEM, Clarkson JA. The recurrence of whooping cough: Possible implications for assessment of vaccine efficacy. Lancet. 1982 Mar 20;1(8273):666–9.

76. LAMBERT HJ. EPIDEMIOLOGY OF A SMALL PERTUSSIS OUTBREAK IN KENT COUNTY, MICHIGAN. Public Health Rep. 1965 Apr;80(4):365–9.

77. Linnemann CC, Perlstein PH, Ramundo N, Minton SD, Englender GS, McCormick JB, et al. USE OF PERTUSSIS VACCINE IN AN EPIDEMIC INVOLVING

HOSPITAL STAFF. Lancet. 1975 Sep 20;306(7934):540–3.

78. GORDON JE, HOOD RI. Whooping cough and its epidemiological anomalies. Am J Med Sci. 1951;222(3):333–61.

79. Kurt TL, Yeager AS, Guenette S, Dunlop S. Spread of Pertussis by Hospital Staff. JAMA J Am Med Assoc. 1972 Jul 17;221(3):264–7.

80. Gale JL, Thapa PB, Wassilak SG, Bobo JK, Mendelman PM, Foy HM. Risk of serious acute neurological illness after immunization with diphtheria-tetanus-pertussis vaccine. A population-based case-control study. JAMA. 1994 Jan 5;271(1):37–41.

81. Walker AM, Jick H, Perera DR, Knauss TA, Thompson RS. Neurologic events following diphtheria-tetanus-pertussis immunization. Pediatrics. 1988;81(3):345–9.

82. Shields WD, Nielsen C, Buch D, Jacobsen V, Christenson P, Zachau-Christiansen B, et al. Relationship of pertussis immunization to the onset of neurologic disorders: a retrospective epidemiologic study. J Pediatr. 1988 Nov;113(5):801–5.

83. Robinson RJ. The whooping-cough immunisation controversy. Arch Dis Child. 1981;56:577–80.

84. Heininger U, Kleemann WJ, Cherry JD. A controlled study of the relationship between Bordetella pertussis infections and sudden unexpected deaths among German infants. Pediatrics. 2004 Jul 1;114(1):e9–15.

85. Gangarosa EJ, Galazka AM, Wolfe CR, Phillips LM, Gangarosa RE, Miller E, et al. Impact of anti-vaccine movements on pertussis control: The untold story. Vol. 351, Lancet. Lancet Publishing Group; 1998. p. 356–61.

86. Romanus V, Jonsell R, Bergquist S-O. Pertussis in Sweden after the cessation of general immunization in 1979. Vol. 6, The Pediatr Infect Dis J. Nftf; 1987.

87. Amirthalingam G, Gupta S, Campbell H. Pertussis Immunisation and control in England and Wales, 1957 to 2012: A historical review. Eurosurveillance. 2013;18(38):1–9.

88. Klein NP. Licensed pertussis vaccines in the United States: History and current state. Vol. 10, Human Vaccines and Immunotherapeutics. Landes Bioscience; 2014. p. 2684–90.

89. McGirr A, Fisman DN. Duration of pertussis immunity after DTaP immunization: A meta-analysis. Pediatrics. 2015 Feb 1;135(2):331–43.

90. Skoff Clark TH, Liko J, Zell E, Martin S, Messonnier NE, Sara TA, et al. Waning immunity to pertussis following 5 doses of DTaP. Am Acad Pediatr. 2013;

91. Misegades LK, Winter K, Harriman K, Talarico J, Messonnier NE, Clark TA, et al.

Association of childhood pertussis with receipt of 5 doses of pertussis vaccine by time since last vaccine dose, California, 2010. JAMA - J Am Med Assoc. 2012 Nov 28;308(20):2126–32.

92. Cherry JD. The Resurgence of Pertussis: Facts, Fiction, Myths, and Misconceptions. J Vaccines Vaccin. 2017;08(04).

93. Cherry JD. The science and fiction of the “resurgence” of pertussis [Internet]. Vol. 112, Pediatrics. 2003. p. 405–6. Available from: http://pediatrics.aappublications.org/cgi/doi/10.1542/peds.112.2.405

94. Warfel JM, Zimmerman LI, Merkel TJ. Acellular pertussis vaccines protect against disease but fail to prevent infection and transmission in a nonhuman primate model. Proc Natl Acad Sci U S A. 2014;111(2):787–92.

95. Furuta M, Sin J, Ng ESW, Wang K. Efficacy and safety of pertussis vaccination for pregnant women – a systematic review of randomised controlled trials and observational studies. BMC Pregnancy Childbirth. 2017 Dec 22;17(1):390.

96. Walls T, Graham P, Petousis-Harris H, Hill L, Austin N. Infant outcomes after exposure to Tdap vaccine in pregnancy: An observational study. BMJ Open. 2016 Jan 1;6(1):e009536.

97. Kennedy DA, Read AF. Why does drug resistance readily evolve but vaccine resistance does not? Proc R Soc B Biol Sci. 2017 Mar 29;284(1851).

98. Welch TJ, Verner-Jeffreys DW, Dalsgaard I, Wiklund T, Evenhuis JP, Cabrera JAG, et al. Independent emergence of Yersinia ruckeri biotype 2 in the United States and Europe. Appl Environ Microbiol. 2011 May 15;77(10):3493–9.

