Nanopore Sequencing and Phylogenetic Analysis of *Treponema pallidum* from Clinical Specimens in Manitoba.

by

Adam Hedley

A Thesis submitted to the

Faculty of Graduate and Postdoctoral Studies

of the University of Manitoba

In partial fulfillment of the requirements for the degree of

Master of Science

Department of Medical Microbiology and Infectious Diseases

University of Manitoba

Winnipeg

Abstract

Syphilis is a systemic infection caused by the bacteria *Treponema pallidum* subspecies *pallidum* (TPA) that became largely controlled in Canada following the introduction of penicillin treatment in the mid 20th century. Up until the 2000s, syphilis incidence remained below 1 case per 100,00 people but since then there has been a nationwide resurgence. Manitoba has some of the highest rates in the country, reaching 136.4 cases per 100,000 people in 2022, nearly four times the national average. This epidemic has also seen a demographic shift, with women now representing more than 50% of new cases along with the re-emergence of congenital syphilis. Given these rates and the return of previously rare outcomes, understanding the genomic diversity of TPA could enhance surveillance, identify transmission networks and guide public health responses.

This project established a workflow for sequencing TPA directly from clinical swabs collected in Manitoba using Oxford Nanopore Technology (ONT). Screening specimens with a *tp47* PCR and implementation of a multiplex Lesion Panel assay improved diagnostic efficiency and enabled prioritization of samples for downstream genomic analysis. Extraction and amplification methods were evaluated, with total nucleic acid extraction and selective whole-genome amplification (SWGA) providing the best recovery of TPA DNA. Sequencing optimizations enhanced read depth and assembly quality by incorporating adaptive sampling, flow cell reloading and sample multiplexing. Reference-guided assembly and quality assessment produced thirteen high-quality Manitoba genomes.

The results of the phylogenetic analysis aligned with a previous study that included specimens from British Columbia and Alberta, Manitoba genomes clustered within the SS14 lineage. Further *in-silico* analysis confirmed the genetic stability of key diagnostic targets and the presence of macrolide resistance. Overall, this study demonstrates the feasibility of producing TPA genomes from metagenomic clinical samples and highlights the potential of Nanopore sequencing for the genomic surveillance of syphilis in Canada.

Acknowledgments

Firstly, I would like to express my gratitude to Dr. Derek Stein for his willingness to take me on as a student. Thank you for your guidance, encouragement, and patience throughout this project. Your mentorship and confidence in me have helped shape not only this project, but my growth as a scientist.

I am equally thankful to my co-supervisor, Dr. Lyle Mckinnon, and Committee members Dr. Souradet Shaw and Dr. Jared Bullard. Their support and constructive feedback provided valuable direction for this project and their enthusiasm for my work was a source of encouragement.

My sincere thanks go to the team at Cadham Provincial Laboratory, especially Dr. Kerry Dust, for her trust and confidence in my abilities, as well as her constant encouragement and mentorship in both my research and career development. I am also grateful to Dr. Ayo Bolaji for her guidance and knowledge in helping me navigate bioinformatics and computational analysis. I would also like to thank Dr. Paul Van Caeseele for his ongoing support in pursuing this degree, and the Virus Department staff for the incredible work they do and for saving me from performing countless DNA extractions while screening thousands of samples.

I am grateful to the Department of Medical Microbiology and Infectious Diseases for accepting me into the program and the dedication and support they have for students. In particular, I want to thank Angela Nelson for ensuring all forms, paperwork and milestones were completed properly, and for keeping me on track toward graduation. I am also

thankful to Dr. Denice Bay and my MMIC 7050 classmates for giving me a place to develop my presentation skills and improve my scientific communication.

I would also like to thank Dr. Raymond Tsang for providing the NML Nichols strain, which gave me confidence in my sequencing and phylogenetic analyses.

Finally, I would like to thank my family. Their unwavering love, patience, and support have given me the time and opportunity to pursue this degree. Without my wife Ashley, none of this would have been possible, while my kids, Isla and Milo, provided a light at the end of the tunnel.

Table of Contents

List of tables

List of figures

Abbreviations

TPA	*Treponema pallidum* subspecies *pallidum*
TPE	*Treponema pallidum* subspecies *pertenue*
TEN	*Treponema pallidum* subspecies *endicum*
ONT	Oxford Nanopore Technologies
PCR	Polymerase chain reaction
RFU	Relative fluorescent units
Ct	Cycle threshold
DNA	Deoxyribonucleic Acid
CPL	Cadham Provincial Laboratory
NML	National Microbiology Laboratory
UTM	Universal Transport Media
HSV	Herpes Simplex Virus
VZV	Varicella Zoster Virus
BGB	Human β-globin
qPCR	Quantitative Polymerase chain reaction
dPCR	Digital Polymerase chain reaction
MLST	Multi-locus sequence testing
Indel	Insertion or deletion
MSA	Multiple sequence alignments
WGS	Whole genome sequencing
MDA	Multiple displacement amplification
rpWGA	Repli-g whole genome amplification
SWGA	Selective whole genome amplification
AS	Adaptive Sequencing
STD	Standard Oxford Nanopore Technologies Sequencing
rRNA	Ribosomal ribonucleic acid
SNP	Single nucleotide polymorphism
Tpr	Treponemal repeat protein
SUP	Super accurate basecalling
Bp	Base pair
Kbp	Kilobase pair
GB	Gigabytes
Gb	Gigabases
Gbp	Gigabase pair
MB	Megabytes
Mbp	Megabase pair
kDa	Kilodalton

Contributing authors

Chapter 1: This study was conceptualized and designed by Dr. Derek Stein (DS). Thesis writing was completed by Adam Hedley (AH) and edited by DS.

Chapter 2: Thesis writing was completed by AH and edited by DS. The 5,107 DNA samples used for the screening PCR were prepared by the Virus Detection section of Cadham Provincial Laboratory. Dr. Kerry Dust (KD) performed the Lesion panel validation on the Hologic Panther Fusion. Bioinformatics training and support was provided by Dr. Ayooluwa Bolaji (AB). All remaining experimental work was performed by AH.

Chapter 3: AH performed all data analysis and edited by DS. Thesis writing was completed by AH and edited by DS.

Chapter 4: Thesis writing was completed by AH and edited by DS.

List of Copyright Material / Permission Obtained

Chapter 1: Introduction

1.1. The History of Syphilis

1.1.1. Disputed origins

Syphilis, caused by the bacteria *Treponema pallidum* subsp. *pallidum* (TPA), is a disease whose origin is still debated today. Despite the contested origin, historical records suggest that syphilis emerged in Europe during the late 15th century. This occurred around the same time that Christopher Columbus returned from his expedition in the New World and is the basis for the “Columbian Theory” (1,2). Evidence supporting this theory includes accounts from physicians of the time documenting the arrival of sailors and the emergence of the disease, along with recently conducted genetic studies revealing a close relationship between TPA and other *Treponema* species (1,3).

However, an alternate hypothesis, the "pre-Columbian theory" asserts that before Columbus's voyages syphilis was circulating in Europe but had been misdiagnosed, most commonly as leprosy (1,4). Cases of venereal Leprosy seemed to disappear after the 15th century as the identification of Syphilis became more common (4). Advocates of the Pre-Columbian Hypothesis suggest that *Treponema carateum* arose as the skin-limited disease pinta, causing scaly and non-ulcerative lesions, around 15,000 BC (1,5). This was followed by the divergence and global spread of an ancestral *Treponema pallidum* subsp. *pertenue* (TPE) around 10,000 BC. Then around 7,000 BC, arid climates gave rise to *Treponema pallidum* subsp. *endicum* (TEN). As larger communities developed in Asia, a sexually transmitted strain of TEN evolved into TPA. Initially a mild disease, this strain likely mutated in 15th-century Europe, resulting in the more severe form of syphilis (1,4).

Despite the disputed origin, the theories agree with syphilis rapidly gaining notoriety following its European emergence, particularly during a devastating outbreak among French troops in 1495 during the Italian Wars (1,2). Italy, along with Germany and the UK, blamed the French army for spreading it throughout the country. This led to the name “the French disease”, while at the same time the French coined the term “the Neapolitan disease” and “the Polish disease” was being used in Russia. It caught on that each country would blame their neighbors (6,7). Regardless of naming, the conditions of the war created an ideal environment for transmission. The movements of displaced peoples and travelling soldiers enabled the rapid spread of the initial outbreak, which often progressed to tertiary syphilis within months rather than years, resulting in severe disease with a high mortality rate (1,8).

1.1.2. The first epidemic and following centuries of global spread

The first major syphilis outbreak in Europe happened in 1495 among the troops of Charles VIII, the French King, during his invasion of Naples (1,3). The disease spread quickly among the soldiers and then throughout Europe as they returned to their home countries after fighting. The mobility of mercenary armies and the growing trade networks of the time enabled the swift transmission of the disease (9).

By the 17th century, scholars were documenting the rapid spread of syphilis across, and subsequently beyond, Europe. Reports suggest that the disease likely reached Africa and Asia through maritime expeditions led by explorers like Vasco da Gama (8,10). The trade networks established through European exploration further facilitated the spread of TPA

between continents (11,12). A change in the clinical presentation of syphilis also occurred during this time period. After the initial virulent outbreaks of the 16th century, a more protracted but milder form of the disease gained prevalence. This attenuation was observed by the attending physicians and has been noted in modern historical analyses (8,12).

Through the 17th century, religious reform in Europe began framing syphilis as punishment for immoral behavior. This interpretation contributed to a period of silence and stigma where the disease was met with contempt rather than open discussion or intervention (8). Similar views arose in Asia, where historical records link the severity of syphilis symptoms to the degree of the perceived moral transgression (8,12).

Syphilis had emerged as one of the most significant public health challenges worldwide by the late 18th century. In Europe, the disease was particularly widespread where it was estimated that up to 20% of London's population was infected by the 1770s (13). In East Asia, syphilis reached similarly alarming levels and by 1886 it was considered the most prevalent disease in Korea (14). This forced a turning point in societal attitudes where there was growing interest in understanding syphilis, developing treatments and offering support for those affected.

The 19th and early 20th centuries saw rapid urbanization, global migration and the first World War. These events created densely populated cities and expanded commercial sex work, which enhanced syphilis transmission (12). Public health responses included the regulation of prostitution, awareness campaigns, mercury-based treatments and the

creation of specialized hospitals for venereal diseases (12,15). During this era, syphilis was among the most common diagnoses in hospitalized patients, reinforcing the extent of its clinical and societal burden (12).

In 1905, Fritz Schaudinn and Erich Hoffmann found TPA to be the etiological agent of syphilis (16). However, it would still be several decades before an effective treatment was discovered. By the 1940s, as World War II intensified global movements of people, syphilis remained a major threat to both the military and civilian populations. Efforts to combat the disease took on greater urgency, ultimately leading into clinical trials with penicillin which ended up revolutionizing syphilis treatment and dramatically reducing its global impact (15,17). Unlike earlier therapies, penicillin was well tolerated and highly effective (18).

1.1.3. The discovery of penicillin and decline of syphilis in Canada

With the effectiveness of penicillin, syphilis became manageable and was not the destructive force it once was. Incidence in Canada fell quickly during the 1950's and remained relatively low through to the 1980s, around 12 per 100,000 people (19). Increases were seen in the beginning of the 1980's and were speculated to be caused by endemicity in the MSM community. A subsequent outbreak in Winnipeg, during 1984, was seen primarily in heterosexual groups. Three theories for this situation were put forward; the result of a "bisexual bridge" linking the MSM and heterosexual persons, migration of female sexworkers from Alberta due to economic reasons, or biological changes to the bacteria conferring increased virulence (20).

Outside of those occasional outbreaks, incidence further fell in both male and female populations during the 1990's, with less than 1 per 100,000 people (21). However, the National trends during the following decade would begin to show increasing incidence predominantly among men ages 20-59. This was consistent with what was happening in other Western countries (22).

1.1.4. Recent re-emergence of syphilis in Manitoba

In recent years, Manitoba has experienced a concerning resurgence of syphilis cases, well exceeding the national trends in Canada. In 2022, syphilis rates in the province were 136.4 per 100,000 people while the national average was 36.1 per 100,000 people (23). Another alarming trend, first noticed in Manitoba, was the demographic shift from men who have sex with men, into heterosexual groups and women. By the end of 2021 over half (51.9%) of cases were from women (24).

With the dramatic increases in cases in women the re-emergence of congenital syphilis was inevitable, with Manitoba marking its first congenital syphilis case in over 30 years in 2015 (25). By 2020, 11% of all still births in Winnipeg had maternal syphilis listed as a contributing factor (26). Subsequent provincial surveillance has documented more than 320 infants having probable or confirmed congenital syphilis from 2019 to 2024 (27). Later maternal diagnosis in pregnancy and insufficient treatment prior to delivery have been found to be major risk factors for congenital syphilis (28). Adverse fetal outcomes include stillbirth or neonatal death, while surviving infants may present with neurologic impairment,

bone abnormalities, or deafness (29,30) These outcomes not only directly impact the quality of life for affected children but also impose long-term societal costs (31).

1.2. Overview of *Treponema pallidum*

1.2.1. *T. pallidum* Biology

Treponema pallidum is described as a Gram-negative, spiral-shaped bacteria measuring 6–15 μM in length and belonging to the family Spirochaetaceae (29,32,33). The cell has a thin peptidoglycan layer located between an inner and outer membrane. However, unlike typical Gram-negative bacteria, it has a dearth of outer-membrane proteins and does not possess lipopolysaccharide which contributes to its ability to evade the immune response (34–36). Another distinguishing feature is the presence of endo-flagella, located within the peri-plasmic space. These flagella provide the characteristic corkscrew motion, best viewed using dark-field microscopy, which enables effective tissue penetration and dissemination(34,37).

The genome of T. pallidum was found to be circular genome of 1.14 megabase pairs (Mbp) with 1041 open reading frames (ORFs) (38–40). Earlier annotation suggested that ~55% of the open reading frames (ORFs) encode proteins with known biological functions, while 17% are classified as hypothetical proteins and 28% have no homologs, reflecting its highly specialized and auxotrophic nature (38). Recent transcriptome profiling has confirmed transcription of 98% of these features, including nearly all hypothetical proteins (138 of 146) and all five annotated pseudogenes, providing experimental support for the coding potential of the genome (40). Consistent with its reductive metabolism, TPA can generate ATP through glycolysis but lacks genes encoding enzymes for the electron

transport chain and tricarboxylic acid cycle (38). It is also incapable of producing its own nucleotides, fatty acids and enzyme cofactors, relying on host-derived macromolecules acquired through various transporters (38,41).

1.2.2. The three subspecies of T. pallidum

The species *Treponema pallidum* comprises several closely related subspecies that are morphologically and antigenically similar but have distinct clinical presentations (29,42). These include TPA, which causes venereal syphilis; TPE, the agent of yaws; and TEN, responsible for endemic syphilis (Bejel). Collectively, these infections are referred to as treponematoses (29,41).

Although these subspecies vary in transmission and clinical manifestations, they share over 99.5% genomic identity (41,43,44). The genetic differences that do exist, particularly in genes encoding surface-exposed proteins such as the TPR family, are hypothesized to contribute to variations in host interaction, tissue tropism, and disease progression (41,45). However, because they cannot be routinely continuously cultured in vitro, directed mutagenesis of these genes is not available to confirm their roles in pathogenesis (43).

A significant distinction between the subspecies lies in their modes of transmission. TPA is primarily transmitted through sexual contact, entering the host via intact mucous membranes or microscopic abrasions. In contrast, TPE and TEN are spread through non-venereal routes. TPE spreads via direct skin contact with infectious lesions in humid tropical regions, predominantly affecting children, while TEN is transmitted through non-

sexual skin contact or the sharing of utensils in arid regions (46–48). Despite these differences in transmission, all three subspecies progress through primary, secondary, latent and tertiary disease stages when left untreated. The various stages of treponemal disease have numerous clinical presentations frequently mimicking other diseases. The diagnostic complexity led Sir William Osler to describe syphilis as ”the great imitator” (49,50). However, TPA also has the ability to progress to neurosyphilis and/or cause congenital syphilis, as it can cross the maternal-fetal placental and blood-brain barriers (41,48).

1.2.3. Disease stages of syphilis

The primary stage of syphilis develops 10 to 90 days after infection (typically around three weeks) and is marked at the site of inoculation by the appearance of a chancre (50). The chancre is a painless ulcer and can be found in the anogenital region, oral mucosa or other points of contact (51,52). The chancre heals spontaneously within four to six weeks, even in the absence of treatment (50). However, this resolution does not mean clearance of the infection, as the bacteria will continue to spread systemically.

Following the disappearance of the chancre, secondary syphilis develops three to twelve weeks later signifying hematogenous and lymphatic dissemination of TPA (52). The hallmark of this stage is a generalized rash, which affects the soles and palms, and may or may not be pruritic (50,53). Infectious lesions may also develop as oral ulcers, mucous patches, or raised wart-like nodules called condylomata lata (50,52). In addition to dermatological findings, systemic symptoms such as fever, sore throat, headache,

malaise, myalgia and generalized lymphadenopathy commonly appear (53) The rash and systemic symptoms typically resolve within weeks, at which point the infection enters the latent stage (51).

Latent syphilis is defined by the absence of clinical symptoms despite ongoing infection, detectable only through serologic testing. This stage is further classified based on occurrence as either early latent syphilis, within the first year of infection, or late latent syphilis, which persists beyond one year (52). While individuals in the early latent stage may experience symptomatic relapses, they can still transmit the infection congenitally when asymptomatic (51). Those in the late latent phase remain asymptomatic. However, if left untreated, one-third of individuals will progress to tertiary syphilis, which can manifest years or decades post initial infection (50,52).

Tertiary syphilis is marked by severe, often irreversible complications affecting multiple organ systems (50). One of the most well-known manifestations is gummatous lesions, which present as chronic, granulomatous ulcers capable of causing significant tissue destruction usually affecting the skin, and less frequently bones and soft tissue (51). Cardiovascular involvement can result in myocarditis, coronary vessel disease and syphilitic aortitis, leading to aortic aneurysms (52). Neurosyphilis, though capable of occurring in any stage, is frequently seen in tertiary syphilis, where it presents as tables dorsalis or general paresis (51) The widespread tissue destruction caused by tertiary syphilis historically contributed to severe deformities depicted in artwork and medical

literature for centuries (1). These late-stage manifestations were frequent before the advent of antibiotics but are now rarely observed (1,51).

In addition to its impacts on adults, Syphilis poses a major congenital risk when transmitted vertically from mother to fetus at rates from 66% to 100% (54,55). The highest risk occurs during primary, secondary or early latent syphilis, where TPA crosses the placenta and may lead to stillbirth, pre-term birth, or severe neonatal abnormalities (51,55). Babies born with congenital syphilis can present with a myriad of symptoms including hepatosplenomegaly, bone damage, skin lesions, anemia and a bullous rash (51,56). Early screening and treatment with penicillin during pregnancy are the most effective strategies for preventing congenital syphilis (31). **Figure 1.1** provides an overview of the clinical stages of syphilis and their associated risks of neurosyphilis and congenital transmission.

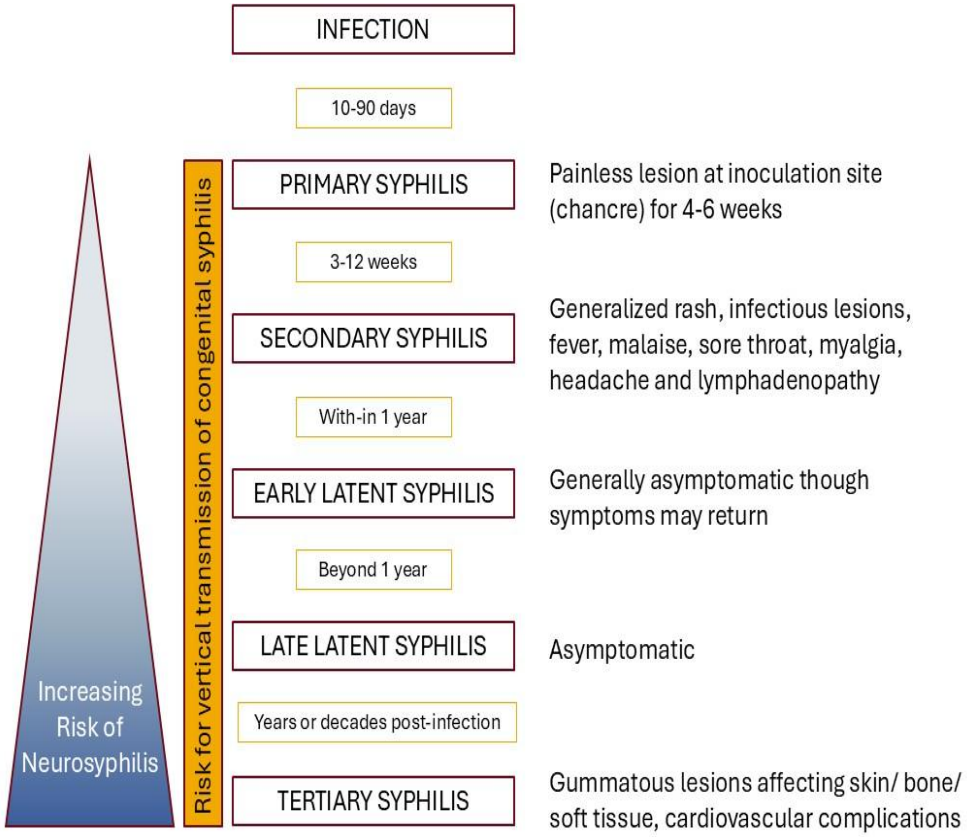


Figure 1.1. Clinical stages of syphilis and associated risks.

The disease progression of syphilis from infection, primary, secondary, latent, and tertiary stages. The risk of vertical transmission is present throughout all stages, while advancing to late disease increases the likelihood of neurosyphilis.

1.2.4. Historical and modern treatments for syphilis

Early treatments for Syphilis were often crude and dangerous. The most widely used treatment during the 16th century was mercury, given as an ointment, ingested, or vaporized. There were questions about the true efficacy, but the side effects of severe toxicity, including neurological and gastrointestinal damage, were common (1,57). Then the early 20th century saw the introduction of arsenic-based compounds like Salvarsan, discovered by Sahachiro Hata and tested by Paul Ehrlich in 1910. This was seen as a significant step forward as it was the first treatment recognized as an effective cure for syphilis (58). Salvarsan was used for the next few decades, though it's drawbacks included complex treatment administration and toxic side effects (59,60).

The discovery of penicillin by Alexander Fleming and its subsequent testing during the 1940s was a turning point in the treatment of syphilis. In 1943, Mahoney and colleagues demonstrated penicillin's ability to effectively treat early syphilis with minimal side effects, outperforming earlier therapies (15,17). After World War II, penicillin was widely available, making syphilis treatment globally accessible. This led to dramatic declines in incidence and mortality. By the mid-20th century, syphilis, once one of the most feared infectious diseases, had become a treatable condition, with penicillin remaining the gold standard of therapy.

Despite the continued effectiveness of penicillin and no documented resistance, azithromycin gained attention in the 1990s as an alternative treatment. It offered a convenient oral alternative to intramuscular benzathine penicillin and was available to

patients with penicillin allergies (29,61). Early studies showed promising results; however, concerns emerged in the 2000s as mass treatment programs appeared less effective, raising questions about azithromycin's reliability in practice (61,62). In 2004, it was conclusively demonstrated that TPA could develop macrolide resistance through point mutations in the 23S rRNA gene, prompting caution in the use of azithromycin in regions with documented resistance (63).

1.2.5. Genomic features of *T. pallidum*

The genome of TPA is compact and specialized, spanning approximately 1.14 Mbp and encoding 1,041 proteins and containing duplicate ribosomal RNA (rRNA) operons. This reduced genome reflects its evolution as an obligate parasite, lacking genes for many biosynthetic pathways, including those for amino acids, fatty acids, and nucleotides (38). Instead, TPA is entirely dependent on the host for essential nutrients, a reliance supported by the presence of numerous transport proteins that comprise nearly 5% of its genome (32). Despite this minimalism, recent analyses suggest that its outer membrane protein (OMP) repertoire is more diverse and functionally redundant than previously recognized, likely facilitating nutrient acquisition across a variety of environments within the host. This may aid pathogen persistence through regulation of surface proteins in response to host environmental pressures (33).

Given the truncated genome, a small number of genes have become common in TPA research, particularly in diagnostics, epidemiology, and understanding pathogenic mechanisms. Among these, *polA*, *tp47*, *arp, tp0470*, and the *tpr* gene family are of

particular importance, **Figure 1.2** (64). The *polA* gene, which encodes DNA polymerase I, has been a valuable diagnostic target due to its unique sequence features that differentiates the *T. pallidum* subspecies from non-pathogenic treponemes and other microorganisms (65).

The *tp47* gene, encoding a 47-kDa membrane-bound protein, has been widely used as a diagnostic PCR target due to its specificity and sensitivity (66–68). It‘s unique to pathogenic treponemes and has no homology to known bacterial or eukaryotic proteins, representing a novel class of penicillin-binding proteins (PBPs) (66,69). Importantly, *tp47* was shown to be the most sensitive PCR target for detecting TPA DNA in clinical specimens, outperforming others (70). Biochemically, the Tp47 protein hydrolyzes penicillin by breaking the β-lactam ring, but the resulting products strongly inhibit further activity by binding to the protein. Fortunately, this effectively limits its β-lactamase function through product inhibition (71). To date, no *tp47* variants conferring penicillin resistance have been identified, but the gene remains an important target for ongoing surveillance to ensure such mutations do not emerge.

The *tp0433* gene, also known as the *acidic repeat protein* (*arp*) gene, is one of the few loci in the TPA genome that exhibits substantial sequence variability across strains, primarily due to differences in the number and composition of internal 60-bp tandem repeats (72). Studies have reported between 4 and 25 repeat copies among *Treponema* species, and up to 22 copies in TPA, resulting in substantial length heterogeneity (73,74). Although its precise function remains unclear, *arp* is routinely used as a molecular typing

marker due to its polymorphic nature (75). However, the repetitive structure of this gene poses challenges for both genome assembly and multiple sequence alignment in short-read datasets (76).

The *tp0470* gene also exhibits substantial length polymorphism due to variation in the number of internal 24-bp tandem repeats (74,77). These repeats encode the highly charged EAEEARRK motif, and repeat counts range from 4 to 29 among TPA genomes, with even higher counts observed in TPE (74). Although the biological role of *tp0470* remains unclear, the gene shows clade-associated repeat patterns, with Nichols-like strains frequently harboring more repeat units than SS14-like strains (74). This clade-specific length variation, coupled with its repetitive sequence, makes *tp0470* a useful target for molecular subtyping and comparative genomic analyses.

The *Treponema pallidum repeat* gene (*tpr*) family, comprises twelve genes that can be grouped into three subfamilies based on sequence homology. The most well studied member is *tprK*, which is important to TPA's capacity for virulence and persistence in infections (78). *tprK* undergoes extensive antigenic variation through non-reciprocal gene conversion with *tprD* among 7 variable sites, providing a significant level of protein diversity, hypothetically improving immune evasion (79). Other *tpr* genes are also implicated in pathogenesis. Members of Subfamily I (*tprC, tprD, tprF* and *tprI)* show extensive variation in the sequences of the predicted surface exposed regions, suggesting roles in host interaction and immune recognition (80). Subfamily II members (*tprE, tprG, tprJ*) have been found to be under transcriptional regulation by the non-tpr

protein TP0262, indicating expression is not static and may be controlled during infection or in response to host-mediated cues (81).