99. Martin SW, Pawloski L, Williams M, Weening K, DeBolt C, Qin X, et al. Pertactin-negative bordetella pertussis strains: Evidence for a possible selective advantage. Clin Infect Dis. 2015 Jan 15;60(2):223–7.

100. Barkoff AM, Mertsola J, Pierard D, Dalby T, Hoegh SV, Guillot S, et al. Pertactin-deficient Bordetella pertussis isolates: Evidence of increased circulation in Europe, 1998 to 2015. Eurosurveillance. 2019 Feb 14;24(7).

101. Williams MM, Sen KA, Weigand MR, Skoff TH, Cunningham VA, Halse TA, et al. Bordetella pertussis strain lacking pertactin and pertussis toxin. Emerg Infect Dis. 2016;22(2):319–22.

102. Bouchez V, Brun D, Cantinelli T, Dore G, Njamkepo E, Guiso N. First report and detailed characterization of B. pertussis isolates not expressing pertussis toxin or pertactin. Vaccine. 2009 Oct 9;27(43):6034–41.

103. Barkoff AM, Guiso N, Guillot S, Xing D, Markey K, Berbers G, et al. A rapid ELISA-based method for screening Bordetella pertussis strain production of antigens included in current acellular pertussis vaccines. J Immunol Methods. 2014 Jun 1;408:142–8.

104. Gates I, DuVall M, Ju H, Tondella ML, Pawloski L. Development of a qualitative assay for screening of Bordetella pertussis isolates for pertussis toxin production. Hozbor DF, editor. PLoS One. 2017 Apr 10;12(4):e0175326.

105. Weigand MR, Pawloski LC, Peng Y, Ju H, Burroughs M, Cassiday PK, et al. Screening and genomic characterization of filamentous hemagglutinin-deficient Bordetella pertussis. Young VB, editor. Infect Immun. 2018 Jan 22;86(January):IAI.00869-17.

106. Hegerle N, Dore G, Guiso N. Pertactin deficient Bordetella pertussis present a better fitness in mice immunized with an acellular pertussis vaccine. Vaccine. 2014 Nov 20;32(49):6597–600.

107. Jayasundara D, Lee E, Octavia S, Lan R, Tanaka MM, Wood JG. Emergence of pertactin-deficient pertussis strains in Australia can be explained by models of vaccine escape. Epidemics. 2020 Feb 9;100388.

108. Wilk MM, Borkner L, Misiak A, Curham L, Allen AC, Mills KHG. Immunization with whole cell but not acellular pertussis vaccines primes CD4 TRM cells that sustain protective immunity against nasal colonization with Bordetella pertussis. Emerg Microbes Infect. 2019 Jan 1;8(1):169–85.

109. Sigera S, Perera J, Rasarathinam J, Samaranayake D, Ediriweera D, Arav-Boger R, et al. Seroprevalence of Bordetella pertussis specific Immunoglobulin G antibody levels among asymptomatic individuals aged 4 to 24 years: a descriptive cross sectional study from Sri Lanka. BMC Infect Dis. 2016;16(1):729.

110. Cherry JD. Adult pertussis in the pre- and post-vaccine eras; lifelong vaccine-induced immunity? Expert Rev Vaccines. 2014;0584(November):1–8.

111. Caro V, Bouchez V, Guiso N. Is the Sequenced Bordetella pertussis strain Tohama I representative of the species? J Clin Microbiol. 2008 Jun 1;46(6):2125–8.

112. Bart CJ, Zeddeman MJ, Van Der Heide AGJ, Heuvelman H, Van Gent K, Mooi MR. Complete genome sequences of Bordetella pertussis isolates B1917 and B1920, representing two predominant global lineages. Genome Announc. 2014;2(6):1301–15.

113. Duchêne S, Holt KE, Weill F-X, Le Hello S, Hawkey J, Edwards DJ, et al. Genome-scale rates of evolutionary change in bacteria.

114. Sealey KL, Harris SR, Fry NK, Hurst LD, Gorringe AR, Parkhill J, et al. Genomic

analysis of isolates from the United Kingdom 2012 pertussis outbreak reveals that vaccine antigen genes are unusually fast evolving. J Infect Dis. 2015;212(2):294–301.

115. Etskovitz H, Anastasio N, Green E, May M. Role of Evolutionary Selection Acting on Vaccine Antigens in the Re-Emergence of Bordetella Pertussis. Diseases. 2019 Apr 16;7(2):35.

116. Rocha EPC, Smith JM, Hurst LD, Holden MTG, Cooper JE, Smith NH, et al. Comparisons of dN/dS are time dependent for closely related bacterial genomes. J Theor Biol. 2006 Mar 21;239(2):226–35.

117. Octavia S, Sintchenko V, … GG-J of I, 2012 undefined. Newly Emerging Clones of Bordetella pertussis Carrying prn2 and ptxP3 Alleles Implicated in Australian Pertussis Epidemic in 2008–2010. academic.oup.com.

118. Mooi FR, van Loo IHM, Van Gent M, He Q, Bart MJ, Heuvelman KJ, et al. Bordetella pertussis strains with increased toxin production associated with pertussis resurgence. Emerg Infect Dis. 2009 Aug;15(8):1206–13.

119. Petersen R, Dalby T, … DD-E infectious, 2012 undefined. Temporal trends in Bordetella pertussis populations, Denmark, 1949–2010. ncbi.nlm.nih.gov.

120. King AJ, van Gorkom T, van der Heide HGJ, Advani A, van der Lee S. Changes in the genomic content of circulating Bordetella pertussis strains isolated from the Netherlands, Sweden, Japan and Australia: adaptive evolution or drift? BMC Genomics. 2010 Jan 26;11(1):64.

121. Rasko DA, Rosovitz MJ, Myers GSA, Mongodin EF, Fricke WF, Gajer P, et al. The pangenome structure of Escherichia coli: Comparative genomic analysis of E. coli commensal and pathogenic isolates. J Bacteriol. 2008 Oct;190(20):6881–93.

122. Long SW, Linson SE, Ojeda Saavedra M, Cantu C, Davis JJ, Brettin T, et al. Whole-Genome Sequencing of Human Clinical Klebsiella pneumoniae Isolates Reveals Misidentification and Misunderstandings of Klebsiella pneumoniae , Klebsiella variicola , and Klebsiella quasipneumoniae . mSphere. 2017 Aug 2;2(4).

123. Holt KE, Wertheim H, Zadoks RN, Baker S, Whitehouse CA, Dance D, et al. Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in Klebsiella pneumoniae, an urgent threat to public health. Proc Natl Acad Sci U S A. 2015 Jul 7;112(27):E3574–81.

124. Andersson DI, Hughes D. Gene Amplification and Adaptive Evolution in Bacteria. Annu Rev Genet. 2009 Dec 17;43(1):167–95.

125. Pettersson ME, Sun S, Andersson DI, Berg OG. Evolution of new gene functions:

simulation and analysis of the amplification model. Genetica. 2009 Apr 22;135(3):309–24.

126. Michel B, Grompone G, Florès MJ, Bidnenko V. Multiple pathways process stalled replication forks [Internet]. Vol. 101, Proceedings of the National Academy of Sciences of the United States of America. 2004. p. 12783–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15328417

127. Kuzminov A. Recombinational Repair of DNA Damage inEscherichia coli and Bacteriophage λ. Microbiol Mol Biol Rev. 1999;63(4):751–813.

128. Rocha EPC, Cornet E, Michel B. Comparative and evolutionary analysis of the bacterial homologous recombination systems. Vol. 1, PLoS Genetics. Public Library of Science; 2005. p. 0247–59.

129. Morita R, Nakane S, Shimada A, Inoue M, Iino H, Wakamatsu T, et al. Molecular Mechanisms of the Whole DNA Repair System: A Comparison of Bacterial and Eukaryotic Systems. J Nucleic Acids. 2010;2010.

130. Weigand MR, Peng Y, Loparev V, Batra D, Bowden KE, Burroughs M, et al. The History of *Bordetella pertussis* Genome Evolution Includes Structural Rearrangement. J Bacteriol. 2017;(February):JB.00806-16.

131. Weigand MR, Peng Y, Batra D, Burroughs M, Davis JK, Knipe K, et al. Conserved Patterns of Symmetric Inversion in the Genome Evolution of *Bordetella* Respiratory Pathogens. Gilbert JA, editor. mSystems. 2019 Nov 19;4(6).

132. Tondella ML, Williams MM, Weigand MR, Peng Y, Loparev V, Johnson T, et al. Complete Genome Sequences of Four Bordetella pertussis Vaccine Reference Strains from Serum Institute of India. 2016 Dec 22;4(6).

133. Smith GR. Conjugational recombination in E. coli: myths and mechanisms. Cell. 1991 Jan 11;64(1):19–27.

134. Lorenz MG, Wackernagel W. Bacterial gene transfer by natural genetic transformation in the environment. Vol. 58, Microbiological Reviews. 1994. p. 563–602.

135. Kuzminov1 A, Schabtach2 E, Stahl1 FW. X sites in combination with RecA protein increase the survival of linear DNA in Escherichia coli by inactivating exoV activity of RecBCD nuclease. Vol. 13, The EMBO Journal. 1994.

136. Lovett ST, Clark AJ. Genetic analysis of the recJ gene of Escherichia coli K-12. J Bacteriol. 1984 Jan;157(1):190–6.

137. Mendonca VM, Klepin HD, Matson SW. DNA helicases in recombination and repair: Construction of a ΔuvrD ΔhelD ΔrecQ mutant deficient in recombination and repair. J

Bacteriol. 1995 Mar;177(5):1326–35.

138. Dillingham MS, Kowalczykowski SC. RecBCD Enzyme and the Repair of Double-Stranded DNA Breaks. Microbiol Mol Biol Rev. 2008 Dec 1;72(4):642–71.

139. Michel B, Leach D. Homologous Recombination—Enzymes and Pathways. EcoSal Plus. 2012 Nov 29;5(1).

140. Morimatsu K, Kowalczykowski SC. RecFOR proteins load RecA protein onto gapped DNA to accelerate DNA strand exchange: a universal step of recombinational repair. Mol Cell. 2003 May;11(5):1337–47.

141. Webb BL, Cox MM, Inman RB. Recombinational DNA repair: the RecF and RecR proteins limit the extension of RecA filaments beyond single-strand DNA gaps. Cell. 1997 Oct 31;91(3):347–56.

142. Lusetti SL, Cox MM. The Bacterial RecA Protein and the Recombinational DNA Repair of Stalled Replication Forks. Annu Rev Biochem. 2002 Jun;71(1):71–100.

143. Kowalczykowski SC. Initiation of genetic recombination and recombination-dependent replication. Trends Biochem Sci. 2000 Apr;25(4):156–65.