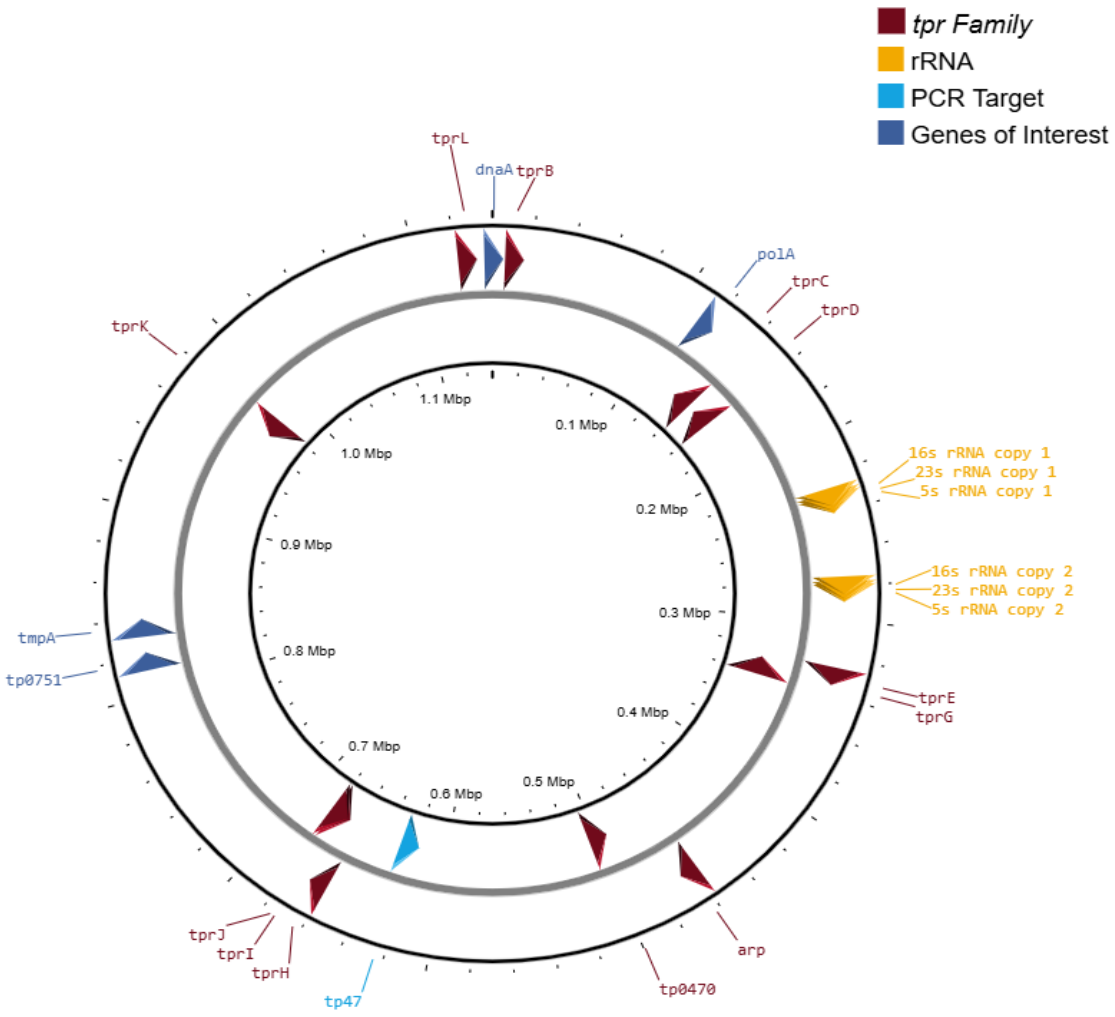


Figure 1.2. Genomic map of *T. pallidum* highlighting genes of interest
Genes are displayed according to their forward (outer track) or reverse (inner track) orientation. Members of the *tpr* gene family are shown in red, the rRNA operons in yellow and selected genes of interest, including PCR or Vaccine targets, in blue. Figure created with Proksee (https://proksee.ca/)

1.3. Molecular Screening of *T. pallidum*

1.3.1. *T. pallidum* Polymerase Chain Reactions

During the late 1980s, the diagnosis of syphilis posed significant challenges: *T. pallidum* could not be isolated in artificial media, rabbit infectivity testing was impractical for routine use, and dark-field microscopy lacked sensitivity (82). With syphilis cases increasing among HIV-positive patients, there was growing concern about the limitations of serological testing for early infections during primary syphilis, as well as the potential for reactivation of latent syphilis in immunosuppressed individuals (82,83). These challenges prompted the development of polymerase chain reaction (PCR) assays, which offered a highly sensitive and specific means of directly detecting treponemal DNA in clinical specimens.

Early PCR assays focused on detecting conserved regions of the *T. pallidum* genome to maximize sensitivity and specificity. Initial targets included the *treponemal membrane protein A* (*tmpA*), *DNA polymerase I* (*polA*), *subsurface lipoprotein 4D* (*4D*), the *basic membrane protein (bmp)* and the *47-kDa membrane protein* (*tp47)* (66,82–84). These targets were chosen because they are conserved across Treponemal species and not present in other common genital pathogens, minimizing the risk of false positives. More recently, the *tp47* and *polA* genes have been the main targets used in TPA PCR (67,85). Comparative studies have shown that both targets provide equivalent sensitivity and specificity in ulcer swabs, confirming their reliability in clinical practice (67).

With improvements in commercial PCR master-mixes there has been a move towards multiplexing the assays to simultaneously detect TPA alongside other pathogens that cause genital ulcer disease, such as *Haemophilus ducreyi* and Herpes Simplex Virus (66,86). This multiplex approach improves diagnostic efficiency in clinical settings where rapid and comprehensive testing is needed. However, a notable limitation of all the previously mentioned assays is that they rely on conserved targets such as *tp47* and *polA*, which cannot distinguish between *T. pallidum* subspecies like TPA, TPE or TEN (82). Because these subspecies share nearly identical sequences in these regions, molecular differentiation requires the use of alternative targets or additional methods, which are not part of routine clinical testing (87).

1.3.2. Multi-Locus Sequence Typing of *T. pallidum*

Multi-locus sequence typing (MLST) was one of the earliest molecular tools used to distinguish the *T. pallidum* subspecies and genotype TPA. With the inability for in vitro culture and limited genetic diversity, MLST schemes were devised around a small number of genomic loci exhibiting variability (88). Initial typing focused on the genes of putative surface-exposed proteins (*tp0136*, *tp0548*, and *tp0705)* and repeat-rich genes (*arp* and the *tpr* family), which show greater sequence heterogeneity across strains (75,89,90).

Despite its utility, even MLST has limitations in accurately differentiating subspecies, particularly when loci have regions susceptible to recombination or convergent evolution. Early typing schemes using *arp*, *tpr*, and *tp0548* helped identify genotypes across global populations, but also led to confusion when isolates displayed conflicting signatures. For

instance, Grange et al. (2013) described an unusual *T. pallidum* isolate (11q/j) from a genital lesion in France that was initially classified as syphilis (91). Upon examination of the published *tp0548* sequence, it was noticed that 11q/j was only a single nucleotide polymorphism (SNP) away from sequences found in TPE and suggested this may be an imported case of yaws (89). Subsequent investigations determined this isolate was a case of sexually transmitted TEN which would be later supported by a Cuban study identifying nine cases of TEN misdiagnosed as syphilis (92,93).

Although the recent decade has seen the emergence of WGS, the simplicity and adaptability of MLST methods have continued to grow, and as late as 2018, optimized schemes were being applied across Europe to monitor the re-emergence and spread of syphilis, allowing international comparisons of TPA strains directly from clinical specimens (94). The results of these MLST studies have extended beyond simple genotyping. Noda et al. (2018) constructed a phylogenetic tree using nine different MLST loci, demonstrating the scheme's potential for understanding broader strain relationships (93).

1.3.3. Whole Genome Sequencing of *T. pallidum*

While MLST and targeted gene studies laid the foundation for understanding the genetic diversity of TPA, these approaches are inherently limited by the small number of loci. The recent advances in next-generation WGS have provided insights into genome structure, population dynamics and evolution.

Until recently, TPA WGS primarily relied on bacterial propagation in rabbit testes to provide sufficient amounts of DNA for Sanger sequencing (38,77). However, this approach limited genetic studies to a small number of laboratory strains and raised concerns about host-adaptive mutations. A series of modern techniques enabled researchers to overcome these hurdles. Strategies such as anti-treponemal antibody enrichment, methyl-directed enrichment, whole genome amplification (WGA), and targeted hybrid-capture have allowed researchers to bypass rabbit passage and directly sequence from clinical TPA samples (76,95–97). These methods have expanded the number of genomes available for analysis and provided insight into the genetic diversity present in modern syphilis.

This early wave of WGS applications has relied exclusively on short-read sequencing technologies, typically Illumina-based platforms (76,96,98,99). While these methods are capable of producing high quality data, they face challenges in resolving repetitive or highly similar genomic regions, such as paralogous *tpr* genes and long tandem repeats (76,100). Nevertheless, they have provided a foundation for global phylogenomic analyses, offered insights into the epidemiology of syphilis outbreaks and identified some mechanisms of antigenic variation.

1.3.4. Phylogenetics of *T. pallidum*

The Nichols strain of TPA, first isolated in 1912 and subsequently maintained through rabbit passage, became the model for syphilis research and was the first TPA genome to be sequenced (38). The SS14 strain became another important reference, particularly

due to its association with macrolide resistance and later complete genome publication. SS14 was isolated from a patient with a penicillin allergy who was not responding to erythromycin treatment in 1977 (77,101). Through accumulated work with MLST and whole genome fingerprinting (WGF), researchers found that TPA samples consistently grouped into two major clades clustering around either Nichols or SS14 (43,102).

Building on those observations, recent studies using WGS have expanded our understanding of TPA phylogeny. A global study by Beale et al. (2021) analyzed more than 700 genomes from 23 countries, confirming the existence of the two clades with distinct sub-lineages (103). It's notable that the SS14 lineage accounts for the majority of contemporary cases worldwide, while the Nichols strains remain less common but are still circulating.

Regional studies have provided further insight into these global patterns. In 2022, analysis of 456 genomes from Australia revealed that multiple sub-lineages spanning both the SS14 and Nichols clades were driving the on-going syphilis epidemic (104). In Japan, nearly all circulating strains belonged to SS14 sub-lineage 1B, with phylogenetic clustering by sexual orientation. This group of closely related strains between Japan and China revealed an East Asian transmission network (105). In Buenos Aires, an unusually high prevalence (37%) of Nichols-like strains were noted, more than global estimates, along with an increase in macrolide resistance present in both TPA strains (106). The results suggested regionally distinct transmission patterns were occurring.(106)

Temporal analyses agree that following the discovery and use of penicillin in the mid-20th century, syphilis incidence declined in many regions. However, the resurgence in the early 2000s was shown to be driven by a few dominant sub-lineages that rapidly expanded after a population bottleneck in the late 1990s (103,107). This increased incidence likely reflects shifts in sexual behavior after the introduction of effective antiretroviral therapies (103). While the diversity of the modern strains is linked to these more recent events, historical phylogenetics indicate that the most recent common ancestor of TPA lineages dates to the 17th century (98). Even with its long history, the modern TPA population shows low overall genetic diversity and slow SNP accumulation, and strains remain highly similar across continents and decades.

1.3.5. Potential Vaccine targets

Not long after TPA was identified as the etiological agent of Syphilis there was a desire to develop a vaccine. In the early 1970's, James Miller was able to provide complete protection from infection, in a rabbit model, for at least 1 year (108). Although his approach was not practical for human immunization, it established a benchmark for evaluating vaccine efficacy in experimental models (109).

The exact mechanism of protection observed in Miller's study remains unknown. However, the family of 12 Tpr proteins have emerged as promising vaccine targets due to their predicted surface localization and porin functions (29,110). Among these, TprK has shown promise in rabbit models as immunization with recombinant TprK results in significantly attenuated lesion development (111) Similarly, recombinant TprF and the N-

terminal conserved region of TprF have also been shown to attenuate lesion development. Unlike the highly variable TprK, there has been a somewhat surprising lack of TprF sequence variation found, potentially making it a better candidate (110).

Other vaccine targets include Tp0751, a protein with dual functions that provides TPA an effective method for dissemination. Tp0751 binds extracellular matrix components, including vascular endothelial cells, improving bacterial adhesion and spread (112). Additionally, it promotes clot dissolution by degrading fibrinogen and laminin, enabling the spirochete to penetrate tissue barriers and circulate more effectively (113). These functions contribute to the virulence of TPA but also make Tp0751 a strong candidate for vaccine development. Multiple studies have demonstrated immunizing rabbits with Tp0751 reduces bacterial organ burden and, when included in a tri-valent vaccine, significantly attenuates the development of chancres (114,115).

Tp0326 (BamA ortholog), part of the β-barrel assembly machinery contains both a periplasmic N-terminal domain and an outer-membrane embedded c-terminal β-barrel domain. Both ends of the protein are targeted by rabbit anti-bodies however, only the periplasmic segment is targeted by human antibodies (116). In Tp0326 immunized rabbits, the animals that had high Tp0326-antibody titers showed some degree of protection in subsequent exposures (117).

1.3.6. Challenges in Studying Syphilis

Research on TPA is hindered by several key limitations due to its biology, genomic complexity, and experimental difficulties. In studies involving bacteria, a common first step

is culture isolation. The small genome of TPA not only results in complex media requirements but also lacks genes encoding the typical enzymes that deal with oxidative stress (38) TPA instead uses proteins hypothesized to have been acquired from hyperthermophilic anaerobic ancestors resulting in the requirement for microaerophilic conditions, around 1.5% O2, in order to maintain viability and growth (41). Even brief exposure to atmospheric oxygen levels can be lethal, necessitating specialized culture environments that mimic mammalian tissues (118,119). Coupled with slow replication, TPA has a doubling time of approximately 30-50 hours, Isolation has been extremely difficult (119–121).

Historically, the only reliable method for propagating TPA involved the use of rabbit models, where the bacteria was maintained through intratesticular inoculation (29,120). However, recent breakthroughs have been demonstrated that build on previous methods of short-term propagation (118,119) In 2018, long-term in vitro culture of TPA was reported using a co-culture system with rabbit epithelial cells (Sf1Ep) and a specialized medium (TpCM-2) in low-oxygen conditions. This system has since been refined to sustain continuous growth for more than three years, with cultures retaining full viability and infectivity (121).

Despite the recent advancements in culture systems they remain highly specialized and are not widely accessible to most laboratories. Consequently, molecular studies typically rely on nucleic acid directly extracted from clinical specimens. This enabled specific and sensitive detection of TPA in cerebrospinal fluid (CSF), during latent and tertiary stages,

using PCR (83,122). However, the low bacterial burden in patient samples severely limits the quantity of recoverable TPA DNA rendering techniques beyond conventional PCR challenging and complicating WGS efforts (123–125).

Differential lysis of samples can be a common technique relying on the strength of the prokaryotic cell wall compared to eukaryotic cell membranes (125–127)However, TPA has been shown to be exceptionally fragile, making laboratory manipulation challenging. The outer membrane is sensitive to mechanical stress induced during centrifugation and non-ionic detergents (36,128). This fragility results in TPA cells being lysed during the same steps that would remove the eukaryotic cells preventing DNA separation.

RNA-bait capture methods have gained popularity when using short-read sequencing platforms. The probes, typically around 120 nucleotides in length, hybridize to fragmented DNA, enabling the enrichment of specific targets (76,103,107,129). This provides an effective means for capturing low-abundance DNA, but the short sequences present a challenge for resolving regions with repetitive genomic elements. For example, the *arp* gene of TPA contains between 4 and 20 copies of 60-base repeats, making it a useful target in MLST schemes. However, when bait capture is used, accurately assembling this region becomes challenging due to the difficulty in correctly mapping repetitive sequences (76). Similarly, bait capture and short-read sequencing struggle to differentiate reads originating from TPA's nearly identical ribosomal RNA operons, which span over 4.8kbp, requiring the use of Sanger Sequencing to resolve the regions (96). This complicates genome assembly by requiring an additional layer of sequencing and

analysis. Due to these assembly challenges, regions are often masked as ambiguous characters (N), limiting the ability to fully characterize TPA's genomic variability. The resulting masked sequences may hinder phylogenetic analyses, genomic surveillance and impede efforts to develop vaccine candidates.

1.3.7. Advancing *T. pallidum* genomics with Oxford Nanopore sequencing

Advances in genomics and DNA sequencing technologies have the potential to significantly improve our knowledge of TPA and Syphilis disease progression. Whole genome sequencing (WGS) can aid in understanding the population structure of the bacteria and the dynamics of infections (10,20,130). Through genome comparisons, it is possible to differentiate strains and identify clusters which allows researchers to uncover patterns of epidemic spread (124).

Genomics also enable the surveillance of antimicrobial resistance (131,132). Although TPA remains susceptible to penicillin, cases of resistance to macrolides have been reported world-wide (103). WGS can identify known mutations, such as those in the 23S rRNA gene, and detect novel ones, allowing the emergence and spread of resistant strains to be tracked (76,103). More recently, the use of doxycycline as post-exposure prophylaxis (doxy-PEP) has shown promising results in reducing bacterial sexually transmitted infections (STI) (133,134). While tetracycline resistance has not yet been documented in TPA, the existence of 23S rRNA mutations proves ribosomal changes imparting resistance can occur. It's already been found in *Neisseria gonorrhoeae*, the V57M mutation in the 30S ribosomal subunit has been shown to provide tetracycline

resistance (135). For that reason, genomic surveillance will be important for monitoring potential doxycycline resistant mutations in the ribosomes.

Additionally, genomic studies have enhanced our ability to identify potential vaccine targets (136). High-throughput sequencing paired with proteomic analyses help uncover conserved and immunogenic proteins, such as outer membrane proteins (OMPs) and treponemal repeat proteins (Tpr), which are important for bacterial survival and immune evasion (96). With the variability seen in these targets, using WGS, it may be necessary to tailor the vaccines to locally circulating strains.

Furthermore, WGS can improve public health surveillance by enabling more precise molecular epidemiology. Genomic data can be used to follow outbreaks in real time, identify high-risk populations, and reveal unknown transmission patterns (130,132,137). Integrating genomic surveillance with traditional epidemiological methods can enhance early detection of outbreaks and facilitate targeted strategies for syphilis control and prevention.

While short-read sequencing has made many of these advances possible, it struggles to resolve repetitive and duplicated elements in the TPA genome. Long-read sequencing with Oxford Nanopore Technologies (ONT) offers specific advantages for addressing these challenges. ONT generates long reads that can span thousands of bases in a single molecule (138) This can resolve problematic genomic regions such as the *arp* and *tp0470* repeat loci, as well as the closely related *tpr* gene family and duplicated rRNA operons, which are often masked in Illumina datasets (97,103,104). By reducing the need to mask

large portions of the genome, long-read sequencing will enable a more complete picture of TPA diversity.

In addition to read length, ONT sequencing offers features well suited to metagenomic clinical specimens, where TPA DNA is often present at very low abundance relative to host DNA (95). Adaptive sampling (AS), a software-based enrichment strategy, enables the selective retention of bacterial reads and active rejection of non-target DNA as sequencing occurs in real time **(Figure 1.3)** (139). This enrichment can increase the proportion of TPA reads several-fold without additional laboratory steps, thereby improving genome coverage from low-input or heavily contaminated samples (139,140). Together, these abilities make Nanopore sequencing uniquely advantageous for furthering TPA genomics and overcoming challenges in syphilis research.

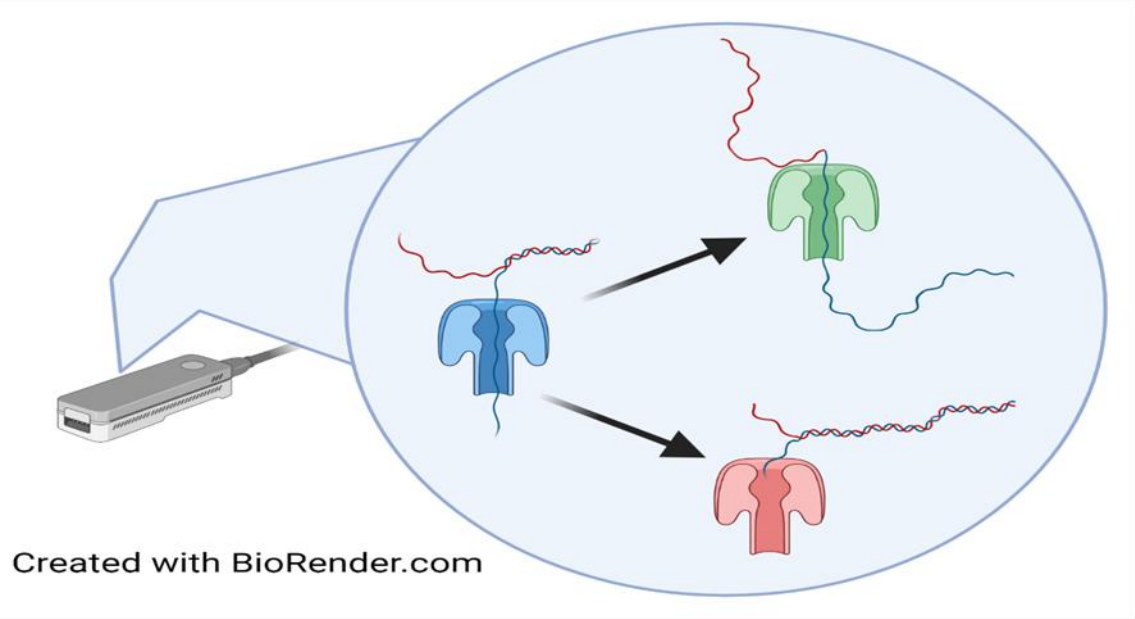


Figure 1.3. Adaptive sampling during Oxford Nanopore sequencing
Adaptive sampling, a real-time enrichment strategy were DNA molecules are evaluated as they enter the nanopore (blue pore). Strands mapping to a user-defined reference genome (TPA) are retained and fully sequenced (green pore), while non-target DNA (host sequences) are rejected and ejected from the pore (red pore). This process increases the proportion of on-target reads.

1.4. Study Rationale and Aims

This study is based on our hypothesis that the SS14 lineage has become the dominant strain of TPA circulating in Manitoba. To test this hypothesis, we established the following objectives:

(1) Identify and collect clinical specimens suitable for TPA analysis

(2) Develop a long-read sequencing protocol for metagenomic samples

(3) Establish a comprehensive bioinformatics pipeline for data processing and genome assembly

(4) Compare long-read assembled TPA genomes from Manitoba to available global genomic data, to investigate local strain diversity and potential epidemiological patterns.

Chapter 2: Materials and Methods

Ethical approval for this study was obtained from the University of Manitoba Health Research Ethics Board (Approval No. HS26285, January 2024). The protocol, Genomic epidemiology and vaccines combined: A comprehensive approach to pre-clinical syphilis vaccine discovery, was designed to describe the genetic epidemiology of *Treponema pallidum* in Manitoba and across the Canadian Prairies. The study aims to sequence *T. pallidum* in retrospective and prospective diagnostic samples to characterize circulating strains and link genomic variation with epidemiological factors. The methods include selective whole genome amplification, Oxford Nanopore long-read sequencing and bioinformatic analysis. Only de-identified metadata (age, sex, specimen type, specimen date) are linked to specimens, and all analyses are conducted at Cadham Provincial Laboratory in collaboration with academic partners.

2.1. PCR screen

Cadham Provincial Laboratory receives mucocutaneous lesion specimens as flocked swabs in Universal Transport Media (UTM) (Becton Dickinson and Company, cat# 220528) as part of routine clinical testing for Herpes Simplex Virus types 1 (HSV1), Herpes Simplex Virus types 2 (HSV2) and Varicella-Zoster Virus (VZV). At the request of the ordering physician, specimens can be further tested for TPA where a specimen aliquot is taken and sent to the National Microbiology Laboratory (NML). All specimens are stored at -80°C and previously identified TPA specimens were used as positive control material for PCR.

2.1.1. Tp47 Screening PCR

Following routine testing for HSV1, HSV2, and VZV, the extracted DNA was screened for TPA (t*p47* gene) using real-time PCR. Reactions were performed in a 20 µl final volume with TaqMan Fast Virus 1-Step Master Mix (Thermo Fisher Scientific, cat# 4444432), containing 5 µl of template DNA, 500nM primers and 250nM probe targeting the Tp47 gene as previously described (70): forward primer 5′-CAACACGGTCCGCTACGACTA-3′, reverse primer 5′-TGCCATAACTCGCCATCAGA-3′, and probe 5′-HEX-CGGTGATGACGCGAGCTACACCA-BHQ1-3′.

The PCR was conducted on a CFX-96 Real-Time System (Bio-Rad Laboratories, cat# C1000) with the following thermal cycling conditions: 50°C for 5 minutes, 95°C for 20 seconds, followed by 40 cycles of 95°C for 5 seconds and 61.4°C for 30 seconds with fluorescence plate reading. Data analysis was carried out using CFX Maestro software (v 2.3). Any specimens that screened positive for TPA, had a cycle threshold (Ct) value <40Ct, were stored for further use.

Synthetic TPA DNA (ATCC, cat# BAA-2642SD) was used as quantified positive control material and serially diluted to produce a standard curve. Starting at an original concentration of 4.7 x 10^5 copies/µL, 50µL was added to 450µL of IDTE pH8.0 1X TE buffer (Integrated DNA Technologies, cat # 11-05-01-13). This was continued producing dilutions of 450µL, from 10^{-1} to 10^{-8}.

2.1.2. Lesion panel

A multiplex real-time PCR was performed on extracted DNA to detect TPA, HSV1, HSV2, VZV, and human Beta-globin (BGB). Each reaction was prepared in a 20 µl volume with TaqMan Fast Advanced Mix (Life Technologies, cat# 4444557), including 5 µl of template DNA. Primers and probes were used to amplify specific genes: the Tp47 gene for TPA, the glycoprotein B (gB) gene for HSV1 and HSV2, the DNA polymerase gene for VZV, and the BGB gene as an endogenous control for human DNA.

For HSV1, the sense primer (5'-GCAGTTTACGTACAACCACATACAGC-3') and the antisense primer (5'-AGCTTGCGGGCCTCGTT-3') targeted the gB gene, with probe 5'-(FAM/ZEN)CGGCCCAACATATCGTTGACATGGC-3'. For HSV2, amplification of the gB gene used the sense primer (5'-TGCAGTTTACGTATAACCACATACAGC-3') with the same antisense primer as HSV1, and probe 5'-(HEX/ZEN)-CGCCCCAGCATGTCGTTCACGT-3'. Detection of TPA was achieved by targeting the Tp47 gene with the forward primer (5'-CAACACGGTCCGCTACGACTA-3'), reverse primer (5'-TGCCATAACTCGCCATCAGA-3'), and probe 5'-(TXRed-XN)CGGTGATGACGCGAGCTACACCA-BHQ2-3'. VZV detection targeted the DNA polymerase gene with primers (5'-CGGCATGGCCCGTCTAT-3') and (5'-TCGCGTGCTGCGGC-3'), and probe 5'-(CY5/TAO)ATTCAGCAATGGAAACACACGACGCC-3'. BGB was amplified as a control using the forward primer (5'-TGGATGAAGTTGGTGGTGAG-3') and reverse primer (5'-CCCAGTTTCTATTGGTCTCCTT-3'), with probe 5'-(TYE705)CCTGGGCAGGTTGGTATCAAGGTT-BHQ2-3'.

Reactions were run on a CFX-96 Real-Time PCR System (Bio-Rad Laboratories) in multiplate 96-well unskirted PCR Plates (Bio-Rad Laboratories, cat# ML9601) sealed with Microseal B film (Bio-Rad Laboratories, cat #MSB1001). Thermal cycling conditions were as follows: initial hold at 50°C for 2 minutes, 95°C for 20 seconds, followed by 40 cycles of 95°C for 3 seconds and 58°C for 30 seconds with plate read. Each run included negative (no template) and positive controls. Data analysis was performed using CFX Maestro software, version 2.3.

Synthetic TPA DNA (ATCC, cat# BAA-2642SD) was used as quantified positive control material and serially diluted to produce a standard curve. Starting at an original concentration of 4.7 x 10^5 copies/µL, 50µL was added to 450µL of IDTE pH8.0 1X TE buffer (Integrated DNA Technologies, cat # 11-05-01-13). This was continued producing dilutions of 450µL, from 10^{-1} to 10^{-8}.

Standard curves generated from Synthetic TPA DNA demonstrated consistent performance across a dynamic range from 5 x 10^4 to 0.5 genome copies per reaction. The regression equation was $Ct = -3.27 \log_{10}(copies) + 34.25$, with an R^2 of 0.999. The average Ct difference between 10-fold dilutions was 3.27 cycles, corresponding to a PCR efficiency of 102.4%.

2.1.3. Digital PCR

To create quantified PCR controls with the previously validated TPA tp47 primers and probe, digital PCR (dPCR) was performed on the Absolute Q dPCR system (Thermo Fisher Scientific, cat# A52864). Using the Absolute Q DNA Digital PCR Mix (Thermo

Fisher Scientific, cat# A52490) with the forward primer (5'-CAACACGGTCCGCTACGACTA-3'), reverse primer (5'-TGCCATAACTCGCCATCAGA-3'), and tp47 probe (5'- FAM-CGGTGATGACGCGAGCTACACCA-ZEN/IABK-3') a master mix was made according to the manufacturers recommendations. 1 µl of sample DNA was added to 9 µl of master mix and 9 µl of that was transferred to the Absolute Q MAP16 plate (Thermo Fisher Scientific, cat # A52688). The assay was run with the cycling conditions recommended by the manufacturer, including an annealing temperature of 58°C. The dPCR data file was analyzed using QuantStudio Absolute Q digital PCR Software, version 6.3.0.

The dPCR-quantified Tp47 standards were subsequently serially diluted and run on the tp47 screening PCR to generate standard curves (Ct versus $\log_{10}$ copy number), which enabled direct calculation of genome copy numbers in experimental samples. The resulting curves had a range from 4 x 10^6 to 4 genome copies per reaction, with the regression equation Ct = -3.361 $\log_{10}$(copies) + 35.678, R^2 = 0.999, and an average 10-fold dilution shift of 3.36 cycles, corresponding to a PCR efficiency of 98.4%.