144. Michel B, Boubakri H, Baharoglu Z, LeMasson M, Lestini R. Recombination proteins and rescue of arrested replication forks. DNA Repair (Amst). 2007 Jul 1;6(7):967–80.

145. Lloyd RG, Sharples GJ. Dissociation of synthetic Holliday junctions by E. coli RecG protein. EMBO J. 1993 Jan;12(1):17–22.

146. West SC. PROCESSING OF RECOMBINATION INTERMEDIATES BY THE R uv ABC PROTEINS . Annu Rev Genet. 1997 Dec 28;31(1):213–44.

147. Ives CL, Bott KF. Characterization of chromosomal DNA amplifications with associated tetracycline resistance in Bacillus subtilis. J Bacteriol. 1990 Sep;172(9):4936–44.

148. Anderson RP, Roth JR. Tandem Genetic Duplications in Phage and Bacteria. Annu Rev Microbiol. 1977 Oct;31(1):473–505.

149. Kuzminov A. Homologous Recombination—Experimental Systems, Analysis, and Significance. EcoSal Plus. 2011 Dec 16;4(2).

150. Rocha EPC. An appraisal of the potential for illegitimate recombination in bacterial genomes and its consequences: From duplications to genome reduction. Genome Res. 2003 Jun 1;13(6 A):1123–32.

151. Furuta Y, Kawai M, Yahara K, Takahashi N, Handa N, Tsuru T, et al. Birth and death of genes linked to chromosomal inversion. Pnas. 2011;108(4):1501–6.

152. Ikeda H, Shiraishi K, Ogata Y. Illegitimate recombination mediated by double-strand

break and end-joining in Escherichia coli. Adv Biophys. 2004;38(Complete):3–20.

153. Shuman S, Glickman MS. Bacterial DNA repair by non-homologous end joining [Internet]. Vol. 5, Nature Reviews Microbiology. 2007. p. 852–61. Available from: http://www.nature.com/articles/nrmicro1768

154. Tl~ty TD, Albertini AM, Millepv JH. Gene Amplification in the kc Region of E. coli. Vol. 37, Cell. 1984.

155. Petit MAA, Mesas JMM, Noirot P, Morel-Deville F, Ehrlich SDD. Induction of DNA amplification in the Bacillus subtilis chromosome. EMBO J. 1992 Apr 1;11(4).

156. Roth: Rearrangements of the bacterial chromosome:... - Google Scholar [Internet]. [cited 2020 Mar 18]. Available from: https://scholar-google-com.ezproxy1.bath.ac.uk/scholar_lookup?hl=en&publication_year=1996&author=JR+Roth&author=N+Benson&author=T+Galitski&author=K+Haack&author=JG+Lawrence&author=L.+Miesel&title=Rearrangements+of+the+bacterial+chromosome%3A+formation+and+applications

157. Prozorov AA. Recombinational Rearrangements in Bacterial Genome and Bacterial Adaptation to the Environment. Microbiology. 2001;70(5):501–12.

158. Bergthorsson U, Andersson DI, Roth JR. Ohno's dilemma: Evolution of new genes under continuous selection [Internet]. 2007. Available from: www.pnas.orgcgidoi10.1073pnas.0707158104

159. Ohno S. Evolution by Gene Duplication Springer, New York. 1970;

160. Taylor JS, Raes J. Duplication and Divergence: The Evolution of New Genes and Old Ideas. Annu Rev Genet. 2004 Dec;38(1):615–43.

161. Cairns J, Foster PL. Adaptive reversion of a frameshift mutation in Escherichia coli. Genetics. 1991;128(4):695–701.

162. Roth JR, Kugelberg E, Reams AB, Kofoid E, Andersson DI. Origin of Mutations Under Selection: The Adaptive Mutation Controversy. Annu Rev Microbiol. 2006 Oct;60(1):477–501.

163. Cairns J, Overbaugh J, Miller S. The origin of mutants. Vol. 335, Nature. Nature Publishing Group; 1988. p. 142–5.

164. Rocha EPC. Inference and Analysis of the Relative Stability of Bacterial Chromosomes.

165. Reams AB, Kofoid E, Duleba N, Roth JR. Recombination and annealing pathways compete for substrates in making rrn duplications in Salmonella enterica. Genetics. 2014 Jan 1;196(1):119–35.

166. Sonti R V, Roth JR. Role of gene duplications in the adaptation of Salmonella typhimurium to growth on limiting carbon sources. Genetics. 1989 Sep;123(1):19–28.

167. Kugelberg E, Kofoid E, Reams AB, Andersson DI, Roth JR. Multiple pathways of selected gene amplification during adaptive mutation. Proc Natl Acad Sci U S A. 2006 Nov 14;103(46):17319–24.

168. Esnault E, Valens M, Espéli O, Boccard F, Espe O. Chromosome Structuring Limits Genome Plasticity in Escherichia coli. PLoS Genet. 2007 Dec;3(12):e226.

169. Matthews TD, Maloy S. Fitness effects of replichore imbalance in Salmonella enterica. J Bacteriol. 2010 Nov 15;192(22):6086–8.

170. Couturier E, Rocha EPC. Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes. Mol Microbiol. 2006 Mar;59(5):1506–18.

171. Rocha EPC. The replication-related organization of bacterial genomes. Microbiology. 2004;150(6):1609–27.

172. Rocha EPC, Danchin A. Gene essentiality determines chromosome organisation in bacteria. Nucleic Acids Res. 2003 Nov 15;31(22):6570–7.

173. Trojanowski D, Hołówka J, Ginda K, Jakimowicz D, Zakrzewska-Czerwińska J. Multifork chromosome replication in slow-growing bacteria. Sci Rep. 2017 Mar 6;7.

174. Dennis P. Modulation of Chemical Composition and Other Parameters of the Cell by Growth Rate Molecular evolution of superoxide dismutase genes in halobacteria. View project Ha Bremer Reinier de Graaf Groep [Internet]. researchgate.net. 1996. Available from: https://www.researchgate.net/publication/237130769

175. Rocha EPC. The Organization of the Bacterial Genome. Annu Rev Genet. 2008 Dec 4;42(1):211–33.

176. Brewer BJ. When polymerases collide: Replication and the transcriptional organization of the E. coli chromosome. Vol. 53, Cell. Elsevier; 1988. p. 679–86.

177. Mao X, Zhang H, Yin Y, Xu Y. The percentage of bacterial genes on leading versus lagging strands is influenced by multiple balancing forces. Nucleic Acids Res. 2012;40(17):8210–8.

178. Ring N, Abrahams J, Preston A, Bagby S. Resolving the complex B . pertussis genome with barcoded nanopore sequencing. 2018;4.

179. Udall JA, Dawe RK. Is it ordered correctly? Validating genome assemblies by optical mapping. Vol. 30, Plant Cell. American Society of Plant Biologists; 2018. p. 7–14.

180. Kaiser MD, Davis JR, Grinberg BS, Oliver JS, Sage JM, Seward L, et al. Automated

Structural Variant Verification in Human Genomes using Single-Molecule Electronic DNA Mapping. bioRxiv. 2017 May 22;140699.

181. Brinig MM, Cummings CA, Sanden GN, Lawrence A, Relman DA, Stefanelli P. Significant Gene Order and Expression Differences in Bordetella pertussis Despite Limited Gene Content Variation Significant Gene Order and Expression Differences in Bordetella pertussis Despite Limited Gene Content Variation †. 2006;188(7):2375–82.

182. Cui T, Moro-oka N, Ohsumi K, Kodama K, Ohshima T, Ogasawara N, et al. Escherichia coli with a linear genome. EMBO Rep. 2007 Feb;8(2):181–7.

183. Itaya M, Tsuge K, Koizumi M, Fujita K. Combining two genomes in one cell: Stable cloning of the Synechocystis PCC6803 genome in the Bacillus subtilis 168 genome. Proc Natl Acad Sci U S A. 2005 Nov 1;102(44):15971–6.

184. Caro V, Hot D, Guigon G, Hubans C, Arrivé M, Soubigou G, et al. Temporal analysis of French Bordetella pertussis isolates by comparative whole-genome hybridization. Microbes Infect. 2006 Jul 1;8(8):2228–35.

185. Dalet K, Weber C, Guillemot L, Njamkepo E, Guiso N. Characterization of adenylate cyclase-hemolysin gene duplication in a Bordetella pertussis isolate. Infect Immun. 2004;72(8):4874–7.

186. Dienstbier A, Pouchnik D, Wildung M, Amman F, Hofacker IL, Parkhill J, et al. Comparative genomics of Czech vaccine strains of Bordetella pertussis. Pathog Dis. 2018 Oct 1;76(7).

187. Heikkinen E, Kallonen T, Saarinen L, Sara R, King AJ, Mooi FR, et al. Comparative Genomics of Bordetella pertussis Reveals Progressive Gene Loss in Finnish Strains. Bahn Y-S, editor. PLoS One. 2007 Sep 19;2(9):e904.

188. Velinov M. Genomic Copy Number Variations in the Autism Clinic—Work in Progress. Front Cell Neurosci. 2019 Feb 19;13:57.

189. Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM, et al. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science (80-). 2008 Apr 25;320(5875):539–43.

190. Wissemann WT, Hill-Burns EM, Zabetian CP, Factor SA, Patsopoulos N, Hoglund B, et al. Association of parkinson disease with structural and regulatory variants in the hla region. Am J Hum Genet. 2013 Nov 7;93(5):984–93.

191. Bowtell DD, Böhm S, Ahmed AA, Aspuria PJ, Bast RC, Beral V, et al. Rethinking ovarian cancer II: Reducing mortality from high-grade serous ovarian cancer. Vol. 15, Nature Reviews Cancer. Nature Publishing Group; 2015. p. 668–79.

192. Ewing A, Semple C. Breaking point: The genesis and impact of structural variation in tumours [version 1; referees: 2 approved]. Vol. 7, F1000Research. F1000 Research Ltd; 2018.

193. Yi K, Ju YS. Patterns and mechanisms of structural variations in human cancer. Vol. 50, Experimental and Molecular Medicine. Nature Publishing Group; 2018.

194. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Vol. 7, Nature Reviews Genetics. Nature Publishing Group; 2006. p. 85–97.

195. Ionita-Laza I, Rogers AJ, Lange C, Raby BA, Lee C. Genetic association analysis of copy-number variation (CNV) in human disease pathogenesis. Genomics. 2009 Jan 1;93(1):22–6.

196. Pinto D, Marshall C, Feuk L, Scherer SW. Copy-number variation in control population cohorts.

197. Li H. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics. 2014;30(20):2843–51.

198. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 2011;21(6):974–84.