2.2. DNA Extractions

Due to the metagenomic nature of mucocutaneous lesion swabs, multiple DNA extraction methods were evaluated to optimize recovery of TPA DNA for downstream sequencing applications. These included both selective DNA extraction methods aimed at enriching bacterial DNA while depleting host DNA, and total nucleic acid extractions to capture all DNA present.

2.2.1. Selective DNA Extractions

The first method to increase the yield of TPA DNA involved using magnetic beads to bind eukaryotic DNA, pellet the beads and transfer the supernatant to a new tube. This effectively enriches for prokaryotic DNA. The second selective extraction involved using differential lysis on the principle that prokaryotic cells contain cell walls that are more resistant to detergents. Using a mild detergent, eukaryotic cells are lysed and the DNA can be degraded with a DNase. Once the DNase is removed the prokaryotic cells are lysed, providing a DNA eluate enriched for bacteria.

2.2.1.1. Enriching Bacterial DNA

Total DNA, including host, microbiome and TPA, was extracted using the Qiagen DNeasy Blood & Tissue Kit (Qiagen, Cat 69504). Briefly, 20 µl of proteinase K (Qiagen) and 200 µl of Buffer AL (Qiagen) were added to 200 µl of UTM containing the sample. The mixture was vortexed and incubated at 56°C for 10 minutes. After incubation, 200 µl of ethanol was added, and the sample was mixed before being transferred to a Qiagen spin column (Qiagen).

The column was centrifuged at 6,000 x g for 1 minute, and the flow-through was discarded. With a new collection tube, 500 µl of Buffer AW1 (Qiagen) was added, and the column was centrifuged at 6,000 x g for 1 minute. Again, the flow-through was discarded, and the column was transferred to a new collection tube. Next, 500 µl of Buffer AW2 (Qiagen) was added, and centrifuged at 20,000 x g for 3 minutes. The column was then placed in

a fresh collection tube and centrifuged at 20,000 x g for an additional minute to ensure removal of Buffer AW2 (Qiagen).

To elute the nucleic acid out of the spin column, it was placed in a clean elution tube and 200 µl of Buffer AE (Qiagen) was added. The column was incubated at room temperature for 1 minute before a final centrifugation at 6,000 x g for 1 minute for total nucleic acid collection.

The NEBNext Microbiome DNA Enrichment Kit (New England Biolabs, cat# E2612S) was then used to remove the host DNA from the sample. Using the manufacturer's recommendations, 160 µl of MBD2-Fc-bound magnetic beads (New England Biolabs) were prepared in 5X Bind/Wash buffer (New England Biolabs) for 1 µg of input DNA. This mixture was gently pipetted and rotated for 15 minutes at room temperature to allow methylated host DNA to bind to the beads. The beads were bound on a magnetic separation rack (New England Biolabs, cat# S1506S) and the supernatant containing enriched microbial DNA was transferred to a clean tube.

Microbial DNA was further purified from the Bind/wash buffer using 1.8X AMPure XP beads (Beckman Coulter, cat# A63880). The mixture was thoroughly mixed by pipetting up and down and incubated at room temperature for 10 minutes to permit DNA binding. Beads were then captured on a magnetic stand for 5 minutes, and the supernatant was carefully removed. Two washes with 200 µl of 80% ethanol were performed, each with a 30-second incubation, followed by removal of the ethanol. After the final wash, the beads were air-dried for 30 seconds. DNA was eluted by adding 50 µl of nuclease-free water

(Integrated DNA Technologies, cat# 11-05-01-14), gently mixing, and incubating for 5 minutes at room temperature. The tube was returned to the magnetic stand, and the cleared eluate was transferred to a clean tube.

2.2.1.2. Depleting Host DNA

Host DNA depletion extractions were performed on swab samples in UTM using the QIAamp DNA Microbiome Kit (Qiagen, cat# 51704), following a protocol tailored for a 1 ml sample volume. Briefly, 500 µl of Buffer AHL (Qiagen) was added to 1 ml of sample in a 2 ml microcentrifuge tube, mixed, and incubated for 30 minutes at room temperature with end-over-end rotation. After incubation, samples were centrifuged at 10,000 x g for 10 minutes, and the supernatant was carefully discarded to avoid disturbing the pellet.

The bacterial pellet was treated with 190 µl of Buffer RDD (Qiagen) and 2.5 µl of Benzonase (Qiagen), then incubated at 37°C for 30 minutes at 600 rpm. Following Benzonase treatment, 20 µl of Proteinase K (Qiagen) was added, and the mixture was incubated at 56°C for another 30 minutes. Next, 200 µl of Buffer ATL with Reagent DX (Qiagen) was added, and the sample was transferred to a Pathogen Lysis Tube L (Qiagen). Mechanical lysis was performed using a TissueLyser II (Qiagen, cat# 85300) for 10 minutes at a frequency of 50/s.

After lysis, the sample was centrifuged at 10,000 x g for 1 minute to reduce foam, and the supernatant was transferred to a fresh tube. An additional 40 µl of Proteinase K (Qiagen) was added, followed by incubation at 56°C for 30 minutes. Next, 200 µl of Buffer APL2 (Qiagen) was mixed into the lysate, incubated at 70°C for 10 minutes, and 200 µl of

ethanol was added. The sample was then loaded onto a QIAamp UCP Mini Column (Qiagen) and centrifuged at 6,000 x g for 1 minute, with the process repeated as needed.

The column was washed with 500 µl of Buffer AW1 (Qiagen) and centrifuged at 6,000 x g, followed by 500 µl of Buffer AW2 (Qiagen) and centrifugation at 20,000 x g for 3 minutes. A final centrifugation step ensured the complete removal of residual wash buffer. DNA was eluted in 50 µl of Buffer AVE (Qiagen) by incubating for 5 minutes at room temperature, followed by centrifugation at 6,000 x g for 1 minute. Eluted bacterial DNA was stored at 4°C for downstream applications.

2.2.2. Total Nucleic Acid Extractions

Total nucleic acid extraction methods are designed to recover all nucleic acids present in a sample, including both host and microbial material. In this study, only DNA was assessed and used for downstream applications. Unlike targeted enrichment or host depletion strategies, these non-selective methods maximize the likelihood of recovering low-abundance organisms such as TPA.

2.2.2.1. Qiagen DNeasy Blood & Tissue Kit

For total nucleic acid extraction, the Qiagen DNeasy Blood & Tissue Kit (Qiagen) was used following the same protocol described in the earlier microbiome enrichment experiments, with one modification: the final elution volume was reduced from 200 µL to 100 µL to increase nucleic acid concentration. Briefly, 20 µL of proteinase K (Qiagen) and 200 µL of Buffer AL (Qiagen) were added to 200 µL of UTM containing the sample. After vortexing and incubation at 56°C for 10 minutes, 200 µL of ethanol was added. The lysate

was then transferred to a spin column and washed sequentially with Buffer AW1 (Qiagen) and Buffer AW2 (Qiagen). Following a final high-speed centrifugation to remove residual wash buffer, DNA was eluted with 100 µL of Buffer AE (Qiagen). No microbiome enrichment or host DNA depletion steps were applied.

2.2.2.2. KingFisher Flex Extraction

Total nucleic acid extractions were performed on the KingFisher Flex instrument (Thermo Fisher Scientific, cat# 5400610) with the Applied Biosystems 5x MagMAX-96 Viral Isolation Kit (Thermo Fisher Scientific, cat#1836-5). Using a modified protocol and the deep-well plate format, 200 µL of swab samples in UTM were added to 620 µL of lysis binding solution/bead mix (Thermo Fisher Scientific). Sample lysis was followed with two rounds of 300 µL of wash 1 (Thermo Fisher Scientific), two rounds of 300 µL of wash 2 (Thermo Fisher Scientific), and a 110 µL elution step (Thermo Fisher Scientific). The extracted nucleic acid was stored at 4°C.

2.2.2.3. BioMérieux eMAG System

Swab samples in UTM were extracted using the BioMérieux eMAG automated system (BioMérieux, cat# 418591) with the Generic_3.0.4 protocol. Initially, 200 µl of sample was manually added to Lysis Buffer (BioMérieux) in individual extraction vessels (BioMérieux, cat# 280135). The vessels were then placed back on the instrument, which performed the remaining steps automatically: addition of 50 µl of silica (BioMérieux, cat# 280133), sequential washing steps, and elution of nucleic acids in 110 µl of Extraction Buffer 3 (BioMérieux) into clean eluate tubes.

2.2.2.4. DNA Fragment Size Assessment

DNA fragment size distributions were assessed using the Agilent 2200 TapeStation (Agilent, cat# G2964A) with Genomic DNA ScreenTape (Agilent, cat# 5067-5365). For each sample, 1 µL of extracted DNA was mixed with 10 µL of Genomic DNA Sample Buffer (Agilent, cat# 5067-5366) and loaded onto the ScreenTape. A DNA ladder provided with the kit was prepared according to the manufacturer's instructions and served as a reference for fragment sizing. Prepared samples were run on the TapeStation instrument, and electropherogram profiles were generated and analyzed using 2200 TapeStation Software version A.01.05(SR1).

2.2.2.5. DNA Quantification Using Qubit Flex Fluorometer

Genomic DNA was quantified with the Qubit Flex Fluorometer (Invitrogen, cat# Q33327) using the Qubit dsDNA High Sensitivity (HS) assay kit (Invitrogen, cat# Q32851). To calibrate the instrument, two standard tubes were prepared by adding 190 µl of buffer (Invitrogen) and 10 µl of each standard (Invitrogen). For each sample, a 0.5 ml Qubit assay tube was prepared by adding 198 µl of Qubit dsDNA HS buffer (Invitrogen) and 2 µl of DNA sample. All tubes were briefly vortexed and incubated at room temperature for two minutes, followed by a quick centrifugation to settle contents and remove air bubbles. The Qubit Flex Fluorometer was calibrated with the two standard solutions, after which the DNA concentration of each sample was measured and recorded.

2.3. Whole Genome Amplification

None of the extraction methods tested yielded sufficient TPA DNA for Nanopore sequencing. While RNA bait capture is commonly used to enrich microbial DNA, current protocols are optimized for short-read sequencing and fragment sizes (~120 bp). As an alternative, whole genome amplification (WGA) methods were explored to increase total DNA while preserving long fragment lengths. These methods utilize isothermal multiple displacement amplification (MDA) of entire genomes within a sample. Both random-primed and sequence-specific approaches were evaluated to improve TPA DNA concentration in the sequencing libraries.

2.3.1. Random-Primed Whole Genome Amplification

WGA was performed on extracted DNA samples using the REPLI-g Advanced DNA Single Cell Kit (Qiagen, cat# 150363), which utilizes random hexamer primers and phi29 DNA polymerase. Briefly, 2.5 µl of template DNA (1 ng) was combined with 2.5 µl of Buffer D1 (Qiagen) in a 0.2 ml PCR tube (Diamed, cat# DIATEC420-1377), vortexed, and incubated at room temperature for 3 minutes to denature the DNA. Following denaturation, 5 µl of Buffer N1 (Qiagen) was added, mixed by vortexing, and the sample was placed on ice.

A master mix was prepared with 9 µl of nuclease-free water, 29 µl of REPLI-g Advanced sc Reaction Buffer (Qiagen), and 2 µl of REPLI-g sc DNA Polymerase (Qiagen) per reaction. For each sample, 40 µl of master mix was added to the 10 µl of denatured DNA solution, bringing the final reaction volume to 50 µl. The reactions were incubated at 30°C for 2 hours, followed by inactivation of the polymerase at 65°C for 3 minutes. Amplified

DNA was purified using a 2X AMPure XP bead cleanup as described in Section 2.2.1.1, eluting into 30 µl of nuclease-free water.

2.3.2. Selective Whole Genome Amplification

Selective whole genome amplification (SWGA) was performed to preferentially amplify TPA DNA while minimizing host background. This method uses sequence-specific primers designed with a high affinity to TPA genomic regions and low affinity to host DNA, enhancing target specificity during amplification.

SWGA was carried out using the EquiPhi29 DNA Polymerase (Thermo Fisher Scientific, cat# A39390) with the SWGA Pal 12 primer set developed by Thurlow et al., 2022 (**Table 2.1**)**.** The SWGA primers included phosphorothioate bonds between the final two nucleotides at the 3' end to resist the 3'->5' exonuclease activity of phi29 polymerase.

Table 2.1. SWGA primers developed by Thurlow et. al. 2022.

Primer Set ID	Primer Name	Sequence (5' to 3')	References
SWGA-Pal 12	SWGA-Pal 4.1	CGCGA*A*A	Thurlow et al., 2022
	SWGA-Pal 4.2	CGTAC*C*G	Thurlow et al., 2022
	SWGA-Pal 4.3	CGTAC*G*A	Thurlow et al., 2022
	SWGA-Pal 4.4	CGTAT*C*G	Thurlow et al., 2022
	SWGA-Pal 4.5	TACGC*G*T	Thurlow et al., 2022
	SWGA-Pal 5.1	CGCGT*A*A	Thurlow et al., 2022
	SWGA-Pal 2	CGCGC*A*A	Thurlow et al., 2022

*Nucleotides with phosphorothioate bonds

For each reaction, 2.5 µl of extracted DNA was combined with 1 µl of an 80 µM SWGA Pal 12 primer pool, 1 µl of nuclease-free water, and 0.5 µl of 10x EquiPhi Reaction Buffer (Thermo Fisher Scientific). The mixture was denatured at 95°C for 5 minutes using a VeritiPro 96-well thermal cycler (Thermo Fisher Scientific, cat# A48141), then immediately placed on ice.

The 5 µl of denatured DNA was combined with a master mix consisting of 1.5 µl of 10x Reaction Buffer (Thermo Fisher Scientific), 0.2 µl of 100 mM DTT (Thermo Fisher Scientific), 2 µl of 10 mM dNTPs (Thermo Fisher Scientific), and 1 µl of EquiPhi29 DNA Polymerase (Thermo Fisher Scientific). The final reaction volume was adjusted to 20 µl with nuclease-free water. Amplification was performed at 45°C for 3 hours, followed by heat inactivation at 65°C for 10 minutes. Amplified DNA was cleaned using a 2X AMPure XP bead cleanup as described in Section 2.2.1.1, with elution into 50 µl of nuclease-free water.

2.3.3. Custom Quantification Curve for High-Abundance Samples

Digital-PCR (dPCR) setup, cycling conditions, and data analysis were identical to those described in Section 2.1.3. Briefly, the previously validated Tp47 primer/probe assay was run on the Absolute Q platform using the Absolute Q DNA Digital PCR Mix.

Because several WGA samples yielded Ct values far above the range of the synthetic ATCC standards, a high-yield SWGA product (sample 0314) was selected to construct a broader standard curve. A 1:10 serial dilution series (neat to 10^{-9}) was prepared in nuclease-free water, and each dilution was quantified once by dPCR.

Quantifiable dilutions from 10^{-3} to 10^{-9} produced a consistent 10-fold decrease in copy number per dilution step, confirming dilution accuracy. Based on this verified log-linear trend, concentrations of earlier dilutions that saturated the Absolute Q array (neat, 10^{-1}, 10^{-2}) were extrapolated; the undiluted sample was estimated at 4.0×10^{7} copies per µL.

The resulting calibration curve provided quantification and was used to calculate fold changes and copy numbers for all WGA products in both the random-primed (Section 2.3.1) and sequence-specific (Section 2.3.2) workflows.

2.3.4. De-Branched DNA Structures Using T7 Endonuclease I

Following whole genome amplification, branched DNA structures were removed using T7 Endonuclease I (New England Biolabs, cat# M0302) following the ONT Ligation sequencing gDNA - whole genome amplification (SQK-LSK112) protocol. A total of 3 µl of NEBuffer 2 (New England Biolabs) and 1.5 µl of T7 Endonuclease I (New England Biolabs) were added to 1.5 µg of amplified DNA, and the volume was adjusted to 30 µl with nuclease-free water. The reaction was incubated at 37°C for 15 minutes on a VeritiPro thermal cycler (Thermo Fisher Scientific).

Immediately following incubation, 20 µl of TE buffer (Integrated DNA Technologies, cat# 11-05-01-13) and 50 ul of Ampure beads (1x) were added. DNA purification was performed as described in Section 2.2.1.1, and the final product was eluted into 27 µl of nuclease-free water.

2.4. Library Preparation and Flow Cell Loading

2.4.1. Library preparation

Library preparation was adapted from Oxford Nanopore Technologies' ligation-based workflows, including the Ligation Sequencing gDNA - Whole Genome Amplification (Version: WAL_9154_v112_revC_09Feb2022) and the Ligation sequencing gDNA - native barcoding workflow (Version: NBE_9121_v109_revG_19Jan2021). Modifications were implemented to maximize the amount of DNA carried forward at each step and to optimize sequencing performance for multiplexed clinical samples with variable genome-wide amplification efficiency. After the barcoding stage, samples were split and processed in duplicate to ensure enough DNA for 2 libraries, thus enabling a re-load step halfway through sequencing.

Genomic DNA was normalized to 1000 ng in 53.5 µl of nuclease-free water prior to end-repair and A-tailing. NEBNext Ultra II End-prep Reaction Buffer and Ultra II End-prep Enzyme Mix (New England Biolabs, cat# E7647AA and E7646AA) were thawed to room temperature and mixed. For each reaction, 3.5 µl of reaction buffer and 3 µl of enzyme mix were added to the DNA in a 0.2 ml PCR tube (DIATEC, cat# 420-1377). Samples were mixed by pipetting and incubated on a VeritiPro thermal cycler using the following program: 20 °C for 10 minutes, 65 °C for 10 minutes, followed by a hold at 4 °C. End-prepped DNA was purified with a 1x AMPure XP bead cleanup and eluted in 27 µl of nuclease-free water. DNA concentrations were measured using a Qubit dsDNA High Sensitivity assay.

Barcodes were added using the Oxford Nanopore Technologies Native Barcoding Expansion kit (Oxford Nanopore Technologies, cat# EXP-NBD196). End-prepped DNA was normalized to 500 ng in 22.5 µl of nuclease-free water. Each barcoding reaction consisted of 25 µl of NEB Blunt/TA Ligase Master Mix (New England Biolabs, cat# M0367L) and 2.5 µl of the assigned Native barcode (Oxford Nanopore Technologies). Reactions were incubated on a VeritiPro thermal cycler at 20 °C for 20 minutes, followed by 65 °C for 10 minutes, and held at 4 °C. To prevent barcode carryover, a 1x AMPure XP bead purification was performed, and the barcoded DNA was eluted in 27 µl of nuclease-free water. DNA concentration was again assessed using a Qubit fluorometer.

Barcoded samples were pooled by combining equal volumes into a 1.5 ml DNA Low retention tube (Labcon, cat# 3039-560-000-9). A 1x volume of AMPure XP beads was added to concentrate the pooled DNA to 60 µl. The sample was then split into two 30 µl aliquots to support a two-step sequencing strategy with flow cell reloading. Each 50 µl adapter ligation reaction contained 30 µl of pooled DNA, 5 µl of either ONT Native Adapter Mix (for R10.4.1 flow cells) or AMII Adapter Mix (for R9.4.1 flow cells), 10 µl of NEBNext Quick Ligation Reaction Buffer (New England Biolabs, cat# E6058AA), and 5 µl of Quick T4 DNA Ligase (New England Biolabs, cat# E6057AA). Reactions were incubated at room temperature for 20 minutes. Following adapter ligation, libraries were purified with a 1x AMPure XP bead cleanup using two 250 µl washes of Short Fragment Buffer (SFB) (Oxford Nanopore Technologies) in place of ethanol. Final libraries were eluted in 15 µl of Elution Buffer (Oxford Nanopore Technologies) and quantified using a Qubit 1X dsDNA

High Sensitivity assay. For the first sequencing run, only a single library was prepared, and no reload was performed.

2.4.2. Flow cell Loading

Sequencing was performed on Oxford Nanopore MinION flow cells using either R9.4.1 pore chemistry with AMII adapters (early experiments) or R10.4.1 chemistry with ONT Native Adapters (later experiments). 13 µl of the final library was loaded per run.

2.4.3. Flow Cell Washing and Reload Strategy

To extend sequencing time and improve yield, sequencing runs incorporated a flow cell wash and reload procedure using the ONT Flow Cell Wash Kit (Oxford Nanopore Technologies, cat# EXP-WSH004). After ~20 hours of sequencing, when active pore usage dropped below 10%, the run was paused in MinKNOW and a wash was performed according to the manufacturer's protocol. 400 µL of freshly prepared wash mix (2 µL Wash Mix with DNase I and 398 µL Wash Diluent, Oxford Nanopore Technologies) was loaded into the priming port of the flow cell, and the cell was incubated at room temperature for 60 minutes. The wash removed residual DNA and buffer components, helping to clear idle or "recovering" pores and restore sequencing activity. After washing, the flow cell was re-primed and reloaded with a second aliquot of the barcoded library, and sequencing was resumed. This strategy allowed for re-engagement of pores and an extended sequencing window, typically achieving an additional 20–30 hours of productive runtime.

2.5. Nanopore Sequencing

2.5.1. Standard Sequencing (SS)

Sequencing was performed on the GridION platform (Oxford Nanopore Technologies, cat# GRD-MK1) using either R9.4.1 or R10.4 flow cells, depending on availability. Libraries were basecalled in high-accuracy mode using MinKNOW version 23.11.7, with default parameters, including a minimum Q score of 9 and no filtering on read length. Active pore counts were checked prior to loading.

Each DNA library was split into two equal portions after barcoding (section 2.4.1), with 13 µL loaded per run. Sequencing was conducted for approximately 24 hours before being paused for a flow cell wash and reload step, as described in Section 2.4.3. The second portion of the library was then reloaded, and sequencing resumed for an additional 20–24 hours, or until pore activity declined below 20%, at which point the run was terminated.

2.5.2. Adaptive Sampling (AS)

Adaptive sampling was conducted using the same instrumentation, run parameters, and library preparation strategy as standard sequencing. The only difference was that adaptive sampling mode was enabled in MinKNOW version 23.11.7, allowing real-time enrichment for TPA DNA. The SS14 reference genome (NC_021508.1) served as the enrichment target, guiding read retention during sequencing.

Reads aligning to the reference were sequenced to completion, while off-target reads were ejected early to free sequencing pores. All other aspects of the workflow including flow cell type, run duration, Q score thresholds, and the mid-run wash/reload protocol

(Section 2.4.3), were identical to standard sequencing runs to enable direct performance comparisons.

2.6. Bioinformatics Analysis

2.6.1. Basecalling and Read QC

Following sequencing, raw data files in either FAST5 (for R9 flow cells) or POD5 (for R10 flow cells) format were transferred to a GPU-enabled workstation for high-accuracy basecalling. Dorado version 0.7.1 (Oxford Nanopore Technologies, Oxford, UK) was used to perform basecalling using ONT's super-accurate (SUP) models tailored to each chemistry: dna_r9.4.1_e8_sup@v3.6 for R9 datasets and dna_r10.4.1_e8.2_400bps_sup@v5.0.0 for R10 datasets. The resulting BAM-formatted outputs were then converted to FASTQ format using Samtools version 1.20 for downstream analysis (141).

Initial read quality metrics were assessed using NanoPlot version 1.43.0 (142). For each sample, NanoPlot generated summary statistics, read length distributions, and quality score profiles, all compiled into structured HTML reports and graphical outputs. These metrics were used to evaluate the overall quality of sequencing and guide trimming thresholds.

Taxonomic classification of the basecalled reads was performed using Kraken2 version 2.1.3 against a custom database specific to TPA (143). Classification output included full reports, summary files, and filtered FASTQ files containing only reads assigned to TPA.

These classified reads were used for downstream genome assembly and polishing workflows.

Read trimming was carried out in two stages to improve data quality for assembly. First, Filtlong version 0.2.1 was used to retain reads ≥1,000 bp in length (144). These filtered reads were then processed with Chopper version 0.9.0, which removed 100 bp from both the 5′ and 3′ ends of each read and discarded any remaining reads with average Phred quality scores below 10 (145). This trimming step was applied uniformly across all samples.

2.6.2. Genome Assembly

2.6.2.1. De novo

De novo assembly of classified TPA reads was performed using three long-read assemblers: Flye version 2.9.5, Raven version 1.8.3, and Unicycler version 0.4.4 (146–148). Assemblies were attempted on FASTQ files from twelve high-coverage samples to evaluate the performance and output of each assembler. Due to computational limitations with large datasets, Raven and Unicycler could not be run on full datasets; therefore, a non-random down-sampling strategy was applied to each prior to assembly.

Reads were first aligned to the TPA SS14 reference genome (NC_021508.1) using Minimap2 version 2.24, and the resulting SAM files were sorted and indexed with Samtools version 1.20 (149). Down-sampling was performed using the subsample_bam utility from Pomoxis version 0.3.15 (Oxford Nanopore Technologies, Oxford, UK), which extracted reads from aligned BAM files to achieve approximate genome coverages of 50x

and 100x. These BAM files were then converted back to FASTQ format using Samtools for input into each assembler.

Flye assemblies were generated for each down-sampled FASTQ file using the --nano-hq, --meta, and --genome-size parameters. Raven assemblies were generated using the default parameters. Unicycler was run in long-read only mode with the --mode bold parameter to improve performance on sparse and metagenomic datasets.

QUAST version 5.3.0 was used to generate detailed assembly statistics (150). Each assembly was compared to the SS14 reference genome (NC_021508.1), and metrics such as N50, genome fraction, number of contigs ≥1 kb, and rates of mismatches and indels per 100 kb were compiled.

2.6.2.2. Reference-guided

Reference-guided assembly was performed using the TPA SS14 reference genome (NC_021508.1). All classified reads were aligned to the reference using Minimap2 version 2.24 with parameters optimized for ONT data (-ax map-ont). The resulting SAM files were sorted and indexed using Samtools. Genome coverage, read depth and # of mapped reads was determined with the Samtools version 1.20 Coverage module.

Consensus sequences were generated using the Samtools Consensus module, with separate parameters for R9 and R10 data. With the R9 datasets, the following thresholds were applied: a minimum mapping quality of 10 (--min-MQ), minimum base quality of 7 (--min-BQ), and consensus quality cutoff of 10 (-C 10). For R10 datasets, the configuration

profile r10.4_sup was used instead, as recommended by Samtools for this sequencing chemistry.

The resulting assemblies were then subjected to polishing and quality assessment as described below.

2.6.3. Assembly Polishing

Genome polishing was performed to improve the base-level accuracy of reference-guided assemblies by correcting ambiguous bases and insertion/deletion (indel) errors inherent to ONT sequencing. Several polishing strategies were evaluated, and the final approach was selected based on comparative performance.

2.6.3.1. Racon

The effectiveness of Racon version 1.5.0 was investigated to determine its proficiency at improving consensus accuracy (151). For each sample, long reads were aligned back to their respective draft assemblies, generated in section 2.6.3.2, using minimap2 with the ONT flag -x map-ont to produce PAF-format alignments. These alignments were then processed using Racon with default parameters. Polished assemblies were output in FASTA format and used for downstream assessment.

2.6.3.2. Medaka

Medaka polishing was evaluated as an alternative to Racon. Using the unpolished assemblies from 2.6.3.2, Medaka version 2.0.1 (Oxford Nanopore Technologies, Oxford, UK) was applied using the medaka_consensus module. Chemistry-specific basecalling

models were selected based on sequencing platform: r941_prom_sup_g507 was used for R9 datasets, and r1041_e82_400bps_sup_v5.0.0 was used for R10 datasets. Two rounds of Medaka polishing were applied to each assembly. A third round was evaluated but was not included in the final protocol due to diminishing returns and occasional reintroduction of indel errors.

A final polishing step was performed using Homopolish to correct homopolymer-associated errors remaining after Medaka (152). Homopolish version 0.4 was used along with a bacterial sketch (bacteria.msh) and ONT flow cell–specific model files: R9.4.pkl for R9 and R10.3.pkl for R10.

The final polishing workflow, two rounds of Medaka followed by one round of Homopolish, was applied uniformly to all assemblies used in downstream quality assessment and phylogenetic analysis.

2.6.4. Genome Assembly Assessment

Assembly completeness was assessed using BUSCO version 5.8.0 with the *spirochaetales_odb10* lineage dataset (version: 2024-01-08) (153). Each assembly was evaluated for the presence of single-copy orthologs expected in TPA, and the number of complete versus missing genes was used to estimate assembly quality. Assemblies containing more than 317 complete BUSCO genes were considered high quality. To visually summarize these results, a plot was generated using the generate_plot.py script provided with BUSCO utilities.

QUAST results, described previously in section 2.6.2.1, were used to compare each assembly against the TPA SS14 reference genome (NC_021508.1), with key metrics including N50, genome fraction, and indel/mismatch rates.

2.7. Analysis of TPA in Manitoba

2.7.1. Multiple Sequence Alignment

Polished consensus sequences from this study were combined with 89 publicly available Treponema pallidum genomes downloaded from RefSeq for comparative analysis (Appendix 1.2). Prior to multiple sequence alignment (MSA), all genomes were annotated using Prokka v1.14.6 (154). Functional annotation was guided by a custom protein FASTA file derived from the Nichols reference genome (CP_004010.1), provided via the --proteins parameter. Annotations were standardized by specifying the genus (Treponema), species (pallidum), and a locus tag prefix of “TP.”

During initial MSAs, poor alignment of the duplicated rRNA operons was observed. To resolve this, using a custom python script each genome FASTA file was split at the 5′ end of the 5S rRNA gene within the first rRNA operon, based on coordinates identified from Prokka annotations. This process was automated in Python (see **Supplemental Information 1: split-fasta.py**). The two resulting genome fragments were aligned independently with MAFFT v7.525 using the --auto parameter. The alignments were then rejoined by matching sequence IDs, restoring full-length assemblies (see **Supplemental Information 2: merge-alignments.py**) (155).

Manual curation of minor gaps and assessment of alignment quality were visualized in Geneious Prime v2024.0.7 (https://www.geneious.com). In situations with specimens being sequenced in multiple runs the assembly with the highest quality metrics was used for the final trees.

2.7.2. Phylogenetic Analysis

Phylogenetic reconstruction of TPA genomes was performed using IQ-TREE version 2.3.6 based on the MSA generated with MAFFT (156). The full, unmasked alignment of genome assemblies was used to retain complete genomic variation across samples, including repetitive and divergent regions. IQ-TREE was run with the integrated ModelFinder module (157) to select the best-fitting substitution model, and 10,000 ultrafast bootstrap replicates were performed using the -B 10000 flag to assess branch support (158). The -T AUTO option was used to optimize CPU core usage. The resulting maximum likelihood tree was used as the basis for downstream comparative analysis. All tree visualization and annotation steps were conducted in R version 4.3.2 using RStudio version 2023.09.1 (see **Supplementary Information 3**).

To identify macrolide resistance associated mutations, ONT reads were mapped to the TPA Nichols reference rRNA operon (GenBank CP004010.2, positions 233169–236118) using minimap2 with ONT-specific alignment parameters (-aLx map-ont --cs --MD). Sorted and indexed BAM files were processed with Clair3 using ONT models (R10: r1041_e82_400bps_sup_v500; R9: r941_prom_sup_g5014) depending on flow cell chemistry. Variant calling was performed with the --haploid_precise setting to improve

single-haplotype accuracy, and the --print_ref_calls flag was included to retain reference positions. Variants were specifically reviewed at positions A2058 and A2059 (E. coli numbering; position 3898 in TPA rRNA operon), which are associated with macrolide resistance.

Sample-specific rRNA operon sequences were reconstructed using BCFtools consensus, applying the Nichols reference operon as input (-f) and generating FASTA output with the -o flag. These reconstructed sequences were subsequently used for two purposes: (1) confirmation of macrolide resistance genotyping, and (2) BLAST-based quality control to identify and exclude assemblies with misassembled rRNA regions.

To ensure high-confidence phylogenetic analysis, genome assemblies were subjected to a multi-criteria filtering process. Assemblies with fewer than 316 complete BUSCOs were excluded, based on the expected maximum completeness for TPA reference genomes. Among the genomes passing the BUSCO filter, further quality control was performed by verifying the integrity of the 23s rRNA gene. BLAST+ version 2.16.0 was used to compare each assembly's 23s rRNA sequence to a custom reference database containing 23s rRNA from the TPA reference genomes in Appendix A (159). Assemblies with less than 100% identity to any known 23s sequence were excluded to minimize the risk of mis-assemblies or contamination. The remaining assemblies were manually inspected within the MSA, and additional samples were removed based on clustered SNPs within the rRNA operons. The final tree visualization and annotation steps were conducted in R version 4.3.2 using RStudio version 2023.09.1 (see **Supplementary Information 4**).

2.7.3. *In-silico* analysis of Genes of interest

To evaluate regions of interest such as PCR primer sites and genes with known or suspected variability, an in-silico analysis was performed using assembled genomes from samples with sufficient sequencing depth. For each gene of interest, coding sequence (CDS) coordinates were extracted from the annotated genome assemblies generated using Prokka (Section 2.7.1), and custom Python scripts were used to retrieve sample-specific gene sequences, see **Supplementary Information 5**.

To investigate potential SNPs in the tp47 screening PCR target, samples were filtered based on read depth across the gene using samtools depth version 1.20. Genes with a minimum depth of >50x were extracted and compiled into a multi-FASTA file. The MSA was performed using MAFFT version 7.525 with default parameters. The alignment was visualized in Geneious Prime to inspect nucleotide conservation and assess potential mismatches in the primer and probe binding regions used in the screening qPCR assay (70,160). Primer and probe sequences were annotated on the alignment to facilitate identification of SNPs that could interfere with assay performance.

To assess repeat copy number variation in the *arp* gene, which contains tandem 60-bp repeats relevant to molecular typing, a similar approach was used. CDS coordinates were extracted from Prokka annotations, and the corresponding sequences were retrieved per sample. Mean per-base coverage was calculated using samtools depth, and only *arp* sequences with ≥5x coverage across the full gene were included. Tandem Repeat Finder (TRF) version 4.10 was used to identify and count 60-bp tandem repeat units within each

arp gene using default parameters (161). TRF was run using the recommended quick start parameters (i.e.. 2 5 7 80 10 50 2000).

Chapter 3: Results

The inability to culture and obtain isolates of TPA makes any molecular work more difficult as you are inherently working with a metagenomic sample where the organism of interest is in relative low abundance. Each step of this analysis required careful considerations and optimizations.

Optimization began with the initial DNA extraction step, which required consideration of the fragility and low abundance of TPA (Results 3.1 - 3.2). During library preparation, amplification conditions were adjusted to maximize recovery of TPA DNA to a suitable concentration for sequencing, while also determining the number of samples that could be multiplexed into a single flow cell (Results 3.4). Development of the analysis pipeline involved evaluating numerous bioinformatics programs and establishing an appropriate sequence for their execution (Results 3.6). Finally, high-quality assemblies were carried forward for downstream comparative genomics. This included multiple sequence alignment of assembled genomes, maximum likelihood phylogenetic reconstruction, and in silico analysis of relevant loci such as *arp* and *tp47* (Results 3.7). To ensure reliable phylogenetic inference only 13 assemblies, achieving the highest quality metrics during this project, were retained for the final analysis.

3.1. Clinical Specimen Collection for TPA Analysis

3.1.1. Screening PCR for Identifying TPA-Positive Samples

The Tp47 screening PCR was developed using previously published primer and probe sequences and implemented as a TaqMan-based qPCR assay (70,160). Validation was

performed using both synthetic Treponema pallidum DNA (ATCC) and previously confirmed TPA-positive clinical specimens as positive controls.

To optimize sensitivity, a temperature gradient was applied to refine the annealing temperature, and the primer and probe concentrations were adjusted. Serial dilutions of the synthetic TPA DNA (ATCC) ranging from 10^{-1} to 10^{-8} were prepared to evaluate the assay's amplification efficiency, determine the limit of detection (LoD), and enable accurate quantification of TPA DNA in clinical samples (**Figure 3.1a**). Each dilution was tested in triplicate, yielding an amplification efficiency of 98.6% and an R^2 of 0.995, indicative of a well performing assay (**Figure 3.1b**). Amplification was consistently observed in all three replicates for the first six dilutions (10^{-1} to 10^{-6}), while no amplification was detected in the last two dilutions (10^{-7} and 10^{-8}), establishing an LoD of 0.47 copies/µL.

Following optimization, a total of 5,107 extracted clinical nucleic acid specimens were screened for TPA over a two-year period. The overall positivity rate for TPA was 6.5% (n=334), with genital swabs comprising the majority (76%) of positive specimens. This screening approach also facilitated the retrospective retrieval of clinical swabs stored in UTM, providing sufficient TPA positive material for DNA extraction optimization.

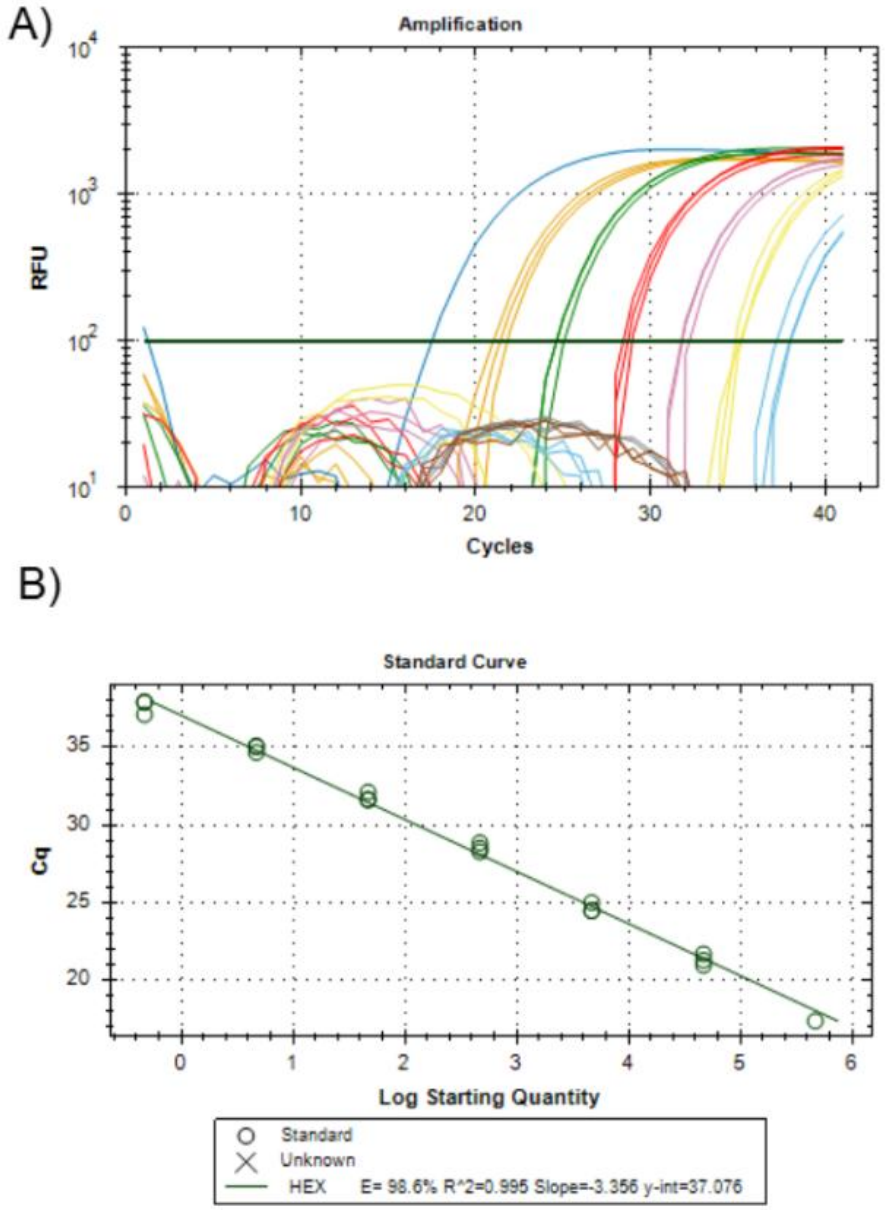


Figure 3.1. Screening PCR for TPA.

A) visualization of the relative fluorescent units (RFU) after each PCR cycle for the control dilution series in log view. B) Cycle threshold (Ct) for each dilution plotted against the TPA DNA copies per µL.

3.1.2. The Multiplex Lesion Panel

The resurgence of syphilis in Manitoba, with incidence rates exceeding the national average, suggested potential gaps in diagnostic testing and the need for a more robust approach to detecting TPA. The initial Tp47 screening PCR revealed a positivity rate of 6.5% (**Table 3.2**) and further investigation found that 36% of TPA-positive samples lacked a corresponding physician-ordered TPA PCR request**.** Even more concerning was 19% lacked syphilis serology within 30 days, demonstrating that case detection was often dependent on clinical suspicion rather than systematic screening (68).

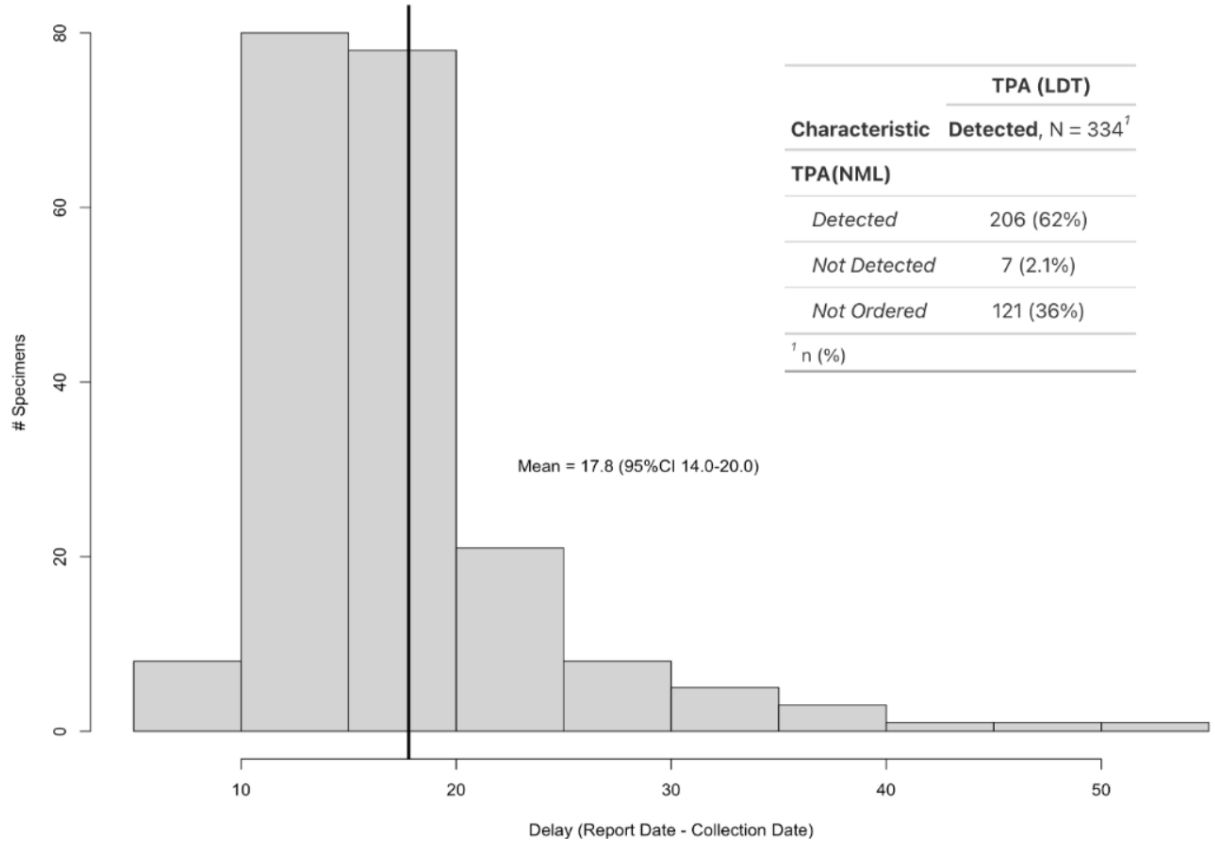
Table 3.1. The demographics of specimens screened (sex, age, regional health authority, specimen type) by the TPA screening assay.

Adapted with permission (68).

Characteristics	Overall *N* = 5,107 (100%)	HSV-1 Detected *N* = 670 (13%)	HSV-2 Detected *N* = 657 (13%)	VZV Detected *N* = 343 (6.7%)	TPA (LDT) Detected *N* = 334 (6.5%)
Sex					
Female	3,042 (60%)	458 (68%)	446 (68%)	212 (62%)	188 (56%)
Male	2,052 (40%)	209 (31%)	211 (32%)	130 (38%)	145 (43%)
Unknown	13 (0.3%)	3 (0.4%)	0 (0%)	1 (0.3%)	1 (0.3%)
Age group					
0–9	625 (12%)	71 (11%)	0 (0%)	11 (3.2%)	16 (4.8%)
10–19	480 (9.4%)	96 (14%)	47 (7.2%)	19 (5.5%)	53 (16%)
20–39	2,097 (41%)	292 (44%)	384 (58%)	116 (34%)	199 (60%)
40–59	1,034 (20%)	127 (19%)	163 (25%)	80 (23%)	59 (18%)
60+	871 (17%)	84 (13%)	63 (9.6%)	117 (34%)	7 (2.1%)
Regional health authority					
Winnipeg	2,854 (56%)	372 (56%)	368 (56%)	207 (60%)	127 (38%)
Prairie Mountain	449 (8.8%)	73 (11%)	57 (8.7%)	30 (8.7%)	34 (10%)
Interlake East	377 (7.4%)	48 (7.2%)	38 (5.8%)	28 (8.2%)	33 (9.9%)
Northern	714 (14%)	64 (9.6%)	128 (19%)	35 (10%)	107 (32%)
Southern	465 (9.1%)	83 (12%)	42 (6.4%)	33 (9.6%)	16 (4.8%)
Unknown	248 (4.9%)	30 (4.5%)	24 (3.7%)	10 (2.9%)	17 (5.1%)
Specimen type					
Female genital	1,287 (25%)	210 (31%)	321 (49%)	46 (13%)	132 (40%)
Male genital	708 (14%)	47 (7.0%)	186 (28%)	19 (5.5%)	119 (36%)
Rectal/perineal	192 (3.8%)	30 (4.5%)	40 (6.1%)	8 (5.5%)	23 (6.9%)
Oral	709 (14%)	227 (34%)	4 (0.6%)	7 (2.0%)	34 (10%)
Cutaneous	2,211 (43%)	156 (23%)	106 (16%)	263 (77%)	26 (7.8%)

To address these diagnostic limitations, a multiplex lesion panel assay was developed for the simultaneous detection of HSV-1, HSV-2, VZV, and TPA on a fully automated platform (Hologic Panther Fusion). This new assay combined previously validated HSV and VZV targets with the Tp47 target for TPA, enabling broad pathogen screening from a single swab.

The implementation of the lesion panel PCR provided several advantages over the previous workflow. TPA testing is done regardless of whether a physician specifically requested syphilis testing, reducing the risk of missed diagnoses of acute infections. Additionally, the turnaround time (TAT) for TPA testing improved substantially. Prior to implementation, TPA testing required an average of 17.8 days as samples were sent out to a reference laboratory **(Figure 3.2).** With the lesion panel, TAT was reduced to just 4 days, allowing for faster diagnosis and earlier treatment. Furthermore, the single-swab testing approach eliminated the need for separate swabs for HSV/VZV and TPA PCR, streamlining sample processing and improving laboratory efficiency.



Characteristic	TPA (LDT) Detected, N = 334[1]
TPA(NML)	
Detected	206 (62%)
Not Detected	7 (2.1%)
Not Ordered	121 (36%)

[1] n (%)

Figure 3.2. The time from collection to result for TPA request forwarded to the reference lab.

Screening samples for TPA identified 334 positives, 36% did not have an accompanying request for TPA reference testing. Of the 213 samples sent for reference testing, the average TAT was 17.8 days, with most results being received within 14-20 days after collection. Adapted with permission (68).

In the lesion panel PCR, the BGB target serves as an internal control to assess sample quality and DNA integrity. It can also provide a relative idea of the ratio of Human to TPA DNA in a sample. The average BGB Ct values were consistently lower than Tp47 Ct values (average of 27.4 vs. 32.7, respectively), demonstrating the higher abundance of human DNA in clinical specimens, **figure 3.3**. This difference was statistically significant (paired t-test, M = 5.34 cycles, 95% CI: 4.80–5.88, $p < 0.001$). Given that the human genome is approximately 2800 times larger than that of TPA (162), and that a Ct difference of ~3.3 cycles corresponds to a 10-fold difference in DNA concentration, the observed 6–8 cycle gap between BGB and Tp47 indicates that human DNA levels were roughly 100–300 times higher than TPA DNA in these samples.

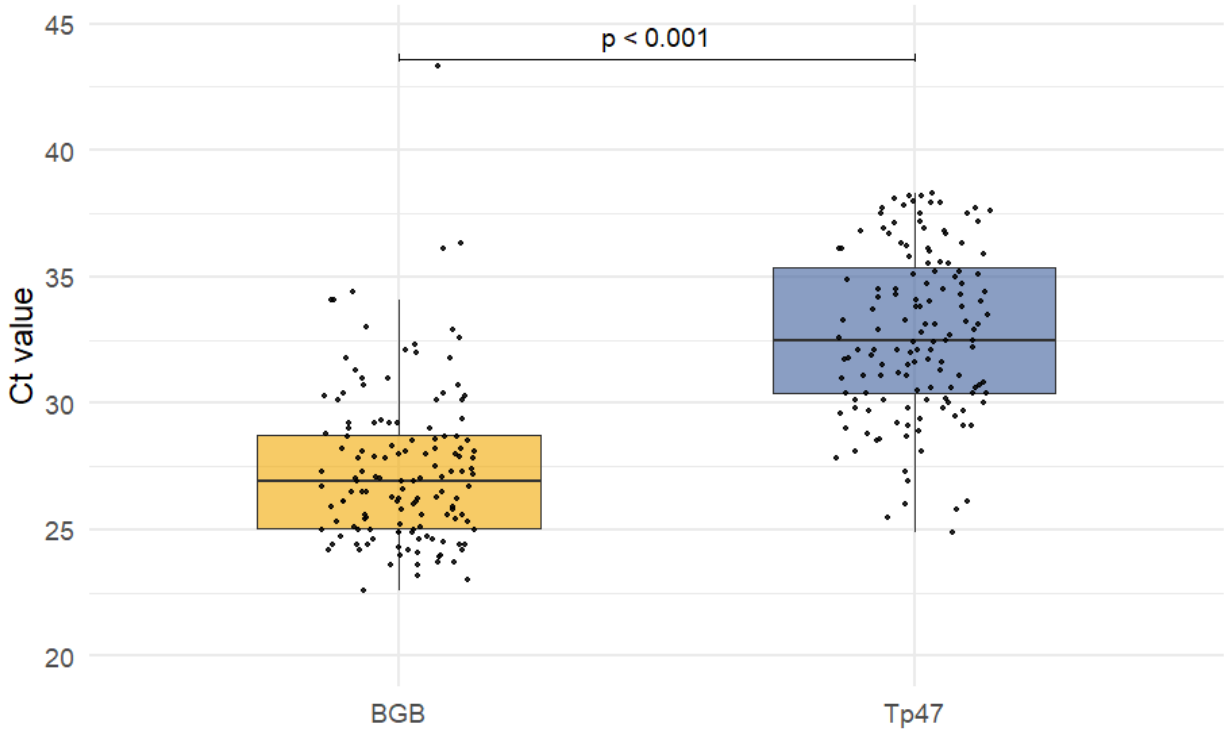


Figure 3.3. Average Ct values for Human BGB and TPA TP47, representing TPA DNA in samples.

Boxplots show the distribution of cycle threshold (Ct) values for the human β-globin (BGB) internal control and the TPA Tp47 PCR target across clinical specimens. Median BGB Ct values were lower than Tp47, showing a higher abundance of host DNA relative to TPA DNA. A paired t-test confirmed that the difference between targets was statistically significant ($p < 0.001$).

3.2. DNA Extraction Optimization for Long-Read Sequencing

3.2.1. Selective DNA Extractions

For microbiome enrichment, testing was performed using the Qiagen DNeasy Kit in combination with the NEBNext Microbiome. Enrichment Kit. Five samples were evaluated, and the method produced only a minimal improvement in TPA detection, with the Tp47 target showing an average gain of 0.8 Ct. The human beta-globin (BGB) target, however, decreased by an average of 2.6 Ct. These modest changes were insufficient to overcome the overwhelming background of human DNA or to generate the yields required for library preparation. As noted in the PCR section, the predominance of human DNA in clinical specimens remained a major obstacle for downstream sequencing.

The impact of host DNA depletion using the Qiagen Microbiome Kit was evaluated by comparing TPA (Tp47) and BGB Ct values to the no enrichment screening results. Overall, the microbiome enrichment process consistently increased Tp47 Ct values by an average of 5.03 cycles (SD = 1.59), indicating a substantial reduction in detectable TPA DNA (**Table 3.2**). The most pronounced Ct shift was observed in Swab 9 (7.24 Ct increase), while the smallest shift occurred in Swab 5 (3.38 Ct increase). Two samples (Swabs 3 and 4) did not yield microbiome kit results for Tp47, likely reflective of the low-abundance TPA DNA in metagenomic samples. A paired t-test confirmed that the increase in Tp47 Ct values was statistically significant ($p < 0.001$).

For human BGB DNA, the microbiome kit resulted in an even greater average Ct increase of 6.35 cycles (SD = 2.77) compared to the original screening values (**Table 3.2**). The largest reduction in BGB DNA was observed in Swab 8 (10.11 Ct increase), while the

smallest was in Swab 3 (0.77 Ct increase). A paired t-test confirmed that the reduction in host DNA was statistically significant ($p < 0.001$). The results indicate that while the microbiome enrichment kit effectively reduced host DNA, it also led to a measurable loss of TPA DNA.

These findings demonstrate the challenge of working with metagenomic samples dominated by host DNA. The NEB Microbiome Enrichment Kit achieved minimal reductions in host DNA, while the Qiagen Microbiome Kit demonstrated improved host DNA depletion but also led to significant loss of TPA DNA. Neither approach was effective enough to recover sufficient TPA DNA for downstream long-read sequencing.

Table 3.2. Comparing extraction results (tp47 and BGB) between the Qiagen Microbiome kit and the no enrichment screening.

	Tp47 Ct values		
Swab	No enrichment screening result	Microbiome kit	Ct difference
Swab 1	28.53	32.66	-4.13
Swab 2	32.1	36.4	-4.30
Swab 3	36.28	N/A	-
Swab 4	30.91	N/A	-
Swab 5	32.35	35.73	-3.38
Swab 6	30.8	34.43	-3.63
Swab 7	29.61	35.13	-5.52
Swab 8	27.61	34.64	-7.03
Swab 9	28.73	35.97	-7.24
		Average Ct difference	-5.03
		Standard Deviation	1.59
	BGB Ct values		
Swab	No enrichment screening result	Microbiome kit	Ct difference
Swab 1	24.35	29.13	-4.78
Swab 2	27.98	35.39	-7.41
Swab 3	30.69	31.46	-0.77
Swab 4	27.35	35.36	-8.01
Swab 5	23.57	30.56	-6.99
Swab 6	26.68	31.6	-4.92
Swab 7	24.89	30.17	-5.28
Swab 8	23.24	33.35	-10.11
Swab 9	25.15	34.01	-8.86
		Average Ct difference	-6.35
		Standard Deviation	2.77

3.2.2. Total Nucleic Acid Extractions

Considering the results from **Section 3.2.1**, which highlights the limitations of selective DNA extractions, a comparative evaluation of three total nucleic acid extraction methods was conducted. Tp47 qPCR was used to assess TPA DNA recovery across the KingFisher Flex, BioMérieux eMAG, and Qiagen DNeasy platforms. Friedman testing signaled significant overall differences were present among methods ($p = 0.0018$), shown in **Figure 3.4.** Pairwise Wilcoxon signed-rank tests found that eMAG produced significantly lower Ct values than both the KingFisher ($p = 0.002$) and Qiagen ($p = 0.006$), representing the highest recovery of *T. pallidum* DNA. Although KingFisher outperformed Qiagen in 8 of 10 paired samples, the difference in Ct values did not reach statistical significance ($p = 0.084$).