199. Huang W, Li L, Myers JR, Marth GT. ART: A next-generation sequencing read simulator. Bioinformatics. 2012;28(4):593–4.

200. Plotly Technologies Inc. Collaborative data science [Internet]. Plotly Technologies Inc. . 2015. Available from: https://plot.ly

201. Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal. 2006;Complex Systems:1695.

202. Fruchterman TMJ, Reingold EM. Graph drawing by force-directed placement. Softw Pract Exp. 1991 Nov 1;21(11):1129–64.

203. Stainer DW, Scholte MJ. A Simple Chemically Defined Medium for the Production of Phase I Bordetella pertussis. J Gen Microbiol. 1970 Oct 1;63(2):211–20.

204. Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference.

205. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014 May 1;30(9):1312–3.

206. Letunic I, Bork P. Interactive Tree Of Life (iTOL): An online tool for phylogenetic tree display and annotation. Bioinformatics. 2007;23(1):127–8.

207. Crispell J, Balaz D, Gordon SV. Homoplasyfinder: A simple tool to identify

homoplasies on a phylogeny. Microb Genomics. 2019 Jan 1;5(1).

208. Nicoloff H, Hjort K, Levin BR, Andersson DI. The high prevalence of antibiotic heteroresistance in pathogenic bacteria is mainly caused by gene amplification. Nat Microbiol. 2019 Mar 11;4(3):504–14.

209. Hjort K, Nicoloff H, Andersson DI. Unstable tandem gene amplification generates heteroresistance (variation in resistance within a population) to colistin in *Salmonella enterica*. Mol Microbiol. 2016 Oct 1;102(2):274–89.

210. Cerquetti M, Cardines R, Ciofi Degli Atti ML, Giufré M, Bella A, Sofia T, et al. Multiple Capsule Genes in H. influenzae • JID [Internet]. 2005. Available from: https://academic.oup.com/jid/article/192/5/819/802390

211. Mekalanos JJ. Duplication and amplification of toxin genes in Vibrio cholerae. Cell. 1983 Nov 1;35(1):253–63.

212. Lam C, Octavia S, Sintchenko V, Gilbert GL, Lan R. Investigating genome reduction of Bordetella pertussis using a multiplex PCR-based reverse line blot assay (mPCR/RLB). BMC Res Notes. 2014;7(1):727.

213. Weigand MR, Peng Y, Cassiday PK, Loparev VN, Johnson T, Juieng P, et al. Complete Genome Sequences of Bordetella pertussis Isolates with Novel Pertactin-Deficient Deletions. Genome Announc. 2017 Sep 14;5(37):5–6.

214. Weigand MR, Peng Y, Loparev V, Johnson T, Juieng P, Gairola S, et al. Complete Genome Sequences of Four Bordetella pertussis Vaccine Reference Strains from Serum Institute of India. Genome Announc. 2016 Dec 22;4(6):1404–20.

215. Weigand MR, Pawloski LC, Peng Y, Ju H, Burroughs M, Cassiday PK, et al. Screening and Genomic Characterization of Filamentous Hemagglutinin-Deficient *Bordetella pertussis*. Young VB, editor. Infect Immun. 2018 Jan 22;86(4).

216. King AJ, van Gorkom T, van der Heide HGJ, Advani A, van der Lee S. Changes in the genomic content of circulating Bordetella pertussis strains isolated from the Netherlands, Sweden, Japan and Australia: Adaptive evolution or drift? BMC Genomics. 2010 Jan 26;11(1):64.

217. Weigand MR, Peng Y, Loparev V, Batra D, Bowden KE, Cassiday PK, et al. Complete Genome Sequences of Four Different *Bordetella* sp. Isolates Causing Human Respiratory Infections: TABLE 1 . Genome Announc. 2016;4(5):e01080-16.

218. Bowden KE, Weigand MR, Peng Y, Cassiday PK, Sammons S, Knipe K, et al. Genome Structural Diversity among 31 Bordetella pertussis Isolates from Two Recent U.S. Whooping Cough Statewide Epidemics. mSphere. 2016;1(3).

219. Weigand MR, Pawloski LC, Peng Y, Ju H, Burroughs M, Cassiday PK, et al. Screening and genomic characterization of filamentous hemagglutinin-deficient Bordetella pertussis. Young VB, editor. Infect Immun. 2018 Apr 1;86(4):IAI.00869-17.

220. Whooping RUS, Statewide C, Bowden KE, Weigand MR, Peng Y, Cassiday PK, et al. Genome Structural Diversity among 31 Bordetella pertussis Isolates from Two Epidemics. 1(3):1–15.

221. Zhang L, Bai W, Yuan N, Du Z. Comprehensively benchmarking applications for detecting copy number variation. Ioshikhes I, editor. PLOS Comput Biol. 2019 May 28;15(5):e1007069.

222. Darling AE, Mau B, Perna NT. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. Stajich JE, editor. PLoS One. 2010 Jun 25;5(6):e11147.

223. Domenech P, Rog A, Moolji J, Radomski N, Fallow A, Leon-Solis L, et al. Origins of a 350-kilobase genomic duplication in Mycobacterium tuberculosis and its impact on virulence. Infect Immun. 2014 Jul;82(7):2902–12.

224. Weiner B, Gomez J, Victor TC, Warren RM, Sloutsky A, Plikaytis BB, et al. Independent Large Scale Duplications in Multiple M. tuberculosis Lineages Overlapping the Same Genomic Region. 2012 Feb 7;7(2):e26038.