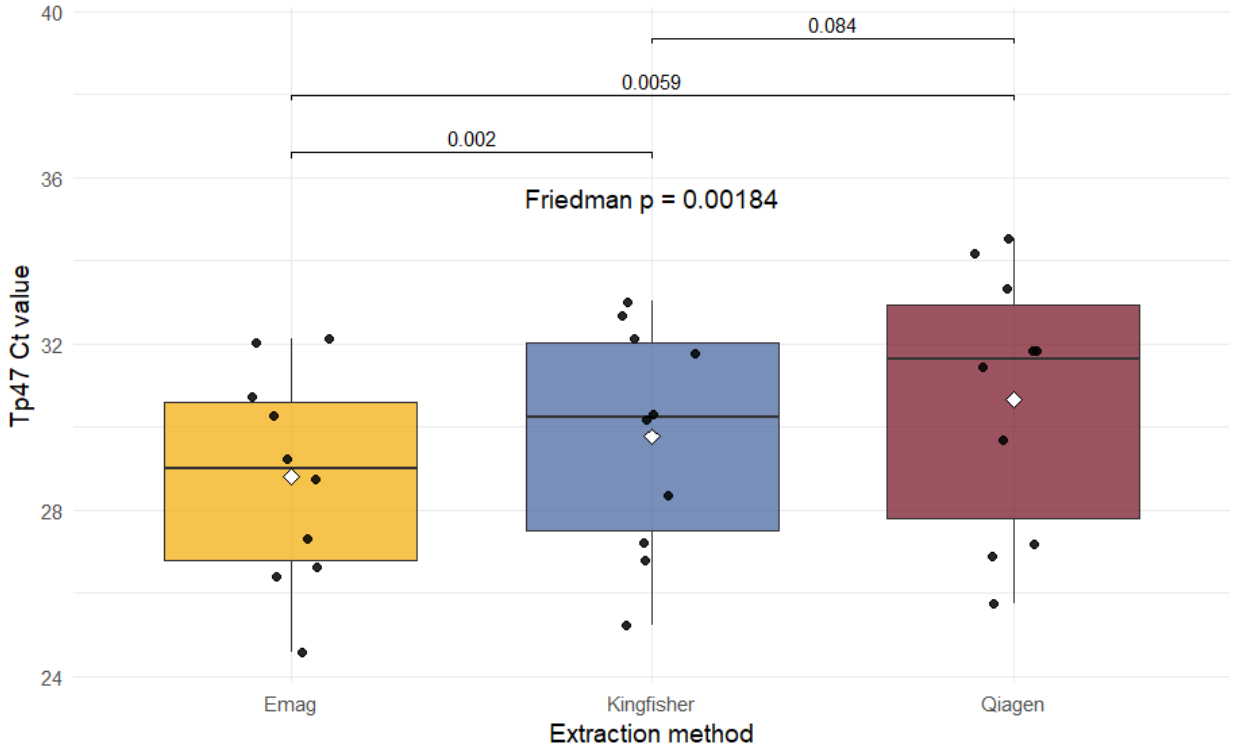


Figure 3.4. Comparison of Tp47 Ct values across extraction methods.

Tp47 qPCR was used to assess TPA DNA recovery from 10 clinical swabs extracted with the KingFisher Flex, bioMérieux eMAG, and Qiagen DNeasy methods. Friedman testing found significant overall differences ($p = 0.0018$). Wilcoxon signed-rank test indicated that eMAG yielded lower Ct values than both KingFisher ($p = 0.002$) and Qiagen ($p = 0.006$), while KingFisher and Qiagen did not differ significantly ($p = 0.084$). Lower Ct values indicate greater DNA recovery.

To complement the qPCR analysis of DNA recovery, total nucleic acid yield was also quantified across extraction methods. DNA concentration was measured using the Qubit dsDNA High Sensitivity assay (**Figure 3.5**). The KingFisher method produced the highest yields, with an average of 6.71 ng/μL across the 10 samples, compared with 4.23 ng/μL for eMAG and 4.40 ng/μL for Qiagen. The Friedman test again confirmed significant overall differences were present ($p = 0.0055$), and Wilcoxon signed-rank test showed that KingFisher concentrations were significantly higher than both eMAG ($p = 0.014$) and Qiagen ($p = 0.020$), while eMAG and Qiagen did not differ ($p = 0.846$).

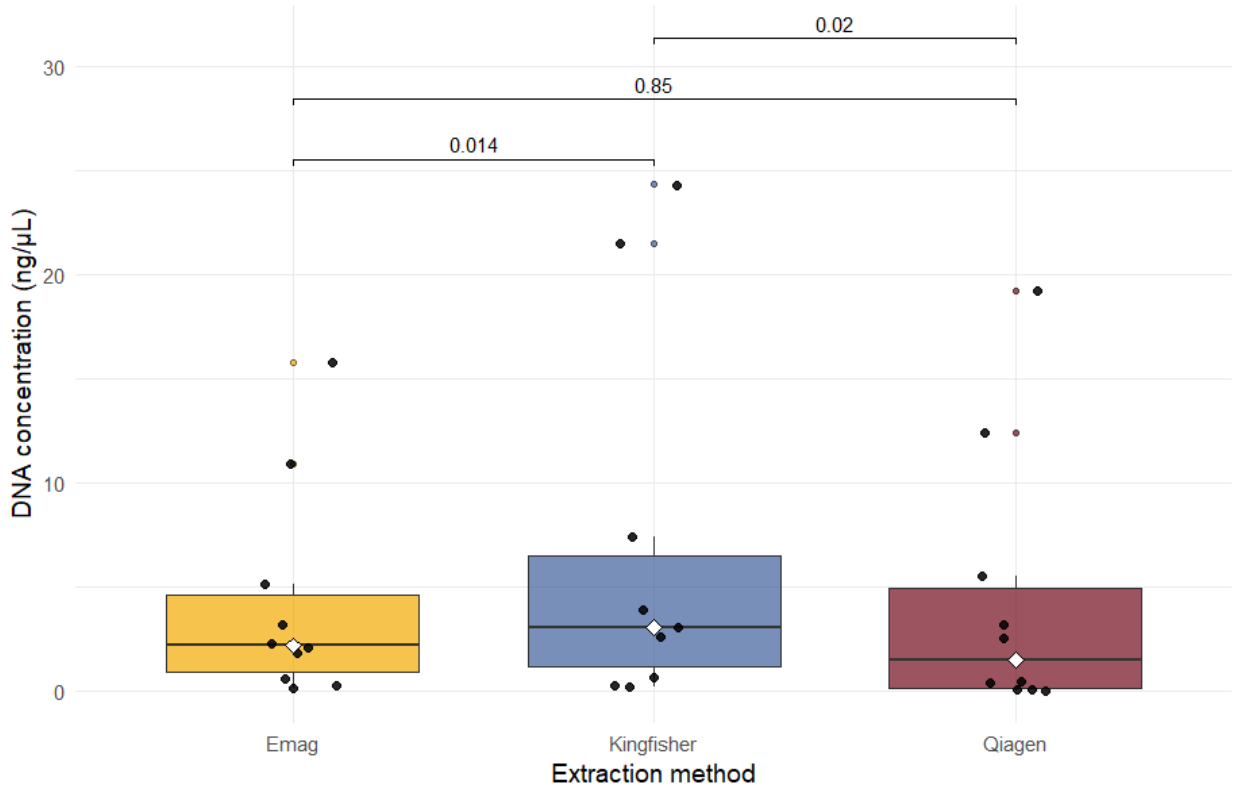


Figure 3.5. DNA concentration by extraction method.

DNA yields were measured using the Qubit dsDNA HS assay across 10 swab samples. Friedman testing found significant overall differences ($p = 0.0055$). Wilcoxon signed-rank test indicated that KingFisher produced higher concentrations than both eMAG ($p = 0.014$) and Qiagen ($p = 0.020$), while eMAG and Qiagen did not differ ($p = 0.846$).

Lastly, DNA integrity was assessed by comparing fragment sizes across extraction methods. DNA fragment size was evaluated using the Agilent 2200 TapeStation to assess suitability for long-read sequencing. KingFisher produced the longest fragments on average (22,629 bp), compared with eMAG (11,707 bp) and Qiagen (6,127 bp). Across the dataset, KingFisher produced longer fragments than Qiagen in every measurable sample (**Figure 3.6**). The Friedman test again identified significant overall differences ($p = 0.0057$), and Wilcoxon signed-rank test confirmed that KingFisher yielded significantly longer fragments than both eMAG ($p = 0.031$) and Qiagen ($p = 0.031$). Differences between eMAG and Qiagen were not significant ($p = 0.094$). Several samples (4, 5, 6, and 10) failed to register detectable fragments, which typically correlated with Qubit concentrations <1 ng/μL.

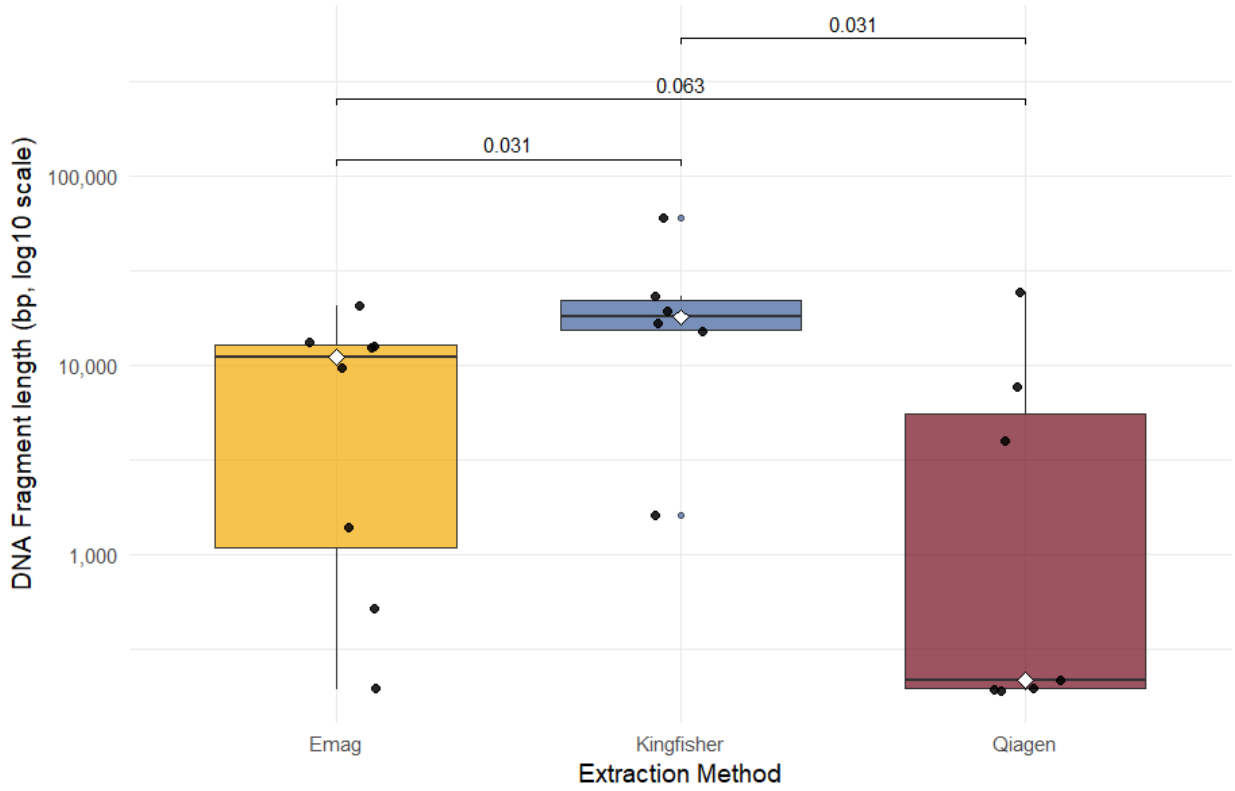


Figure 3.6 DNA fragment length by extraction method.

Fragment size was assessed using the Agilent TapeStation. Friedman testing found significant overall differences ($p = 0.0057$). Wilcoxon signed-rank test indicated that KingFisher produced longer fragments than both eMAG ($p = 0.031$) and Qiagen ($p = 0.031$), while eMAG and Qiagen did not differ ($p = 0.094$). Several low-yield samples failed to register detectable fragments.

Taken together, eMAG demonstrated the strongest performance in terms of DNA recovery by qPCR, whereas KingFisher Flex consistently provided higher DNA concentrations and longer fragment lengths. Because long-read sequencing platforms benefit substantially from longer DNA fragments, and KingFisher additionally offers higher throughput via a semi-automated 96-well format, this platform was selected for all subsequent TPA DNA extraction and sequencing experiments. Its ready availability in the laboratory further supported its adoption as the primary method.

3.3. Whole genome Amplification Optimization

Due to the low concentration of TPA DNA recovered from clinical swabs, no extraction method proved sufficient to meet the minimum input requirements for Oxford Nanopore Technologies (ONT) library preparation. While hybridization-based bait capture methods can enrich low-abundance bacterial DNA in metagenomic samples, they are optimized for short-read platforms such as Illumina and rely on ~120 bp fragment capture. These approaches are currently incompatible with long-read sequencing, which requires preservation of high-molecular-weight DNA.

To address this limitation, multiple displacement amplification (MDA) was explored as a method to amplify the entire DNA content of each sample while preserving long DNA fragments (>10 kbp). MDA uses random hexamer primers and the high-fidelity phi29 DNA polymerase to perform isothermal, strand-displacing replication of genomic DNA. This method is well-suited for whole genome amplification from low-input samples and has been successfully used in other microbial metagenomic contexts (163–165). An additional

approach, selective whole genome amplification (SWGA), was also evaluated to improve amplification specificity for TPA DNA while minimizing host DNA background.

3.3.1. Random-Primed Whole Genome Amplification

Random-primed WGA (rpWGA) using the REPLI-g Advanced Single Cell Kit was performed on six TPA–positive swab extracts with Ct values from 25.31 to 33.36 (M = 27.93, SD = 2.90). While two samples showed substantial reductions in Tp47 Ct values (up to ~9.5 cycles, corresponding to >700-fold amplification of TPA DNA), most exhibited only modest improvements. Four of the six samples had ΔCt values ≤3 cycles (≤10-fold change), and two were below 1.5 cycles (~2-fold), indicating that rpWGA generally provided limited improvement of target abundance.

Following WGA, samples were treated with T7 Endonuclease I to remove branched DNA structures and facilitate Nanopore library preparation. Qubit quantification confirmed sufficient total DNA for sequencing (>20 ng/μL). However, when reads were mapped to the TPA SS14 reference genome using Minimap2 (as described in the initial mapping step of **Methods 2.6.2.2**) and visualized using Geneious, genome coverage was poor. The resulting assemblies achieved only ~12% genome breadth at an average depth of 0.9X (**Figure 3.7**).

These results indicate that rpWGA with the REPLI-g system did not effectively enrich TPA DNA. It is likely that the amplification was non-specific, favoring human genomic DNA or other non-target sequences present in the sample. Unfortunately, BGB (host DNA) was

not assessed post-amplification, but the low recovery of TPA reads strongly supports a predominance of off-target amplification.

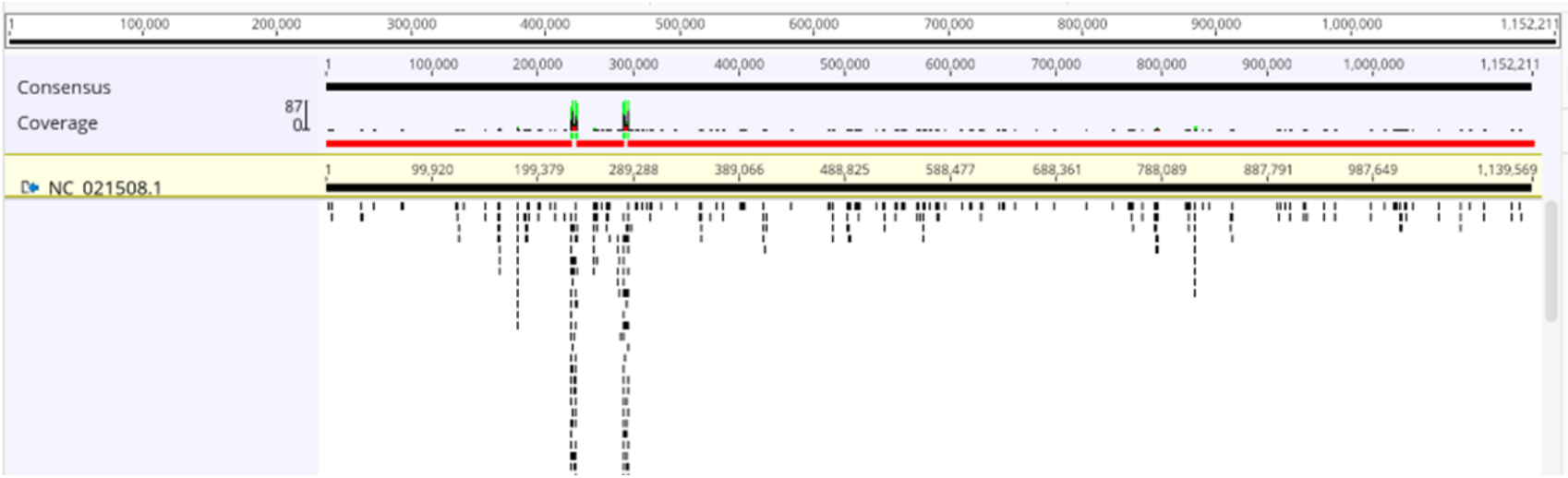


Figure 3.7. Mapping rpWGA reads to the SS-14 reference genome.

The reference sequence (NC_021508.1) is the black bar highlighted in yellow. The red bar above the reference sequence indicates regions of the genome with less than 2 supporting reads. The coverage histogram is barely visible with only two positions showing >2x depth (histogram indicates a maximum depth of 87x). The sequencing reads are represented as black boxes under underneath the reference sequence, illustrating the poor depth and coverage.

3.3.2. Selective Whole Genome Amplification

Selective whole genome amplification (SWGA) using EquiPhi29 DNA polymerase and the Pal 12 primer set from Thurlow et al. was applied to 11 clinical swab samples to selectively enrich TPA DNA (Thurlow 2022). All samples showed a reduction in Tp47 Ct values following amplification, indicating enrichment of the TPA target. Ct reductions ranged from 5.5 to 12.1 cycles (M = 9.5, SD = 2.22), corresponding to an estimated 45- to >4,000-fold increase in target abundance, see table 3.2.

BGB Ct values increased by 1.4 to 11.2 cycles (M = 6.60, SD = 2.39) following amplification, indicating that host DNA was not amplified, see **Table 3.3**. This pattern demonstrates that the Pal 12 primers selectively enriched TPA DNA, providing strong evidence for their specificity.

Table 3.3. Changes in Tp47 and BGB Ct values before and after selective whole genome amplification (SWGA) of clinical swab extracts using the Pal 12 primer set.

	Tp47 Ct values		
Sample ID	Pre-SWGA	Post-SWGA	Ct difference
1	25.23	13.54	11.7
2	27.23	15.14	12.1
3	26.80	17.66	9.1
4	32.12	24.09	8.0
5	25.58	15.15	10.4
6	25.65	20.20	5.5
7	25.65	19.65	6.0
8	26.58	17.25	9.3
9	25.65	14.26	11.4
10	25.73	15.81	9.9
11	24.19	13.25	10.9
		Average Ct difference	9.49
		Standard Deviation	2.22

	BGB Ct values		
Swab	Pre-SWGA	Post-SWGA	Ct difference
1	23.76	29.40	-5.64
2	27.06	28.50	-1.44
3	23.19	31.11	-7.92
4	25.07	30.20	-5.13
5	28.81	35.67	-6.86
6	28.79	40.00	-11.21
7	26.28	32.54	-6.26
8	24.64	32.72	-8.08
9	25.07	31.00	-5.93
10	24.78	31.13	-6.35
11	23.85	31.63	-7.78
		Average Ct difference	-6.60
		Standard Deviation	2.39

Qubit measurements confirmed sufficient total DNA for sequencing (>32 ng/μL). After T7 endonuclease digestion to remove branched structures. Nanopore sequencing, reference-guided mapping (as described in the initial mapping step of **Methods 2.6.2.2**) and visualization using Geneious showed that some SWGA-enriched samples achieved near-complete genome recovery. Several samples reached 100% genome breadth with >100x mean coverage, and one exceeded 1,000x mean depth with a maximum of 28,857x in one region (**Figure 3.8**). These results confirm that SWGA is an effective method for selectively amplifying TPA DNA from low-yield clinical samples, enabling successful long-read sequencing and downstream genomic analyses.

However, despite successful enrichment and high coverage, read depth across the genome remained uneven. Certain regions were amplified to extreme depths (>10,000x), while others had low or no coverage, resulting in disproportionate coverage. This uneven distribution complicated downstream analysis, particularly subsampling for genome assembly and variant calling, is further discussed in Chapter 4.3.

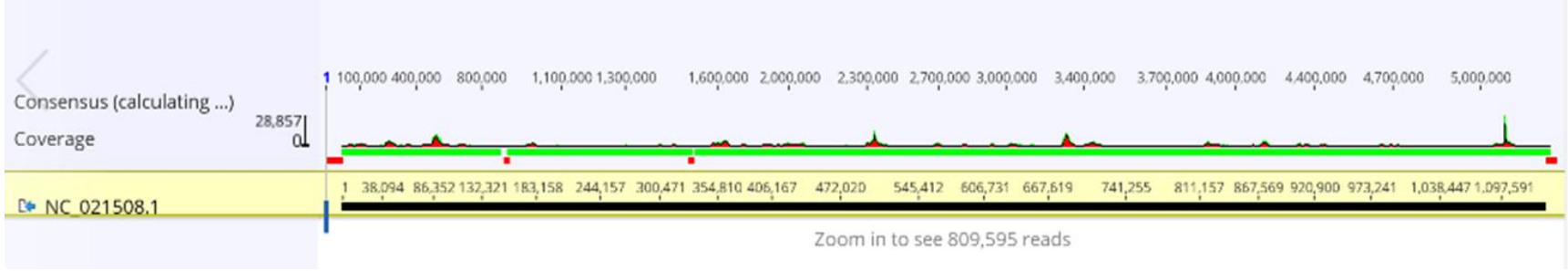


Figure 3.8. Mapping SWGA reads to the ss-14 reference genome.

The reference sequence (NC_021508.1) is the black bar highlighted in yellow. The green bar above the reference sequence indicates regions of the genome with more than 50x depth, while the red boxes indicate depth of less than 2 reads. Mean read depth was 1100x coverage (histogram indicates a maximum depth of 28,857x). The total number of reads cannot be visualized at this scale.

3.3.3. Quantification of Amplification Efficiency Across WGA Methods

Absolute quantification of Tp47 copies/µL before and after WGA was performed to assess the relative amplification efficiency of rpWGA versus SWGA. Copy number estimates were derived from Tp47 Ct values using a custom dPCR-based quantification curve (see Section 2.3.3).

For rpWGA (n = 6), copy number gains were modest. Post-WGA concentrations ranged from 122 to 2.81×10^5 copies/µL, with a median fold-change of 5.5 (IQR = 6.5). In several cases, enrichment was minimal (<2-fold), indicating the varying efficiency of this method for amplifying TPA DNA. A Wilcoxon signed-rank test confirmed that rpWGA significantly increased copy number compared to input ($p = 0.016$), though the overall gains were small (**Figure 3.9**, blue bars).

In contrast, samples processed with SWGA (n=20) showed dramatically improved yields, with post-amplification copy numbers between 7.33×10^3 and 2.13×10^7 copies/µL (**Figure 3.9**, dark red bars). Median fold-change was 2.6×10^3 (IQR = 4.2×10^3), several orders of magnitude higher than rpWGA. A Wilcoxon signed-rank test confirmed that SWGA significantly increased copy number relative to input ($p = 1.9 \times 10^{-6}$). Direct comparison of fold-change distributions between rpWGA and SWGA by Mann–Whitney testing further supported the superior performance of SWGA ($p = 0.0015$). One exception, WGA-sample 22, yielded an apparent copy number of <1 copy/µL despite adequate input, likely reflecting DNA loss during cleanup rather than amplification failure.

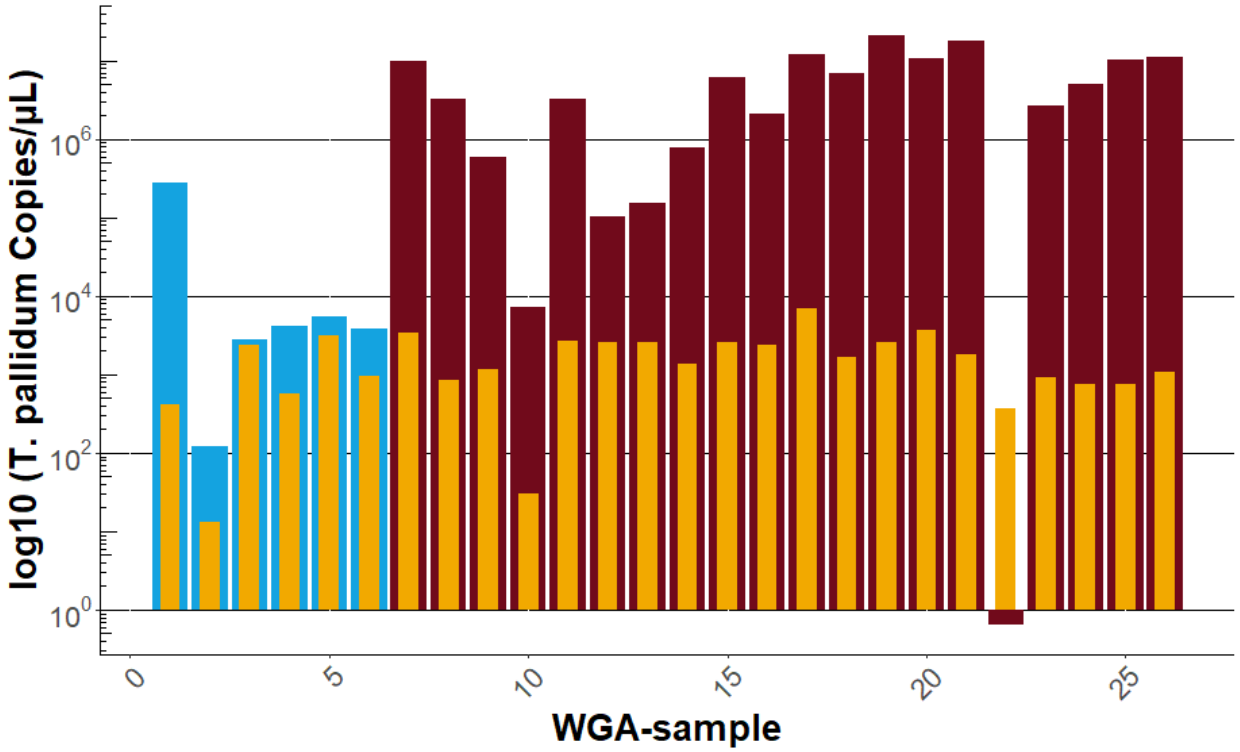


Figure 3.9. Comparison of T. pallidum Tp47 copy numbers before and after whole genome amplification.

Yellow bars represent original (pre-WGA) copy number for each sample. Colored bars indicate post-WGA copy numbers obtained using either random-primed WGA (rpWGA, blue) or selective WGA (SWGA, dark red). Copy numbers are plotted on a log10 scale. SWGA consistently resulted in several orders of magnitude greater amplification than rpWGA.

Quantification of total DNA by Qubit confirmed that both rpWGA and SWGA significantly increased DNA concentrations to sufficient amounts for ONT library preparation. For rpWGA, median concentrations rose from 3.3 ng/μL pre-amplification to 30.5 ng/μL post-amplification (median fold-change = 7.9, IQR = 23.9; Wilcoxon signed-rank test, $p = 0.031$). For SWGA, median concentrations increased from 3.5 ng/μL to 47.5 ng/μL (median fold-change = 8.3, IQR = 11.1; Wilcoxon signed-rank test, $p < 0.001$). A Mann–Whitney test found no significant difference between the fold-change distributions of rpWGA and SWGA ($p = 0.72$), see **Figure 3.10**.

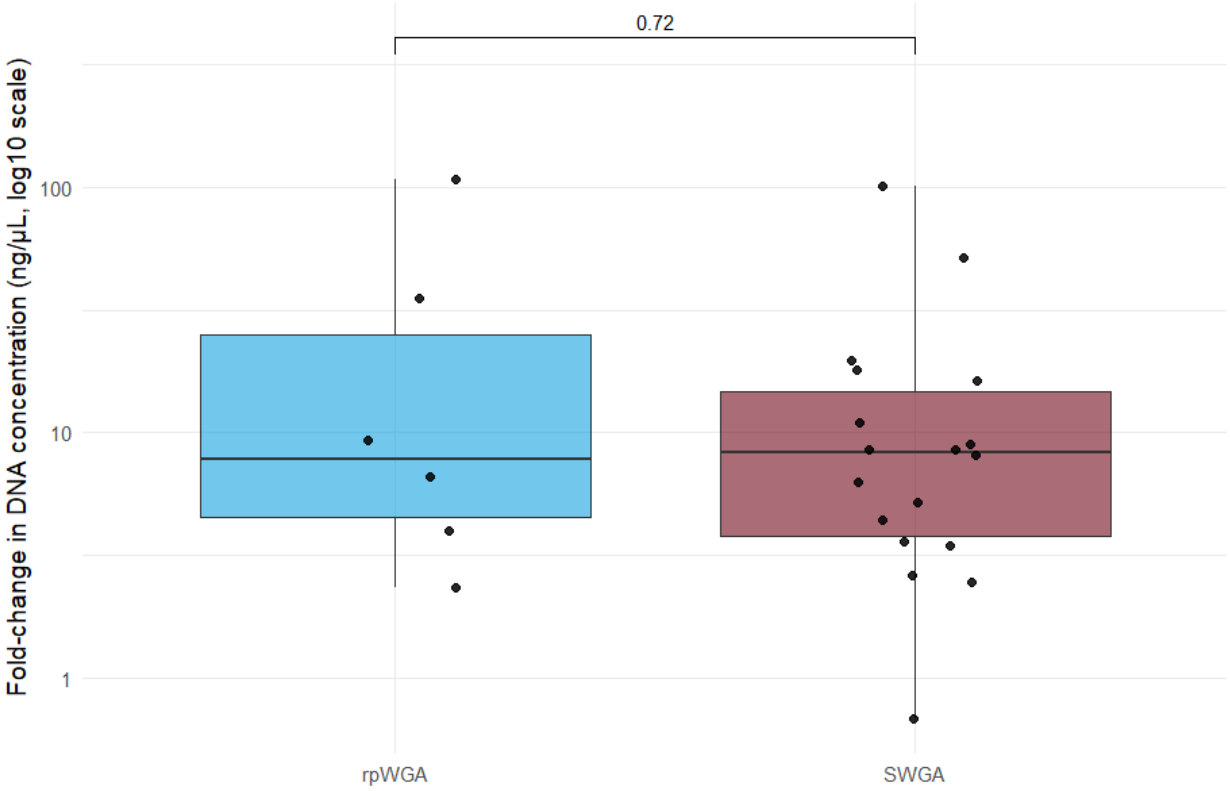


Figure 3.10. Total DNA concentration before and after WGA.

DNA concentration was measured for samples amplified using rpWGA (blue) and SWGA (dark red). Both methods significantly increased DNA concentration (rpWGA: $p = 0.031$; SWGA: $p < 0.001$). Median fold-changes were comparable and a Mann–Whitney test found no significant difference between methods ($p = 0.72$).

When considering both measures together, rpWGA and SWGA produced comparable gains in total nucleic acid concentration, but SWGA yielded substantial and statistically significant increases in TPA copy number. This indicates that while rpWGA amplifies bulk DNA indiscriminately, SWGA provides greater specificity for TPA, consistently generating significantly higher concentrations of target DNA. This specificity makes SWGA the more suitable amplification strategy for downstream ONT sequencing applications.

3.4. Development of a Long-Read Sequencing Protocol for TPA

To optimize sequencing output from low-input clinical samples, We adopted a library preparation workflow that preserved as much DNA as possible at each step. Although SWGA improved TPA DNA recovery, a substantial proportion of the total DNA remained host or commensal in origin. Libraries were prepared using an ONT ligation-based protocol for whole-genome amplification (SQK-LSK112), incorporating the Native Barcoding Expansion kit (EXP-NBD196) for multiplexing. The workflow, described in **Methods 2.4.1**, was further modified to maximize DNA recovery at each step and allow two-step loading of each flow cell.

3.4.1. DNA Requirements and Sequencing Strategy

The initial sequencing experiments were performed on TPA–positive clinical swab extracts that had been amplified using rpWGA. Libraries were prepared using the ONT ligation-based workflow with native barcoding as described in **Methods 2.4.1**, and sequencing was carried out on R9.4.1 flow cells under standard conditions (without adaptive sampling). These runs were performed to evaluate whether sufficient data could

be obtained from rpWGA-amplified material and to optimize strategies for extending sequencing throughput.

The first sequencing run used a flow cell loaded with four samples and was configured for 72 hours without a duplicate library available for reloading. Initial pore activity began at approximately 50%, indicating that half of the sequencing channels were actively processing DNA strands. Activity declined to <10% within 18 hours, and with output plateaued, the run was terminated early yielding 2.42 Gb of DNA.

For the second sequencing run, enough DNA was available to prepare two libraries from the same barcoded pool. This allowed implementation of a flow cell wash and reload strategy. As before, pore activity started at ~50% and dropped below 10% within 23 hours. The run was paused, the flow cell was washed using ONT's flow cell wash kit, and the second aliquot of the same barcoded library was reloaded. This restored pore activity to ~40%, allowing an additional 20 hours of sequencing before activity again declined. Although sequencing output plateaued around 44 hours, the run was allowed to continue for the full 60 hours. The total yield reached 7.38 Gb of DNA.

The third run followed a similar pattern. Pore activity dropped below 10% by 20 hours, prompting a wash and reload that restored activity to ~50%. Sequencing continued for an additional 20 hours before activity again declined. The final yield was 5.33 Gb.

These results are summarized in **Figure 3.11**, which visualizes the pore activity states across 2-hour time increments for each run. The plots illustrate the early decline in sequencing activity during the first run, which lacked a reload, compared to the partial

recovery and extended runtime observed in subsequent runs where a wash and reload was performed. Implementing the reload strategy substantially improved sequencing throughput, extending active sequencing time from ~18–20 hours to 40–50 hours and increasing total data output by 2- to 3-fold relative to the initial single-loaded run.

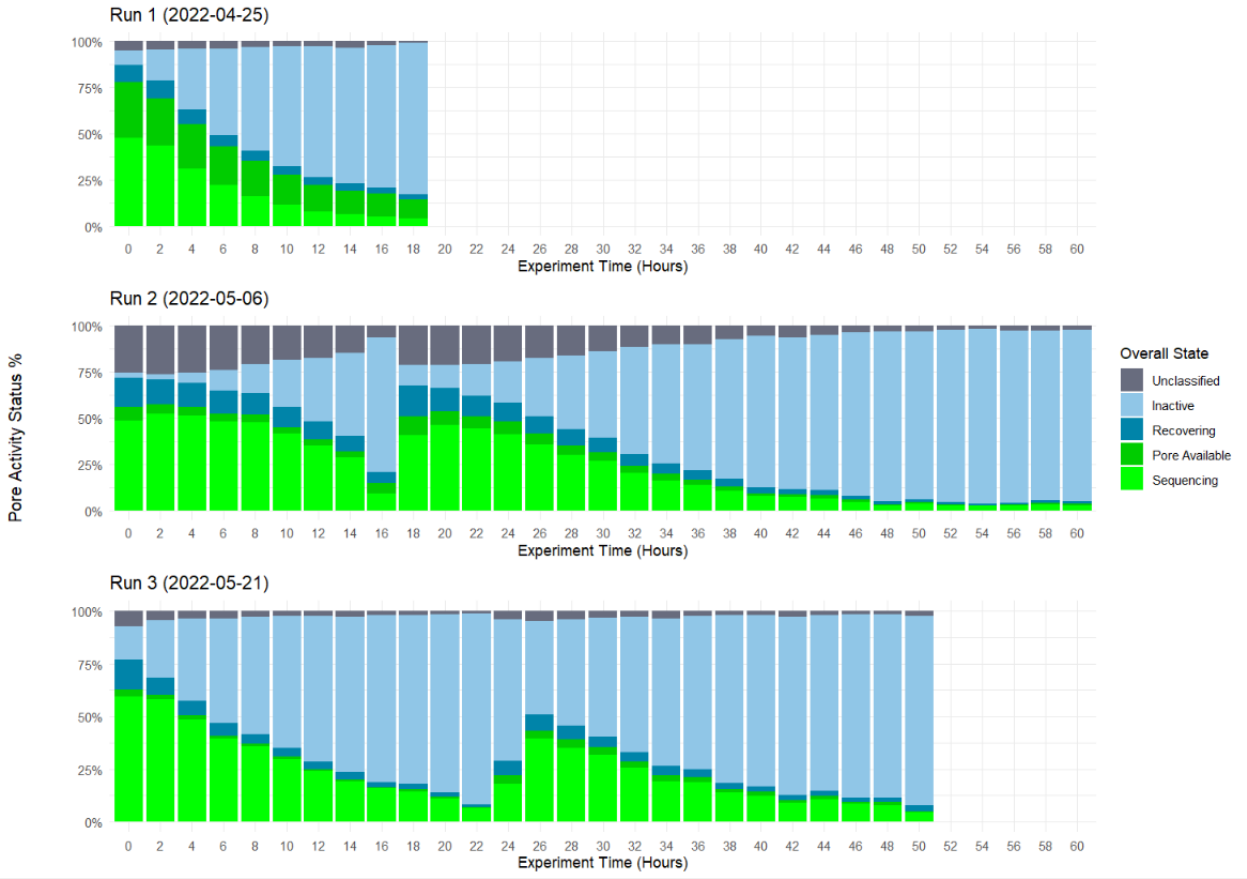


Figure 3.11. Nanopore sequencing activity with and without flow cell reloading.

Run 1) Sequencing was run for 18 hours and discontinued when pore activity fell below 10%. Run 2) Sequencing was run for 18 hours and then a flow cell wash and library reload was performed and sequencing was run till hour 60. Run 3) Sequencing was run for 22 hours and the flow washed-reloaded and sequencing continued until 50 hours, when pore activity fell below 10%.

3.4.2. Multiplexing Sample Libraries

The ONT MinION flow cell has a theoretical capacity of approximately 50Gb (166). In an ideal scenario using pure TPA isolates, this would support multiplexing up to 96 samples at roughly 450x coverage each. However, clinical samples are metagenomic, meaning the majority of sequencing reads may be from non-target DNA. This drastically reduces the effective yield of TPA reads and significantly limits multiplexing capacity.

To evaluate how multiplexing impacts genome recovery, sequencing runs were performed in 2022, using barcoded libraries containing between 4 and 24 TPA-positive samples (**Table 3.4**). Runs with smaller pools (4–6 samples) consistently produced complete assemblies, while increasing the number of multiplexed samples led to diminishing returns. Across 12 sequencing runs, those with 4–6 samples yielded the highest number of near-complete TPA genomes. In contrast, runs with more than six samples often failed to generate sufficient read depth per sample, resulting in incomplete assemblies.

Interestingly, the total number of sequencing reads per flow cell remained relatively constant, averaging approximately 5 million reads, regardless of the number of barcoded samples. This indicates that sequencing output is limited by the physical constraints of the flow cell and increasing the number of samples in libraries will not increase the total data obtained. Thus, as larger sequencing libraries are made, the total yield is distributed across more samples, reducing the sequencing depth per sample.

These findings demonstrate that flow cell sequencing capacity is fixed and that sample multiplexing must be balanced against the required sequencing depth per genome. For TPA, a conservative multiplexing range of 4–6 samples per flow cell provides the best trade-off between cost-efficiency and genome recovery success.

Table 3.4. Effect of sample multiplexing on TPA genome recovery across ONT flow cells.

Run ID	# of Multiplexed Samples	Complete Genomes	# of Reads Obtained (log)
2022-10-14 AS	4	3	6.03
2022-10-14 Mk1C	4	3	6.09
2022-12-28 AS	4	3	6.82
2022-12-28 Mk1C	4	3	6.31
2022-07-08 AS	5	2	6.92
2023-01-27 Adh	6	3	6.71
2023-01-27 ONT	6	3	6.85
2022-09-18 AS	8	2	6.53
2022-09-29 Mk1C	8	2	6.47
2023-03-29 8x	8	1	6.42
2023-03-29 16x	16	0	6.42
2023-03-29 24x	24	1	6.56

3.5. Standard vs. Adaptive Nanopore Sequencing

ONT sequencing enables real-time data acquisition and supports adaptive sampling (AS), a strategy that enriches for target DNA by selectively sequencing molecules that align to a reference genome **(Figure 1.3).** In this study, AS was used to enrich for TPA DNA using the SS14 reference genome (NC_021508.1) as the alignment target. DNA molecules that failed to align were ejected from pores within the first few hundred bases, allowing pores to sequence new molecules. This selective approach was compared against standard sequencing (SS) by preparing four libraries in duplicate and sequencing one set under each condition. The total read count, genome coverage and read depth was calculated with Samtools (**Methods 2.6.2.2).**

Across 20 samples, AS increased the number of TPA reads in 18 of them **(Table 3.5)**. The median AS/SS read count ratio was 1.84 (range: 0.63-2.74), corresponding to median gains of 74,000 reads per sample (range: -396 to +878,547). In several cases, AS more than doubled the number of classified reads, including Sample 34 ($9.9x10^5$ to $1.85x10^6$ reads) and Sample 61 ($4.67x10^5$ to $1.28x10^6$ reads). Samples 33 and 57 showed a net decrease in reads, they also had poor depth and coverage overall indicating low input specimens. A Wilcoxon signed-rank test confirmed that AS produced significantly higher read counts than SS ($p < 0.001$).

Table 3.5. Comparison of TPA read counts obtained by standard sequencing (SS) and adaptive sampling (AS)

Sample ID	SS TPA Classified Reads	AS TPA Classified Reads	Ratio (AS/SS)	Δ TPA Reads (AS−SS)
33	3659	3298	0.90	-361
34	990000	1848663	1.86	858,663
35	31658	55522	1.75	23,864
36	81880	150023	1.83	68,143
37	872504	1645988	1.88	773,484
41	402052	746907	1.85	344,855
42	23886	44009	1.84	20,123
43	5322	9362	1.75	4,040
44	438540	824266	1.87	385,726
45	119681	220496	1.84	100,815
49	873326	986066	1.12	112,740
50	148059	166452	1.12	18,393
51	199422	227988	1.14	28,566
52	212735	241508	1.13	28,773
53	397957	454536	1.14	56,579
57	1072	676	0.63	-396
58	157078	425076	2.7	267,998
59	3974	8394	2.11	4,420
60	4363	10484	2.40	6,121
61	467224	1281353	2.74	814,129

When comparing sequencing depth and genome coverage, AS consistently outperformed SS (**Figure 3.12**). Low coverage samples had the largest benefit, Sample 33 increased from 60.2% (SS) to 75.1% (AS) coverage, and Sample 57 from 20.4% (SS) to 38.3% (AS). Overall, 19 of 20 samples showed improved or equivalent coverage with AS. Sample 52 was the only exception, with 99.9% coverage using SS and 99.8% with AS, both representing near-complete recovery.

Like the significant improvement in total read count, the impact of AS on read depth was also notable. Median read depth across all samples was 268.8x for AS compared to 161.6x for SS, representing a 1.66-fold increase (1.13-fold to 2.74-fold). In some cases, AS more than doubled the read depth; Sample 61 increased from 474x to 1301x.

Together, these results demonstrate that AS improves TPA genome recovery in metagenomic samples with low target abundance. The strategy enhances sequencing efficiency making it a valuable tool in workflows where DNA quantity is a limiting factor.

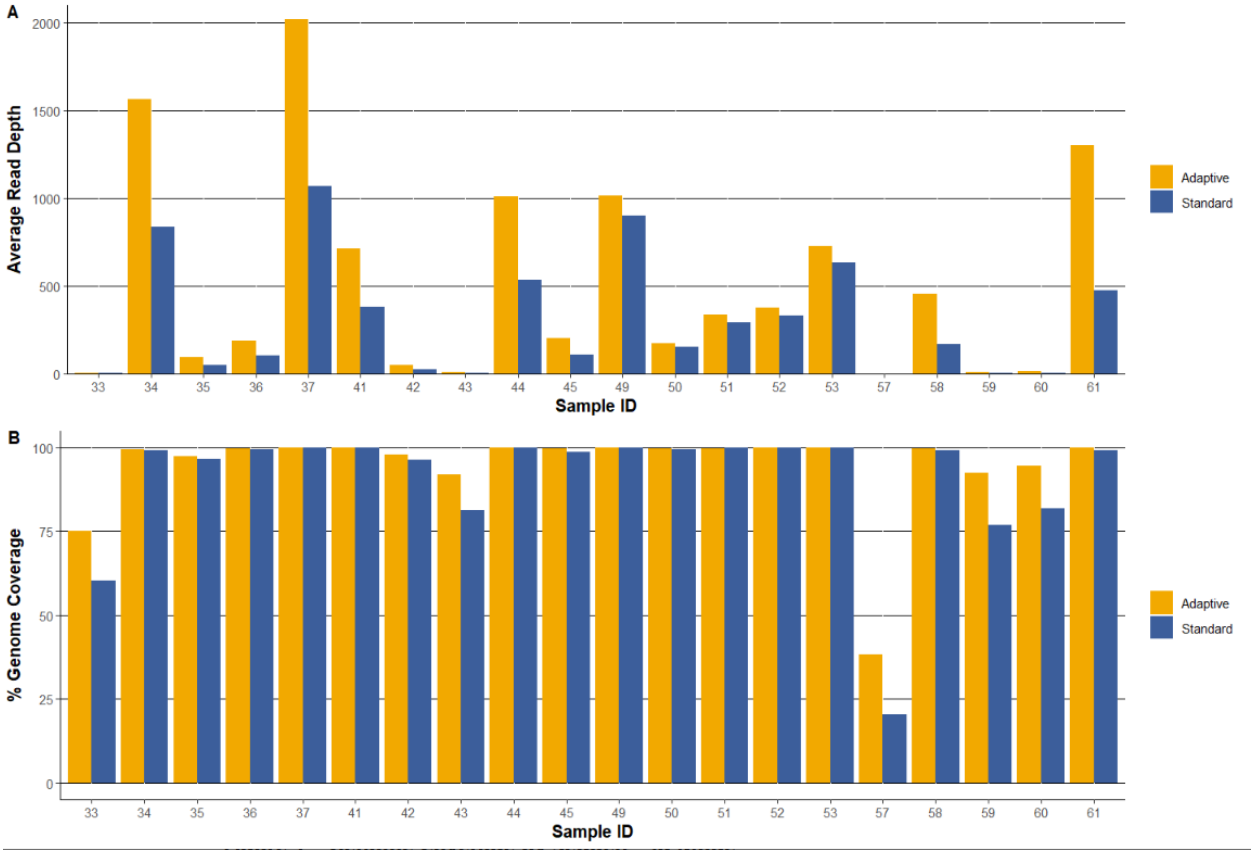


Figure 3.12. Comparative Performance of Adaptive Sampling (AS) and Standard Sequencing (SS) in TPA sequencing

(A) Average read depth per sample for AS (gold) and SS (blue) runs. AS consistently produced higher read depth across all samples.
(B) Percent genome coverage, defined as the proportion of the reference genome covered by at least one read. AS improved coverage in 19 of 20 samples. Sample 52 was the only case where SS slightly outperformed AS (99.9% vs. 99.8%), though both runs achieved near-complete genome recovery.

3.6. Bioinformatics Pipeline for Long-Read Data Processing and Genome Assembly

Following sequencing with ONT, raw reads require systematic processing to ensure accurate downstream analysis. A custom bioinformatics pipeline was developed to manage these steps, beginning with basecalling and demultiplexing, followed by host-read removal and quality control. Reads classified as TPA were then assembled with both *de novo* and reference guided methods, polished, and assessed for quality using multiple tools. The final assemblies were subjected to phylogenetic and comparative analyses as outlined in later sections. An overview of this pipeline is provided in **Figure 3.13**.

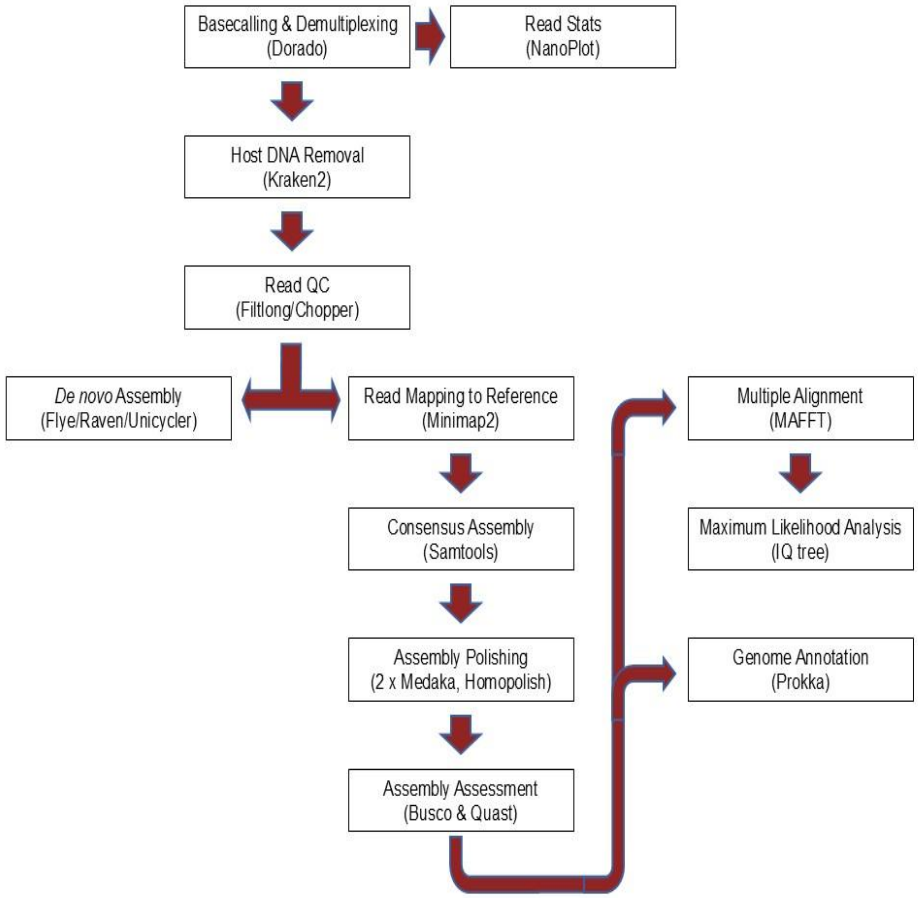


Figure 3.13. Bioinformatics workflow for long-read sequencing data processing and genome assembly.

Raw ONT sequencing reads from metagenomic clinical samples were processed through a custom workflow to enable genome assembly and downstream phylogenetic analysis. Using this pipeline, near-complete genomes were recovered through reference-guided assembly, whereas de novo assembly often produced multiple contigs that could not be fully resolved.

3.6.1. Basecalling and Read Quality

Basecalling is the process of converting raw electrical signals from the pores in the ONT flow cells into nucleotide sequences. The sequences are written as FastQ files that assign Q-scores to each nucleotide base. The Q-score (Phred quality score) is a logarithmic measure of per-base accuracy, where each increase of 10 units represents a ten-fold decrease in error probability (Q10 = 90% accuracy, Q20 = 99%). Early in this project, sequencing was performed using R9.4.1 flow cells, which use a single detection nanopore structure. These datasets had relatively modest read accuracy, with a mean Q-score of 11.62 as determined by Nanoplot (**Methods 2.6.1**), reflecting the expected performance of first-generation nanopore chemistry and models **(Figure 3.14).**

As the project progressed, R10.4.1 flow cells were adopted, offering significant improvements in read accuracy. These newer flow cells employ a dual-read nanopore and are optimized for use with ONT's updated super-accurate (SUP) basecalling models. Read accuracy improved as a result, with a significantly higher mean Q-score of 18.58 compared to the 11.62 with R9.4.1 ($p < 2.2 \times 10^{-16}$). This increase represents an almost 10-fold reduction in base error rates (from ~7.6% to ~1.4%), meaning more accurate consensus and downstream genome reconstruction.

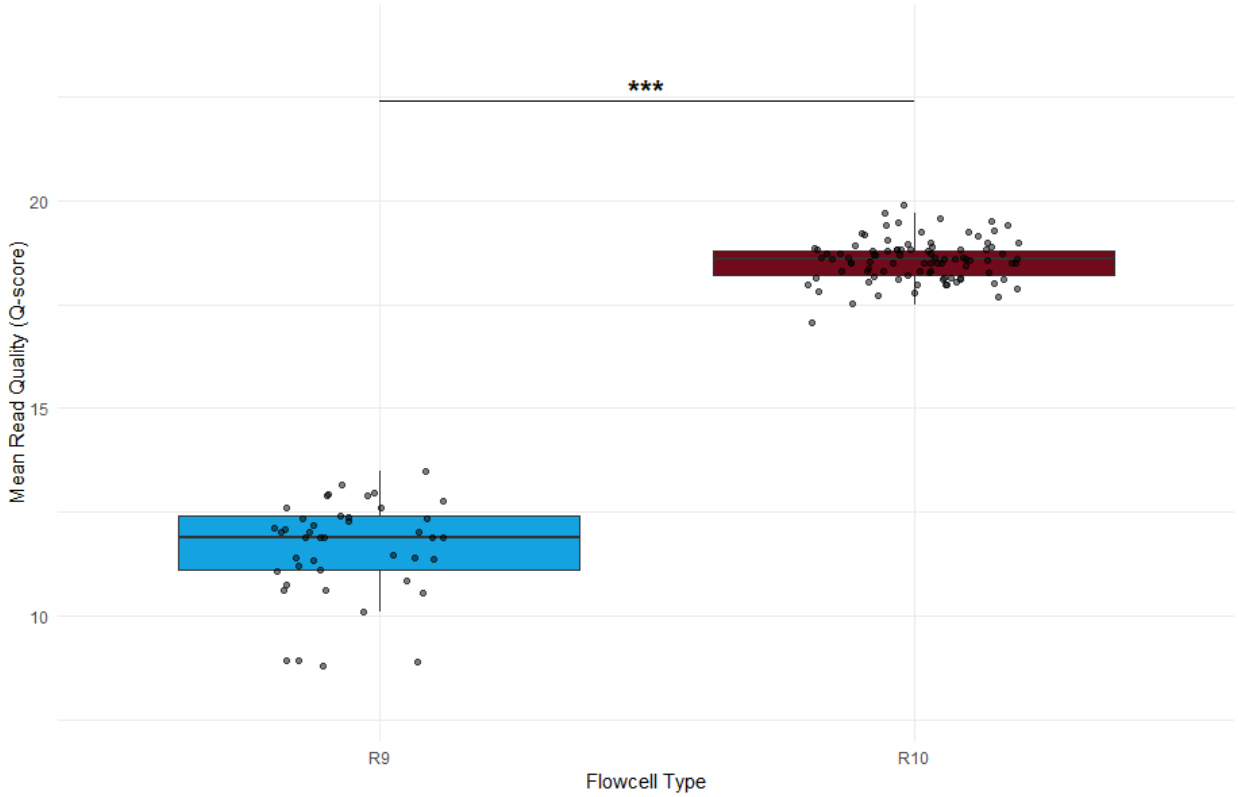


Figure 3.14. Comparing mean read quality (Q-scores) between R9 and R10 flow cells.

Boxplots show the distribution of mean read quality values per sequencing run, with individual points overlaid. R10 flow cells produced significantly higher Q scores than R9 ($p < 2.2 \times 10^{-16}$).

With basecalling complete, the first quality control step was taxonomic classification and filtering of raw reads using Kraken2. Even when both SWGA and AS were used to enrich for TPA, a substantial fraction of sequenced reads still originated from human or microbial background DNA. This further demonstrates the metagenomic complexity of clinical swab samples and the need for enrichment protocols.

Because sequencing runs were extended to maximize yield from low-input samples, the resulting FASTQ files were large (10–30 GB per sample). Kraken2 classification enabled early removal of non-target reads, which in some cases halved the file size, reducing data storage demands and accelerating downstream steps such as alignment and assembly.

Genome recovery outcomes are summarized in **Table 3.6**. Although a high proportion of TPA reads were not required for successful assembly, both read proportion and absolute TPA read count were associated with genome coverage and depth. For example, sample as50 contained only 17% TPA reads yet achieved 99.8% coverage at 176x depth, while as52 and as58 (29% and 26% TPA, respectively) produced similarly complete assemblies. In contrast, samples with <2% TPA reads (as42, as59, as60) were insufficient, producing assemblies with low mean depth (<49X) and incomplete coverage (<99%). These results suggest that while modest read proportions can still yield near-complete genomes, reliable recovery generally required at least 7% TPA reads or 100,000 classified reads.

Table 3.6. Kraken2 Classification and Genome Coverage for Nanopore-Sequenced Samples.

Sample ID	Unclassified Reads	TPA Classified Reads	Proportion of TPA reads	Genome Coverage (%)	Average Depth
as57	1330996	2135	0%	38.3	0.8
as33	1662046	5210	0%	75.1	4.2
as43	1578779	9335	1%	92.1	10.8
as60	1263038	12206	1%	94.5	15.5
as59	884260	9869	1%	92.5	10.9
as42	2105133	40002	2%	97.9	49.0
as35	1607964	67603	4%	97.4	93.0
as36	2009711	157038	7%	99.6	189.1
as45	1122558	198043	15%	99.7	201.6
as50	678666	142470	17%	99.8	175.6
as58	1193986	412331	26%	99.8	456.4
as51	613895	240376	28%	99.7	336.1
as52	657031	270957	29%	99.9	377.4
as41	1175675	685577	37%	99.9	714.4
as53	476607	492216	51%	100.0	725.8
as34	1249420	1750339	58%	99.4	1563.8
as61	802259	1272626	61%	99.9	1301.9
as37	687116	1791293	72%	100.0	2018.7
as49	343096	931639	73%	100.0	1014.8

3.6.2. De novo and Reference Guided Genome Assembly

De novo genome assembly was attempted using the twelve highest coverage TPA samples. The reference-free approach was prioritized to avoid alignment bias and to preserve potential structural variation present in local strains. However, because the clinical specimens required WGA prior to sequencing, downstream issues such as uneven genome coverage and amplification bias were introduced. As a result, sequencing runs produced variable read depths and often required extended runtimes to obtain sufficient data for assembly. The resulting large FASTQ datasets complicated computational assembly even after the removal of non-TPA reads.