225. Hoffman CL, Gonyar LA, Zacca F, Sisti F, Fernandez J, Wong T, et al. Bordetella pertussis Can Be Motile and Express Flagellum-Like Structures. MBio. 2019 Jun 14;10(3):e00787-19.

226. Nakamura MM, Liew S-Y, Cummings CA, Brinig MM, Dieterich C, Relman DA. Growth phase- and nutrient limitation-associated transcript abundance regulation in Bordetella pertussis. Infect Immun. 2006 Oct 1;74(10):5537–48.

227. Archer CD, Elliott T. Transcriptional control of the nuo operon which encodes the energy-conserving NADH dehydrogenase of Salmonella typhimurium. J Bacteriol. 1995 May;177(9):2335–42.

228. Scheller E V, Melvin JA, Sheets AJ, Cotter PA. Cooperative roles for fimbria and filamentous hemagglutinin in Bordetella adherence and immune modulation. MBio. 2015 May 26;6(3):e00500-15.

229. Gonyar LA, Gelbach PE, McDuffie DG, Koeppel AF, Chen Q, Lee G, et al. In Vivo Gene Essentiality and Metabolism in Bordetella pertussis . mSphere. 2019 May 22;4(3).

230. Trost B, Walker S, Wang Z, Thiruvahindrapuram B, MacDonald JR, Sung WWLL, et al. A Comprehensive Workflow for Read Depth-Based Identification of Copy-Number Variation from Whole-Genome Sequence Data. 2018 Jan 4;102(1):142–55.

231. Butler JL, Osborne Locke ME, Hill KA, Daley M. HD-CNV: hotspot detector for copy number variants. Bioinformatics. 2013 Jan 15;29(2):262–3.

232. Locke MEO, Milojevic M, Eitutis ST, Patel N, Wishart AE, Daley M, et al. Genomic copy number variation in Mus musculus. BMC Genomics. 2015 Jul 4;16(1):1–19.

233. Kasak L, Rull K, Vaas P, Teesalu P, Laan M. Extensive load of somatic CNVs in the human placenta. Sci Rep. 2015 Feb 10;5(1):1–10.

234. Nicholson TL, Conover MS, Deora R. Transcriptome Profiling Reveals Stage-Specific Production and Requirement of Flagella during Biofilm Development in Bordetella bronchiseptica. PLoS One. 2012 Nov;7(11).

235. Partridge JD, Harshey RM. More than motility: Salmonella flagella contribute to overriding friction and facilitating colony hydration during swarming. J Bacteriol. 2013 Mar;195(5):919–29.

236. Fahrner KA, Berg HC. Mutations that stimulate flhDC expression in Escherichia coli K-12. J Bacteriol. 2015;197(19):3087–96.

237. Yang X, Thornburg T, Suo Z, Jun S, Robison A, Li J, et al. Flagella Overexpression Attenuates Salmonella Pathogenesis. Kaufmann GF, editor. PLoS One. 2012 Oct;7(10):e46828.

238. COTTER PA, Miller JF. Bvgas-Mediated Signal-Transduction - Analysis of Phase-Locked Regulatory Mutants of Bordetella-Bronchiseptica in a Rabbit Model. Infect Immun. 1994;62(8):3381–90.

239. Stewart MK, Cummings LA, Johnson ML, Berezow AB, Cookson BT. Regulation of phenotypic heterogeneity permits Salmonella evasion of the host caspase-1 inflammatory response. Proc Natl Acad Sci U S A. 2011 Dec;108(51):20742–7.

240. Hendrixson DR. A phase-variable mechanism controlling the Campylobacter jejuni FlgR response regulator influences commensalism. Mol Microbiol. 2006 Sep;61(6):1646–59.

241. Otsuka N, Guiso N, Bouchez V. The length of poly(C) stretch in the bordetella pertussis pfim3 promoter determines the vag or vrg function of the fim3 gene. Microbiol (United Kingdom). 2017 Sep;163(9):1364–8.

242. Mooi FR, Avest A ter, Heide HGJ. Structure of the *Bordetella pertussis* gene coding for the serotype 3 fimbrial subunit. FEMS Microbiol Lett. 1990 Jan;66(1–3):327–31.

243. Nagano K, Hasegawa Y, Murakami Y, Nishiyama S, Yoshimura F. FimB regulates FimA fimbriation in porphyromonas gingivalis. J Dent Res. 2010 Sep;89(9):903–8.

244. Imaia K, Ogata Y, Ochiai K. Microbial interaction of periodontopathic bacteria and Epstein-Barr virus and their implication of periodontal diseases. Vol. 54, Journal of Oral Biosciences. Japanese Association for Oral Biology; 2012. p. 164–8.

245. Avalos Vizcarra I, Hosseini V, Kollmannsberger P, Meier S, Weber SS, Arnoldini M, et al. How type 1 fimbriae help Escherichia coli to evade extracellular antibiotics. Sci Rep. 2016 Jan;6(1):1–13.

246. Anderson P, Roth J. Spontaneous tandem genetic duplications in Salmonella typhimurium arise by unequal recombination between rRNA (rrn) cistrons. Proc Natl Acad Sci. 1981 May 1;78(5):3113–7.