Initial assemblies were performed with three long-read assemblers: Flye, Raven, and Unicycler **(Methods 2.6.2.1).** Raven and Unicycler failed to complete assembly on full datasets due to memory constraints, while Flye was able to process the full datasets but yielded highly fragmented assemblies with up to 255 contigs per sample. None of the Flye assemblies could be resolved into a single, closed contig. Due to uneven coverage introduced by WGA further, random subsampling of the reads disproportionately represented high-coverage regions and produced large gaps across the genome. This precluded the use of standard down sampling approaches that are typically applied to improve assembly efficiency.

To address these limitations, Pomoxis was used to perform targeted subsampling to fixed coverage depths (50x or 100x) based on alignment to the TPA SS14 reference genome. This allowed all three assemblers to complete successfully. Assembly quality with the

down-sampled datasets improved across all tools, with contig counts typically reduced to 10 - 40 per sample.

Figure 3.15 illustrates the number of unresolved contigs for each sample across assemblers and subsampling conditions. Wilcoxon rank-sum tests indicated that subsampling depth (50x vs 100x) did not significantly affect contig number within any assembler (Flye: $p = 0.214$; Raven: $p = 0.817$; Unicycler: $p = 0.077$). In contrast, assembler choice had a strong effect. The Wilcoxon rank-sum tests showed that Flye performed worse than both Raven ($p = 5.6 \times 10^{-5}$) and Unicycler ($p = 0.00033$) with significantly more contigs. Raven and Unicycler had similar performances ($p = 0.056$) but no complete genomes were generated.

Although subsampling improved *De novo* assembly performance, they remained incomplete and highly fragmented, and no genome could be resolved into a single circular contig. To overcome this an alternate assembly method was required. This involved reads being aligned to the TPA SS14 genome and consensus sequences generated for each sample **(Methods 2.6.2.2)**. This approach produced single-contig assemblies, even in low-coverage samples. This was made possible with contig gaps being filled with ambiguous bases (Ns), which accumulated in regions of insufficient read depth.

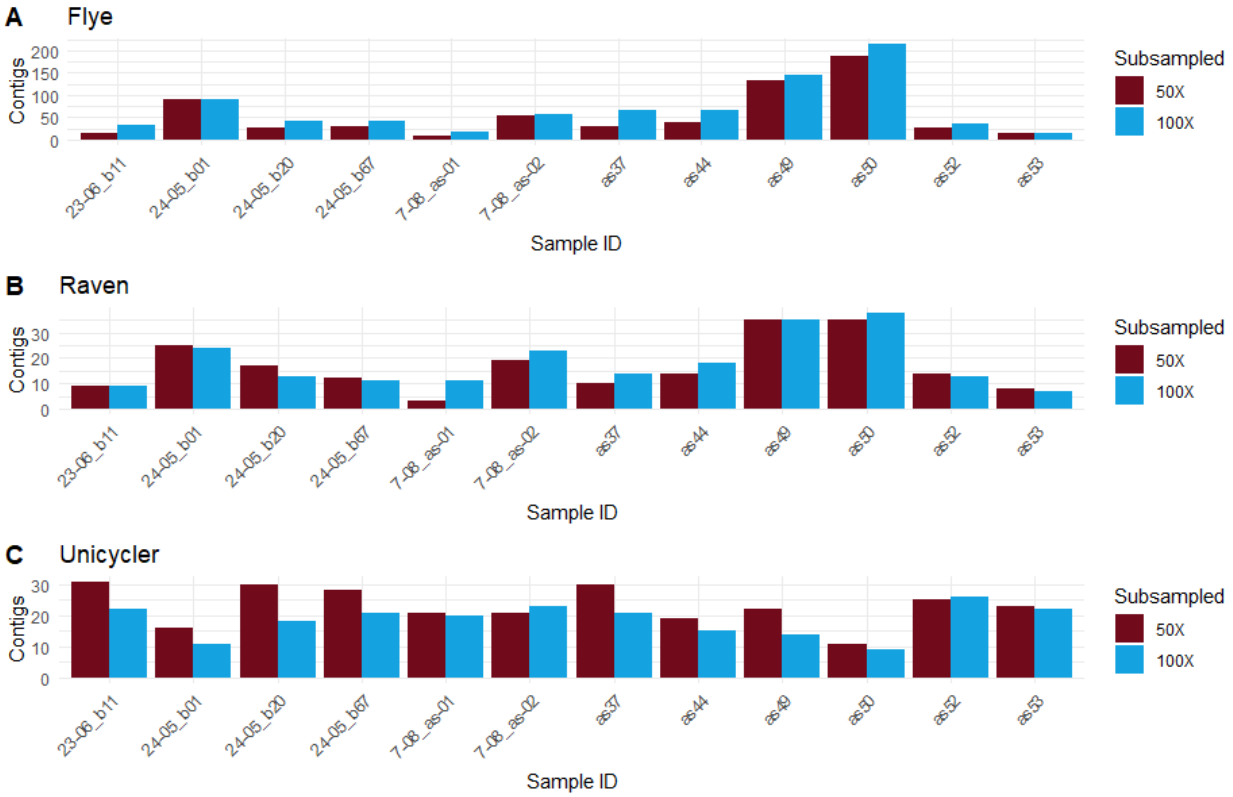


Figure 3.15. The number of contigs produced by each assembler using read depths subsampled to 50x or 100x.

Barplot showing contig counts for each sample (grouped by subsampling depth) within each assembler panel. Note the difference in y-axis scales, Flye is 0-200 contigs, while Raven and Unicycler are 0-30 contigs. Wilcoxon rank-sum tests indicated no significant effect of subsampling depth within assemblers (Flye: $p = 0.214$; Raven: $p = 0.817$; Unicycler: $p = 0.077$). Though, assembler choice significantly impacted the amount of fragmentation. Flye produced higher contig counts than Raven ($p = 5.6 \times 10^{-5}$) and Unicycler ($p = 0.00033$), while Raven and Unicycler did not differ significantly ($p = 0.056$).

3.6.3. Genome Polishing and Error Correction

Although newer ONT chemistries and basecalling models improved raw read accuracy, draft assemblies still contained errors. These include ambiguous bases (Ns), insertion/deletion (indel) and homopolymer-rich regions (stretches of identical bases such as “AAAAAA” or “GGGGGG”). These errors can be mitigated with post-assembly polishing tools, such as Medaka and Racon, that reanalyze the compiled reads and correct nucleotide calls in the assembly. Ns and indel counts were calculated with QUAST using the SS14 reference genome (NC_021508.1; **Methods 2.6.2.1**).

To evaluate polishing performance, four samples with high read depth and coverage were assessed for changes in ambiguous bases (Ns per 100 kbp). Unpolished assemblies contained between 16 and 187 Ns per 100 kbp. Polishing eliminated the ambiguity by determining the correct base, reducing Ns to zero across all methods tested (median reduction of 32.3 Ns per 100 kbp, corresponding to 100% removal) (see **Appendix 1.1** for per-sample values). This confirmed the necessity of polishing for improving base-level accuracy in consensus assemblies.

We next examined indels (**Figure 3.16**). Although one round of Racon (Racon x1) eliminated ambiguous bases, there was a consistent increase in the number of indels across all samples, rendering Racon-polished assemblies unsuitable for downstream analyses. Medaka performed more consistently: one or two rounds (Medaka x1 or x2) had equivalent reductions in indels without introducing additional errors, whereas a third round (Medaka x3) occasionally increased indel counts, so it was not retained. To further

address errors remaining after Medaka x2, adding Homopolish (Medaka x2 + Homopolish) yielded the lowest indel rates in every sample. Based on these findings, the final polishing strategy used for all assemblies comprised two rounds of Medaka followed by one round of Homopolish.

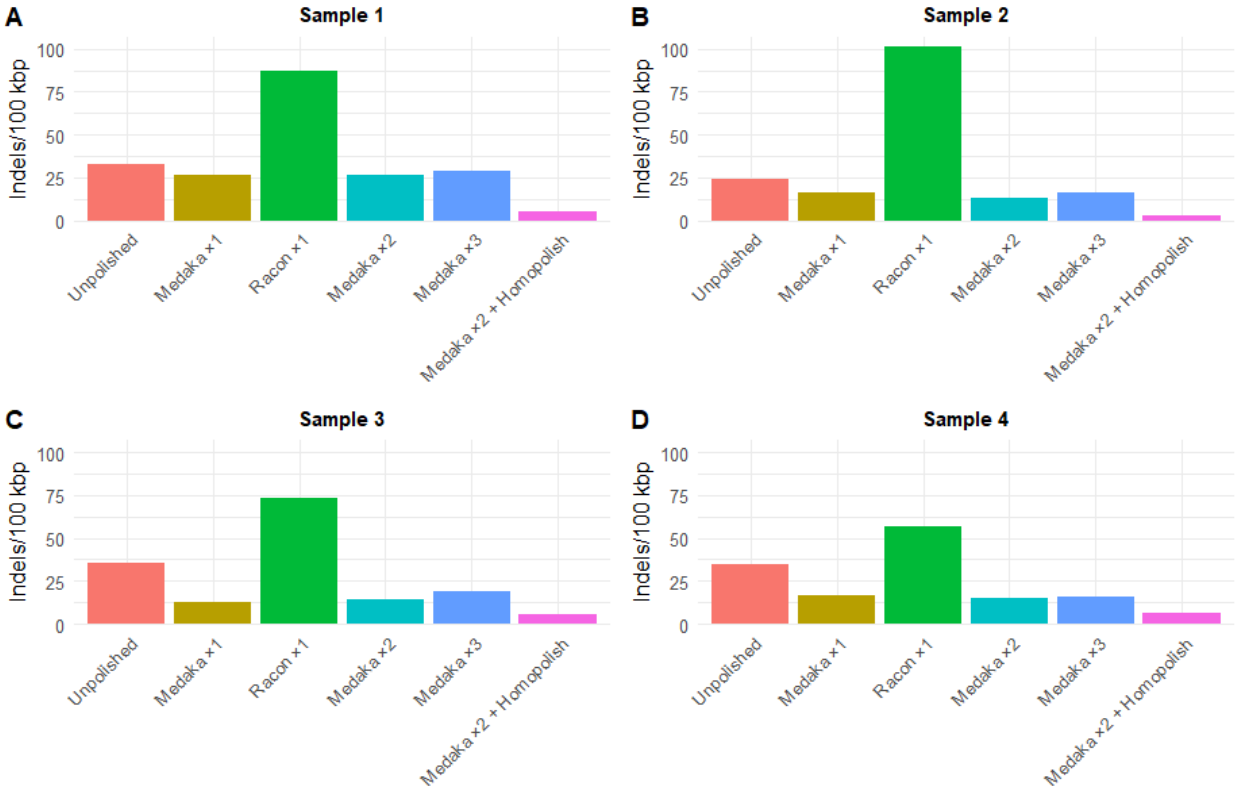


Figure 3.16. The effects assembly polishing on INDELS per 100 kbp.

Bar plots of indels per 100 kbp reported by QUAST relative to the SS14 reference (NC_021508.1). Panels A-D correspond to Samples 1-4. Pipelines shown: Unpolished consensus, Medaka x1, Racon x1, Medaka x2, Medaka x3, and Medaka x2 + Homopolish x1. Across all samples, Medaka x2 + Homopolish x1 produced the lowest indel rates; both Medaka x1 or x2 reduced indels relative to unpolished; Medaka x3 was inconsistent; and Racon x1 increased indels.

3.6.4. Genome Assembly Assessment

Polished assemblies were evaluated with two tools that assess different aspects of genome quality: BUSCO (Benchmarking Universal Single-Copy Orthologs) **(Methods 2.6.4)** and QUAST (Quality Assessment Tool for Genome Assemblies) **(Methods 2.6.2.1).** BUSCO identifies single copy orthologs, genes that occur once per genome in all members of a lineage, as indicators of assembly completeness and accuracy. The Spirochaetales dataset, containing 345 genes, was used as a reference for all TPA assemblies. However, due to the well-characterized genome reduction in TPA (38), all assemblies including the references, consistently lacked some of these genes. Specifically, the maximum number of detectable BUSCOs in complete TPA genomes is 318. Therefore, this value was used as the benchmark for assembly completeness in this study.

For quality assessment, only assemblies with at least 75% genome coverage were included, representing 116 of the 132 total assemblies obtained in this project. In this subset, BUSCO completeness showed a strong inverse correlation with the Ns per 100,000 bp ($r = –0.964$, $p < 0.001$). **Figure 3.17A** illustrates this relationship suggesting that lower sequence ambiguity is closely associated with gene-level completeness. Assemblies with fewer than 316 complete BUSCOs often showed elevated ambiguity, whereas those at or above this threshold consistently had fewer Ns and more reference-like sequence characteristics.

Further assessment using genome fraction (%), from QUAST, showed a strong positive correlation ($r = 0.964$, $p < 0.001$) to BUSCO completeness. **Figure 3.17B** illustrates that assemblies with greater genome coverage contain more BUSCO genes. Again, assemblies with at least 316 complete BUSCOs performed well, exceeding 96% of the genome fraction.

These results show that assemblies with high BUSCO completeness also have good sequence-level quality metrics, such as low ambiguity and broad genome coverage. This also supports using a threshold of ≥316 complete BUSCOs as a criterion for identifying high-quality assemblies in this dataset.

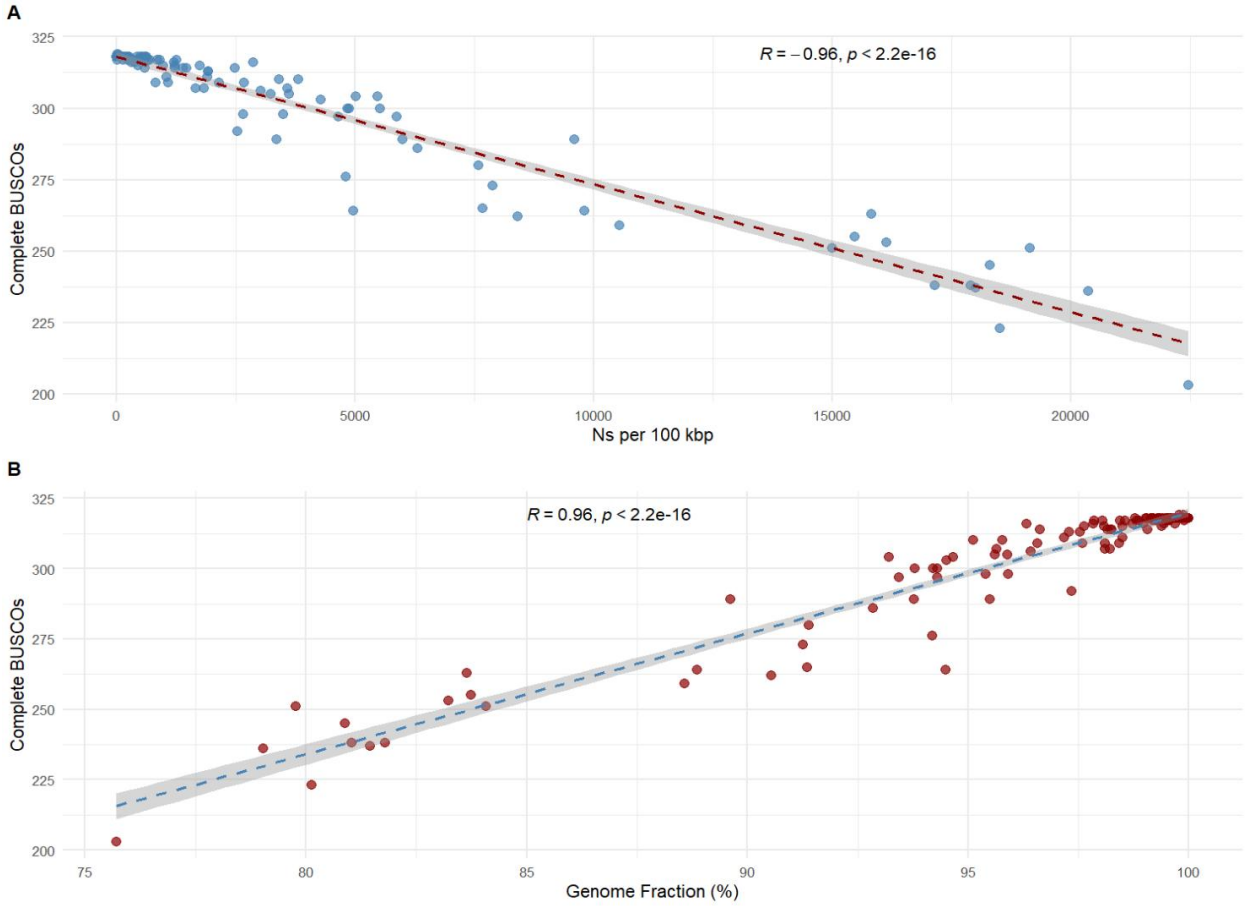


Figure 3.17. Relationship between BUSCO completeness and QUAST assembly metrics.

Only assemblies with ≥75% genome coverage (n = 116) were analyzed. (A) Complete BUSCOs plotted against the number of ambiguous bases (Ns per 100,000 bp), showing a strong inverse correlation (r = –0.964). (B) Complete BUSCOs plotted against the genome fraction (%), showing a strong positive correlation (r = 0.964).

With established assembly metrics, we sought to assess whether AS improved assembly quality relative to SS. The same 18 samples used to compare AS and SS read depth in **Results 3.5** were also evaluated for assembly completeness, **Figure 3.18**. Assemblies generated from AS data consistently showed higher BUSCO scores and lower ambiguous base content than their SS counterparts. Two samples that recovered fewer than 250 complete BUSCOs under SS improved to more than 300 when sequenced with AS, while another rose substantially from 203 to 280. On average, AS assemblies contained 311 complete BUSCOs compared to 282 for SS assemblies. Moreover, 11 AS genomes reached the TPA benchmark of 318 complete BUSCOs, compared with 8 SS genomes. These results provide additional support for the use of adaptive sampling to enhance assembly quality in metagenomic TPA sequencing.

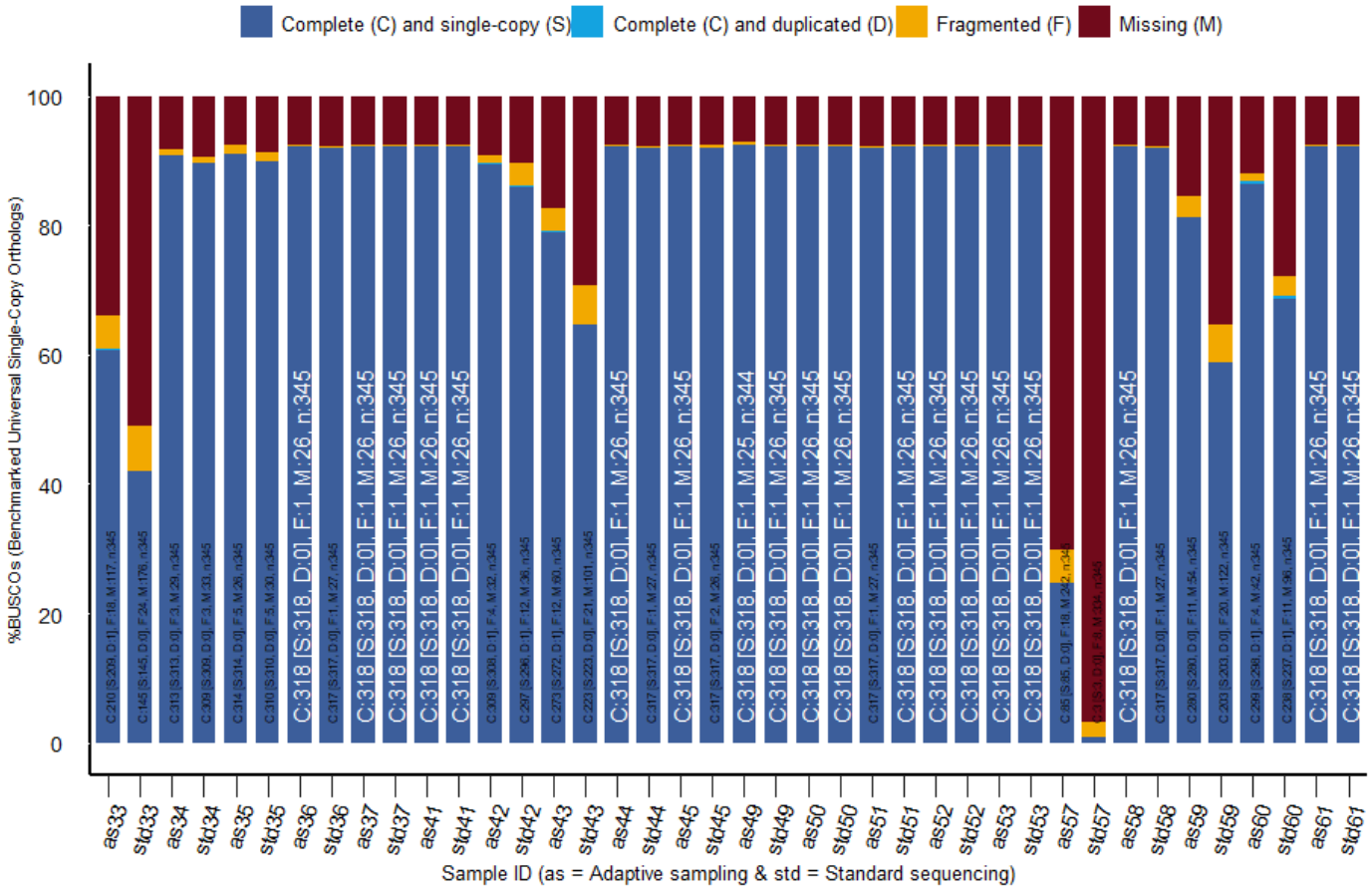


Figure 3.18. BUSCO completeness of paired assemblies generated with adaptive sampling (AS) and standard sequencing (SS).

Stacked bar plots show BUSCO categories; Complete and single copy (S), Complete and duplicated (D), Fragmented (F), and Missing (M) for 18 paired assemblies (x-axis alternates AS and SS for each sample ID). Assemblies generated with AS consistently recovered more complete BUSCOs than SS. 11 AS genomes attained the 318 complete BUSCOs compared with 8 SS genomes (annotated with white text).

3.7. Comparison of TPA Genomes from Manitoba with Global Data

3.7.1. Multiple Sequence Alignments

To compare ONT-derived TPA genomes with global reference strains, multiple sequence alignments (MSA) were generated using MAFFT. However, initial alignments were challenged by large artificial gaps. This distorted downstream phylogenetic analysis by exaggerating divergence among the ONT produced assemblies (**Figure 3.19**).

The issue stemmed from misalignment of the two ribosomal RNA (rRNA) operons, which are present as two identical copies in the TPA genome. MAFFT frequently aligned the first rRNA operon of one genome (located around 230,000 bp) with the second copy (located around 280,000 bp) from another, resulting in mismatched regions and artificial gaps approximately 50 Kbp in length, the genomic distance separating the two operons. As a result, regions of the MSA that should have exhibited near-complete identity instead showed inflated variation between genomes.

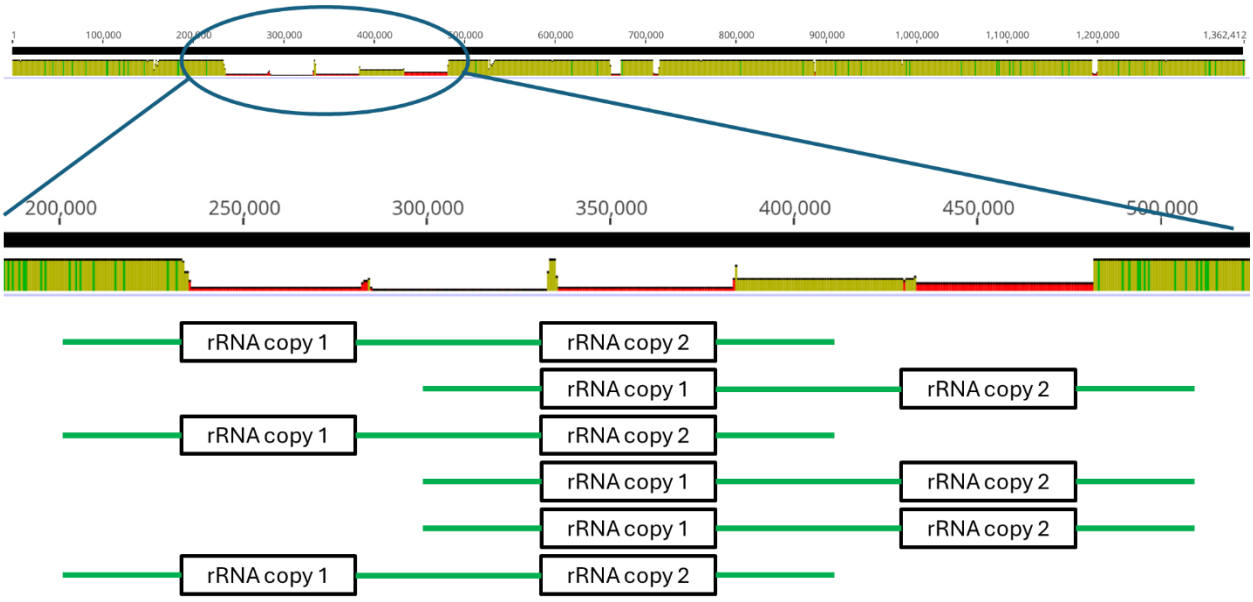


Figure 3.19. Misalignment of ribosomal RNA operons in MAFFT alignments.

Initial MSAs of TPA genomes exhibited large artificial gaps caused by the misalignment of the duplicate ribosomal RNA operons. rRNA copy 1 from one genome was frequently aligned to copy 2 from another, introducing ~50 Kbp shifts and disrupting overall sequence homology.

To address this, assemblies were processed as described in **Methods 2.7.1** to separate the duplicated rRNA operons into distinct segments. Each segment was aligned independently with MAFFT, then rejoined by sample ID to restore full-length genome alignments. This approach improved alignment accuracy and provided confidence in downstream phylogenetic reconstruction (**Figure 3.20**).

Even after correcting the rRNA operon misalignments, four distinct regions of high sequence variability remained in the MSA. The first was located between ~329,000 - 340,000 bp, a region that consistently exhibited poor SWGA amplification across all ONT-sequenced assemblies. This led to reduced sequencing depth, low basecalling confidence, and ultimately poor assembly quality. The other three variable loci: *arp* (~466,000 bp), *tp0470* (~510,000 bp), and *tprK* (~989,000 bp) are the only known regions in the TPA genome with large, variable repeat structures. In *arp* and *tp0470*, variation in the number and composition of tandem repeats generates extensive length heterogeneity across strains (73,74). Separately, *tprK* undergoes antigenic variation through non-reciprocal gene conversion with *tprD* among seven defined variable sites, providing a major source of diversity among strains (79). No other genomic regions have been consistently identified with comparable levels of size variation.

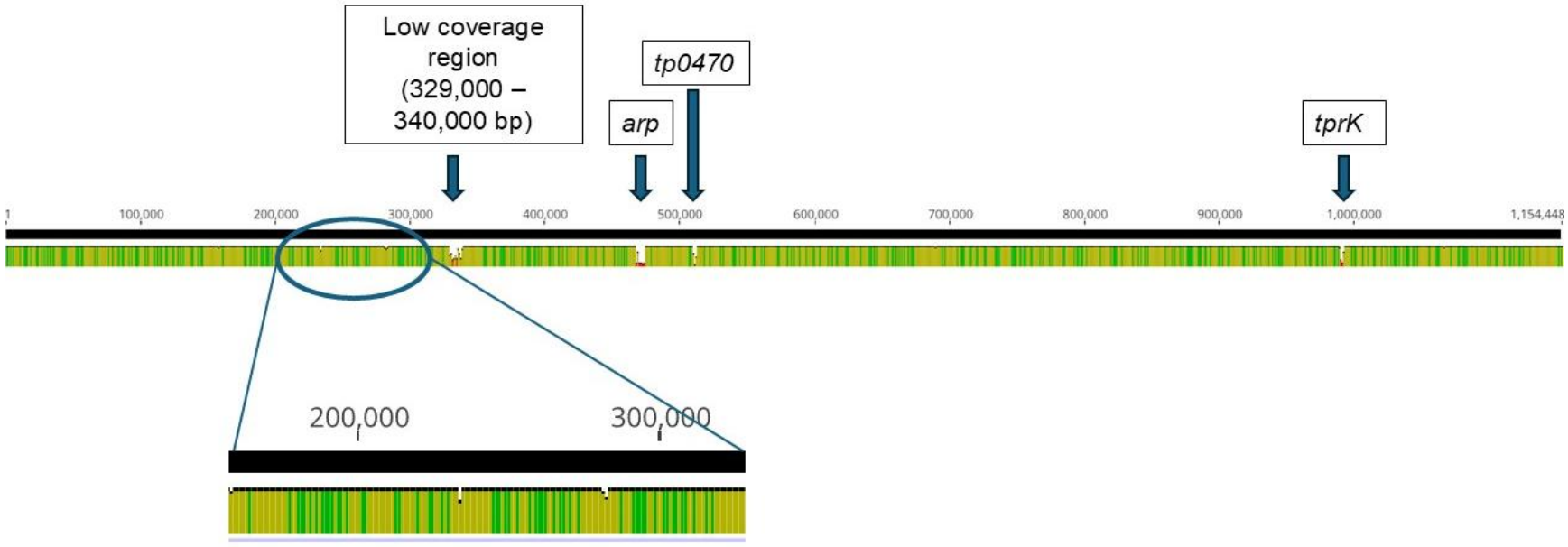


Figure 3.20. Correction of rRNA operon misalignments through genome segmentation and realignment.