247. Patrick WM, Quandt EM, Swartzlander DB, Matsumura I. Multicopy Suppression Underpins Metabolic Evolvability.

248. Anderson RP, Miller CG, Roth JR. Tandem duplications of the histidine operon observed following generalized transduction in Salmonella typhimurium. J Mol Biol. 1976 Aug 5;105(2):201–18.

249. Li WH, Wu CI, Luo CC. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol. 1985 Mar 1;2(2):150–74.

250. Robinson DG, Lee M-C, Marx CJ. OASIS: an automated program for global investigation of bacterial and archaeal insertion sequences. Nucleic Acids Res. 2012 Dec 1;40(22):e174–e174.

251. Yang F, Yang J, Zhang X, Chen L, Jiang Y, Yan Y, et al. Genome dynamics and diversity of Shigella species, the etiologic agents of bacillary dysentery. Nucleic Acids Res. 2005;33(19):6445–58.

252. Ding W, Baumdicker F, Neher RA. panX: pan-genome analysis and exploration. Nucleic Acids Res. 2018 Jan 9;46(1):e5.

253. Ekblom R, Smeds L, Ellegren H. Patterns of sequencing coverage bias revealed by ultra-deep sequencing of vertebrate mitochondria. BMC Genomics. 2014 Jun 12;15(1):467.

254. Tyler AD, Christianson S, Knox NC, Mabon P, Wolfe J, Van Domselaar G, et al. Comparison of Sample Preparation Methods Used for the Next-Generation Sequencing of Mycobacterium tuberculosis. Supply P, editor. PLoS One. 2016 Feb 5;11(2):e0148676.

255. Zhou Y, Bizzaro JW, Marx KA. Homopolymer tract length dependent enrichments in functional regions of 27 eukaryotes and their novel dependence on the organism DNA (G+C)% composition. BMC Genomics. 2004 Dec 14;5:95.

256. Andersson DI, Susan Slechta E, Roth JR. Evidence that gene amplification underlies adaptive mutability of the bacterial lac operon. Science (80-). 1998 Nov 6;282(5391):1133–5.

257. Payne A, Holmes N, Rakyan V, Loose M. Whale watching with BulkVis: A graphical viewer for Oxford Nanopore bulk fast5 files.

258. Quick J. Ultra-long read sequencing protocol for RAD004. Protoc Io. 2018 Jan 22;1–16.

259. MacArthur I, Belcher T, King JD, Ramasamy V, Alhammadi M, Preston A. The evolution of *Bordetella pertussis* has selected for mutations of *acr* that lead to sensitivity to hydrophobic molecules and fatty acids. Emerg Microbes Infect. 2019 Jan 10;8(1):603–12.

260. Xu Y, Lewandowski K, Lumley S, Pullan S, Vipond R, Carroll M, et al. Detection of viral pathogens with multiplex Nanopore MinION sequencing: be careful.

261. Eccles D, White R, Pellefigues C, Ronchese F, Lamiable O. Investigation of chimeric reads using the MinION. F1000Research. 2017;6.

262. Li H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences.

263. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. Phillippy AM, editor. PLOS Comput Biol. 2017 Jun 8;13(6):e1005595.

264. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: interactive visualization of de novo genome assemblies.

265. Services M. UK Standards for Microbiology Investigations Culture of Specimens for Bordetella pertussis and. 2014;1–17.

266. Haack KR, Roth JR. Recombination between chromosomal IS200 elements supports frequent duplication formation in Salmonella typhimurium. Genetics. 1995 Dec;141(4):1245–52.

267. Edlund T, Normark S. Recombination between short DNA homologies causes tandem duplication. Nature. 1981 Jul 1;292(5820):269–71.

268. Adler M, Anjum M, Berg OG, Andersson DI, Sandegren L. High fitness costs and instability of gene duplications reduce rates of evolution of new genes by duplication-

divergence mechanisms. Mol Biol Evol. 2014;31(6):1526–35.

269. Karst SM, Ziels RM, Kirkegaard RH, Albertsen M. Enabling high-accuracy long-read amplicon sequences using unique molecular identifiers and Nanopore sequencing. bioRxiv. 2019 Jan 11;645903.

270. Li H. Minimap2: fast pairwise alignment for long DNA sequences. 2017;1–3.

271. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM [Internet]. 2013. Available from: http://github.com/lh3/bwa.

272. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001 Feb 15;409(6822):860–921.

273. Viehweger A, Krautwurst S, Lamkiewicz K, Madhugiri R, Ziebuhr J, Hölzer M, et al. Direct RNA nanopore sequencing of full-length coron-avirus genomes provides novel insights into structural variants and enables modification analysis.

274. Warwick-Dugdale J, Solonenko N, Moore K, Chittick L, Gregory AC, Allen MJ, et al. Long-read viral metagenomics captures abundant and microdiverse viral populations and their niche-defining genomic islands. PeerJ. 2019 Apr 25;7:e6800.

275. Lees JA, Galardini M, Bentley SD, Weiser JN, Corander J, Stegle O. pyseer: a comprehensive tool for microbial pangenome-wide association studies.

276. samtools-depth(1) manual page [Internet]. [cited 2020 Mar 14]. Available from: http://www.htslib.org/doc/samtools-depth.html

277. Fitch WM. Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. Syst Zool. 1971 Dec;20(4):406.

278. Farris JS. THE RETENTION INDEX AND THE RESCALED CONSISTENCY INDEX. Cladistics. 1989 Dec;5(4):417–9.