After splitting genomes at the annotated 5′ end of the first rRNA operon, re-alignment successfully corrected previous MSA errors in the rRNA operons. The remaining four regions of high variability are due to low coverage or true genomic diversity (*arp, tp00470, tprK*).

3.7.2. Phylogenetic Analysis

A maximum likelihood phylogenetic tree was generated as described in **Methods 2.7.2** (see R script in **Supplementary Information 3**) using IQ-TREE, based on the unmasked MSA of TPA genomes. The analysis included 58 ONT assemblies with 316 or more complete BUSCO genes and 89 publicly available reference genomes. The references represent both Nichols-like and SS14-like lineages, with TPE (Yaws) included as an outgroup, see **Appendix 1.2** for metadata. The alignment preserved full genomic variability (i.e. no gene masking) including repeat and indel-rich regions, and the resulting tree is shown in **Figure 3.21**.

To improve interpretation heatmaps displaying sample metadata and QUAST assembly metrics were included. These include sample origin, indel rate, ambiguous bases (Ns), mismatch rate and genome fraction (%). While some ONT assemblies clustered appropriately among the known references, others displayed elongated branches often exceeding the length of the TPE outgroup. This is despite having passed the more than 316 BUSCO completeness threshold. The inflated branches suggested that genome completeness alone has insufficient resolution to identify error-prone assemblies. However, the QUAST metrics also failed to provide a cut point for quality. This meant further analysis was required for additional quality control measures.

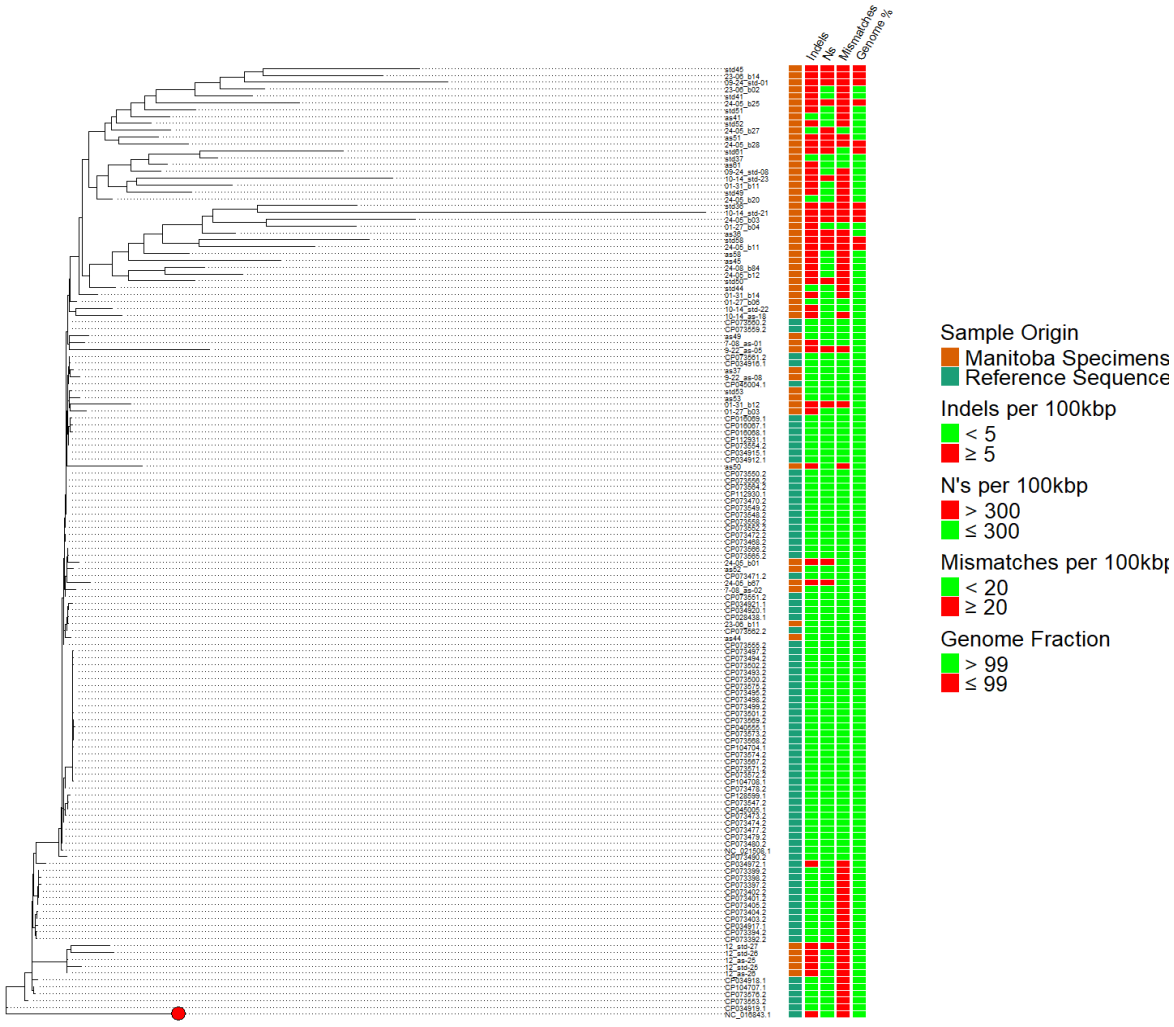


Figure 3.21. Maximum likelihood phylogenetic tree of TPA genomes passing the BUSCO completeness filter, annotated with assembly quality metrics.

The tree includes 58 ONT assemblies with ≥316 complete BUSCOs alongside 89 publicly available reference genomes. TPE is included as outgroup (red circle). Heatmaps display QUAST assembly metrics: indels per 100 kbp, ambiguous bases (Ns) per 100 kbp, mismatches per 100 kbp, and genome fraction (%). Several ONT assemblies display inflated branches, exceeding the yaws sample.

Although the indel rate per 100 kbp showed the closest indicator for branch lengths, several assemblies with low indel counts (<5 per 100kbp) still exhibited inflated divergence. For example, some samples with >5 indels per 100 kbp (such as 24-05_b01 and 24-05_b67) showed short terminal branches, while other assemblies with <5 indels (including as41 and 24-05_b27) had branches approaching the length of the TPE (yaws) outgroup (**Table 3.7**). Closer inspection of the MSA indicated unexpected variability was present in the rRNA operons, regions that are highly conserved among bacterial species (167).

To address this, BLAST searches were performed using the 23S rRNA genes from each assembly (obtained in section 2.7.2) against the 23s genes of the reference genomes in **Appendix 1.2**. Seventeen assemblies failed this QC step (<100% identity) and were excluded. Technical replicates were also removed (N = 19), leaving 22 unique assemblies. The remaining sequences were re-examined in the alignment, **Figure 3.22**, where several still demonstrated dense clusters of SNPs in the rRNA operons (N = 9). In contrast, all reference genomes showed 100% sequence identity in this region.

Table 3.7. Quast Assembly Quality Metrics and rRNA Operon SNP Density for Manitoba ONT Genomes.

Assemblies are separated by rRNA SNP density and then arranged by indels per 100kbp in ascending order. Samples with low SNP density correspond to shorter terminal branch lengths in phylogenetic analysis.

Sample ID	Indels per 100kbp	Ns per 100kbp	Mismatches per 100kbp	% Genome fraction	rRNA operon SNP density
23-06_b11	1.93	0.00	1.93	100.00	Low
01-27_b06	2.46	0.00	4.74	100.00	Low
7-08_as-02	2.72	29.07	5.18	99.87	Low
as37	3.16	0.00	2.28	99.94	Low
as52	3.34	110.38	6.16	99.75	Low
as49	3.52	15.02	7.12	99.79	Low
as53	3.95	0.00	4.13	99.87	Low
24-05_b01	**5.12**	**489.68**	**6.26**	**99.50**	**Low**
24-05_b67	**5.22**	**537.11**	**14.33**	**99.19**	**Low**
7-08_as-01	5.71	0.00	9.41	99.82	Low
12_as-26	5.97	0.00	37.40	99.97	Low
10-14_std-22	6.75	8.76	16.75	99.94	Low
as50	7.24	220.21	35.48	99.41	Low
24-05_b27	**3.63**	**629.91**	**17.26**	**99.15**	**High**
as41	**4.06**	**0.00**	**23.91**	**99.49**	**High**
24-05_b20	4.39	0.00	29.67	99.98	High
as61	5.37	57.80	15.33	99.65	High
24-05_b28	7.28	690.44	33.28	98.88	High
as51	7.31	313.07	27.30	99.53	High
as58	8.02	156.86	37.54	99.62	High
as36	8.13	424.63	40.13	99.29	High
as45	10.27	186.91	44.01	99.16	High

Figure 3.22. Alignment artifacts in the rRNA operon of aligned TPA assemblies.

A zoomed-in region of the MSA displays a portion of the rRNA operon across a subset of ONT assemblies, visualized in Geneious. Red arrows indicate assemblies with numerous SNPs. Each colored base represents a position that differs from the consensus (grey). All reference genomes showed 100% sequence identity in this region, indicating that the observed mismatches are sequencing artifacts.

After assemblies with high SNP density were removed, the MSA of Manitoba assemblies (N = 13) and 89 publicly available reference genomes were used to generate the final maximum likelihood tree (**Figure 3.23**). As described in **Results 3.7.2**, phylogenetic analysis was performed using IQ-TREE based on the unmasked alignment **(Methods 2.7.2 and Supplementary Information 4**)**.** The additional quality filtering improved the alignment and reduced artificial divergence in the final phylogenetic reconstruction. The resulting tree exhibited more realistic branch lengths and consistent clustering among known Nichols-like and SS14-like lineages.

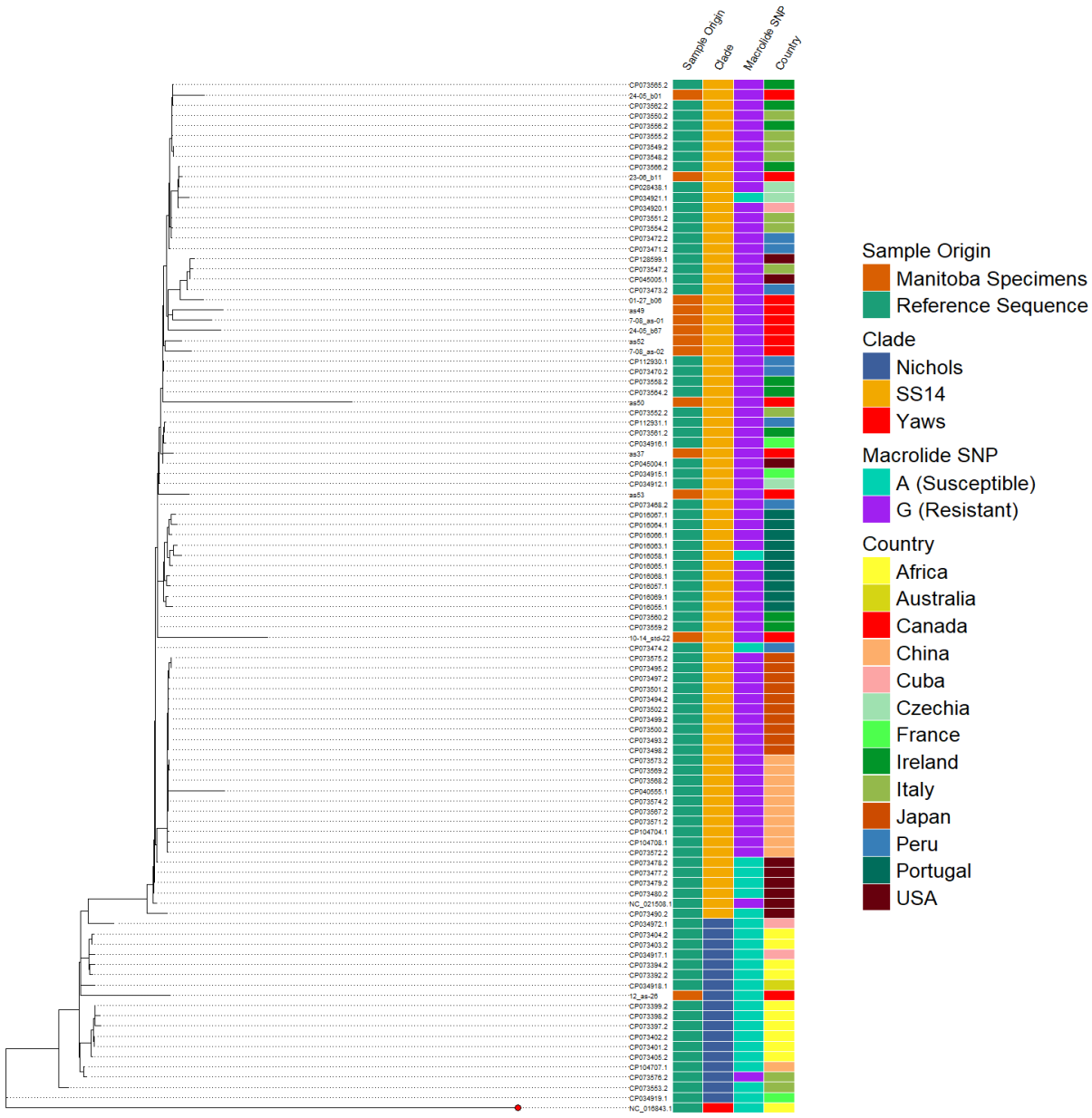


Figure 3.23. Maximum likelihood phylogenetic tree of TPA genomes with ONT assemblies and global references.

The tree includes 13 ONT assemblies and 89 reference genomes. Four heatmaps are displayed alongside the tree: sample origin, clade assignment, 23S rRNA macrolide resistance genotype, sample country of origin. The tree shows the placement of all Manitoba genomes within the SS14 lineage and the expected clustering of clades and 23s genotypes across global samples.

The maximum likelihood phylogenetic tree, based on high-quality ONT assemblies and reference genomes, showed that all Manitoba TPA samples clustered within the SS14 lineage. A single Nichols strain sample, provided by the National Microbiology Laboratory (NML), grouped appropriately with other Nichols references, validating its lineage assignment. The TPE reference remained a distinct outgroup with a longer branch length, and all remaining assemblies displayed shorter terminal branches than TPE, indicating improved sequence accuracy following manual curation and filtering.

Although overall tree topology improved after filtering, two Manitoba samples, 10-14_std-22 and as50, retained noticeably elongated branches compared to other ONT assemblies. These samples also exhibited the highest indel rates among the 13 best ONT assemblies (6.75 and 7.24 per 100 kbp, respectively; **Table 3.7**), suggesting that residual assembly errors persist outside the screened rRNA operon. Despite their longer branches, both assemblies clustered appropriately within the SS14 lineage and did not show artificial divergence beyond the YAWS subspecies observed in earlier tree versions.

Heatmap annotations alongside the tree confirmed that major lineages clustered appropriately by both clade and genotype. Most of the Nichols clade included genomes predicted to be macrolide-susceptible, shown in **Appendix 1.2**. The ONT-sequenced NML Nichols sample clustered among them, also containing the macrolide sensitive 2058A allele (**Table 3.8**). Contrasting that, the SS14 lineage primarily consisted of macrolide-resistant 2058G mutations. This genotype was determined through variant calling of the rRNA operon, as outlined in **Methods 2.7.2.** The thirteen ONT genomes

had the necessary coverage (>50x) for variant calling and the A2058G mutation was detected in each one. Interestingly, a minority of SS14 reference genomes were also macrolide-susceptible, demonstrating that resistance is common but not universal within this lineage.

Table 3.8. Macrolide Resistance (A2058G) Variant Calls in High-Quality ONT Assemblies.

Sample ID	Reference allele	Alternate allele	Read Depth
01-27_b06	A	G	222
10-14_std-22	A	G	77
12_as-26 *	NA	NA	N/A
23-06_b11	A	G	629
24-05_b01	A	G	1621
24-05_b67	A	G	314
7-08_as-01	A	G	293
7-08_as-02	A	G	174
as37	A	G	518
as49	A	G	224
as50	A	G	72
as52	A	G	105
as53	A	G	389

*Sample 12_as-26 is the Nichols strain, no variant allele was detected at that location

In addition to lineage-level separation, some geographic structuring was observed in **Figure 3.23**. SS14 genomes from China and Japan formed a tight cluster, as did samples from Czechia. A group of six Manitoba genomes also clustered together within a single subclade, supporting the presence of a regional epidemiological link. However, available patient metadata for these samples, including age, sex and regional health authority, showed no direct epidemiological connections between these cases. Given the limited diversity of TPA genomes and the slow rate of sequence evolution (103), this clustering may reflect broader regional circulation of a dominant strain rather than recent direct transmission among cases.

3.7.3. In-Silico analysis of genes of interest

A total of 132 TPA genome assemblies were generated throughout this project, with several samples sequenced in multiple independent runs. To assess the sequence stability of the *tp47* gene, a target used in the screening qPCR assay, assemblies were filtered to include only those with a minimum read depth of 50x across the entire gene. This filtering identified 61 assemblies suitable for downstream analysis.

From these 61 assemblies, the MSA of the *tp47* gene revealed near-complete sequence consensus. This high degree of conservation validates the selection of this gene as a diagnostic target, where specificity and stability are crucial for reliable PCR detection **(Figure 3.24).**

A single nucleotide polymorphism at position 124 (T—>G) was observed in the technical replicates of the NML-provided Nichols strain. The presence among all three replicates indicates it is likely a true variant and not a sequencing artifact. Most importantly, this SNP lies outside the PCR amplicon and would not interfere with assay performance. These results confirm that the *tp47* gene is a highly stable target for molecular screening of TPA and remains suitable for continued diagnostic use.

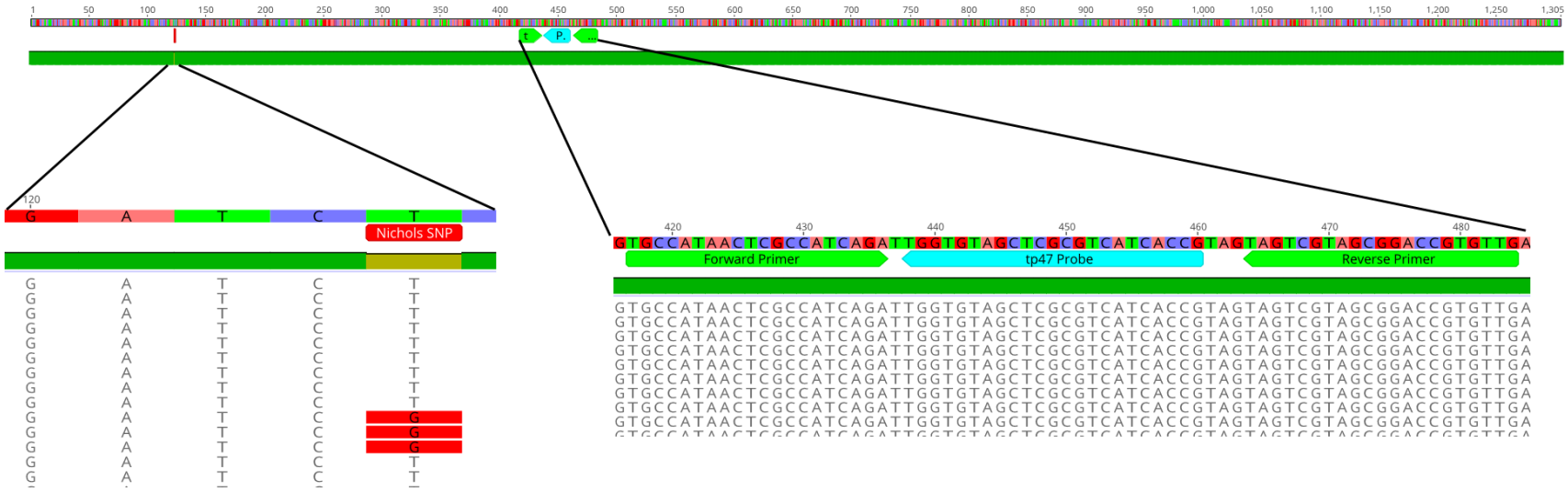


Figure 3.24. Multiple sequence alignment of the tp47 gene from high-depth TPA assemblies (N = 61).

The *tp47* gene was aligned across 61 ONT assemblies. Primer-binding regions for the qPCR assay are indicated by green arrows (forward and reverse primers), and the hydrolysis probe binding site is marked by a blue arrow. A single T→G SNP at position 124 was identified in all replicates of one Nichols strain sample, highlighted in red well outside of the PCR amplicon region (nucleotides 417–484).

The *arp* gene contains a variable number of 60-bp tandem repeats that have historically been used for strain differentiation (43,75,168). To investigate repeat copy number in the ONT dataset, *arp* sequences were analyzed from all 132 genome assemblies using the workflow described in **Methods 2.7.3**. Of these, 34 assemblies had a minimum mean sequencing depth of 5x across the entire *arp* gene and were included in the analysis. With Tandem Repeat Finder, the analysis identified a consistent pattern across all 34 samples, each *arp* gene contained 14 complete 60-bp tandem repeat units. No variation in repeat count or repeat unit length was observed in this subset. This uniform repeat structure suggests limited diversity in the *arp* locus among the sequenced samples in Manitoba.

Chapter 4: Discussion

4.1. Collection of Clinical Specimens

4.1.1. Advantages of Screening PCR

Research on TPA, the causative agent of syphilis, is significantly constrained by its fastidious biology, small genome size, and experimental challenges associated with culture isolation (38,41,119). As a result, identifying clinical samples with a higher likelihood of successful downstream amplification for whole-genome sequencing is critical. When this project was initiated, nested PCR was the most common approach for TPA detection (169). Although nested PCR offers excellent sensitivity, it is less amenable to precise quantification of bacterial load, which is important for prioritizing samples for downstream analysis.

The results of the Tp47 screening assay (**Results 3.1.1**) and subsequent implementations of the Lesion panel (**Results 3.1.2**) streamlined laboratory workflows and significantly improved CPL TATs compared to the traditional nested PCR approach. The qPCR assays enabled accurate quantification of TPA DNA, allowing us to identify samples with higher DNA concentrations for downstream MDA and long-read sequencing.

4.1.2. Human vs. TPA Genome Abundance

Most bacterial WGS using Nanopore technology is performed on isolates, where the total amount of DNA needed for library preparation is typically measured using Qubit fluorometry or Nanodrop spectrophotometry (131,170). The library preparation protocols have high DNA concentrations (>30 ng/µl) to ensure successful sequencing, and it is generally assumed that the majority of the DNA originates from the target bacterial

genome. However, in this study, the samples are metagenomic in nature because there is no culture-based method for isolating TPA. Consequently, although nucleic acid measurements may indicate sufficient DNA for library construction with our samples, these values predominantly reflect the host microbiome, not TPA.

Our PCR results consistently showed a 6–8 cycle difference between the human BGB control and the Tp47 target (**Results 3.1.2** and **Figure 3.3**). This translates to more than a 100-fold difference in DNA concentration, with human DNA vastly outnumbering T. pallidum DNA in the extracted samples. Given that the human genome is approximately 2800 times larger than that of T. pallidum (38,162), the result is an overwhelming excess of human DNA bases in each extraction. The predominance of human DNA effectively monopolizes the Nanopores during sequencing, wasting sequencing capacity and reducing the likelihood of generating sufficient TPA genome coverage. Ultimately, while total nucleic acid measurements (Qubit or Nanodrop) may indicate that there is sufficient DNA for library preparation, the proportion of TPA DNA is minimal. For effective Nanopore sequencing additional strategies to enrich or selectively amplify T. pallidum DNA are needed.

Beyond the immediate goal of developing WGS, the implementation of the Lesion Panel also improved routine diagnostic workflows at CPL. By pairing the TPA and HSV/VZV assays, the laboratory was able to better prioritize specimens, reduce reliance on clinical suspicion, and improve the surveillance of syphilis in Manitoba.

4.2. DNA Extraction Optimization for Long-Read Sequencing

4.2.1. Selective DNA Extractions

The performance of host depletion or microbiome enrichment strategies is necessary for successful long-read sequencing of TPA in metagenomic clinical samples, where human DNA overwhelmingly dominates total nucleic acid content. In this study two selective DNA extraction kits, the NEBNext Microbiome Enrichment Kit and the Qiagen Microbiome Kit, were evaluated alongside total nucleic acid extraction methods.

The NEBNext Microbiome DNA Enrichment Kit is designed to remove CpG-methylated host DNA post-extraction, enriching microbial DNA in metagenomic samples. However, in our study, the kit produced only a modest increase in BGB Ct values (2.6 cycles), indicating minimal human DNA depletion **(Results 3.2.1).** This was accompanied by relatively no improvement in Tp47 Ct values, suggesting that TPA DNA was not enriched to a level sufficient for downstream library preparation. These findings diverge from those of Thoendel et al., who observed 6- to 85-fold enrichment of bacterial DNA in sonicate fluids (125). In contrast, our results are consistent with those of Marotz et al., who reported no significant reduction in human-aligned reads in saliva samples treated with the NEBNext kit as it underperformed to other methods (171).

The Qiagen Microbiome Kit showed better host DNA depletion, achieving an average BGB Ct increase of 6.3 cycles **(Results 3.2.1** and **Table 3.2).** However, this came at the cost of an average increase of 5 cycles in Tp47 Ct values, indicating a parallel reduction in TPA DNA recovery. These findings suggest that the mild detergent-based lysis steps in the Qiagen kit, while effective at removing human DNA, likely also disrupt the fragile

outer membranes of T. pallidum. This is consistent with earlier observations that TPA outer membrane is highly sensitive to mechanical stress during centrifugation and non-ionic detergents (36,128). Such fragility results in TPA cells being lysed during the same steps intended to remove eukaryotic cells, preventing effective separation of bacterial DNA from host DNA. Neither the NEBNext nor Qiagen kits provided sufficient TPA recovery for successful long-read sequencing.

4.2.2. Total Nucleic Acid Extraction Methods

Without a suitable selective extraction method to enrich TPA DNA, total nucleic acid extractions were evaluated to maximize TPA DNA recovery while preserving fragment integrity for long-read sequencing (**Results 3.2.2**). The three methods tested, Qiagen DNeasy, bioMérieux eMAG, and KingFisher Flex, successfully extracted detectable TPA DNA from lesion swabs, as indicated by positive Tp47 qPCR results. However, the methods varied in DNA yield, fragment length, and suitability for downstream sequencing.

The eMAG extraction produced significantly lower Tp47 Ct values, providing the highest TPA DNA recovery by qPCR **(Figure 3.4).** Despite this, eMAG extracts had lower overall DNA concentrations and shorter fragment lengths than those of the KingFisher Flex **(Figure 3.5** and **Figure 3.6).** The Qiagen DNeasy protocol, though commonly used and easy to implement (96,99,107,172), underperformed in all 3 metrics. This may reflect mechanical shearing during column-based elution or reduced binding efficiency for high-molecular-weight DNA compared to magnetic bead-based methods.

The KingFisher extraction achieved a favorable balance of performance: it yielded high concentrations of DNA and consistently longer fragments across. Although the weaker t*p47* Ct values were considered significant in eMAG extracts, the averages were within 1 Ct (less than a 2-fold difference). Another consideration is that the purpose of the DNA extraction is for ONT sequencing, the ability to recover long DNA molecules remains an advantage. Furthermore, the KingFisher offers higher throughput (96-well format) with semi-automation, key benefits for scaling genomic surveillance efforts. Together, these findings support the use of the KingFisher Flex as the preferred extraction platform for TPA long-read sequencing.

4.3. Whole Genome Amplification Strategies for Low-Yield TPA Samples

Despite identifying the KingFisher as the preferred extraction method, The TPA DNA yields remained insufficient for long-read sequencing. To overcome this, whole genome amplification (WGA) was employed to increase TPA DNA quantity. While RNA-bait capture is popular for targeted enrichment, current protocols remain optimized for short-read platforms and short DNA fragments (~120 bp) (173), limiting their compatibility with Nanopore sequencing. Amplicon-based WGA approaches, such as those used for SARS-COV-2 (174), are similarly impractical for bacteria due to the complexity of designing and optimizing hundreds of primer sets for the larger genomes. In contrast, MDA offers a solution by amplifying total DNA using a strand-displacing DNA polymerase like phi29. However, the choice between random and sequence-specific priming has a major impact on genome recovery outcomes.

RpWGA, using the REPLI-g advanced kit, failed to deliver meaningful enrichment of TPA DNA. Though the reactions generated enough total DNA to proceed to sequencing, quantitative PCR showed minimal improvement in t*p47* abundance. This was further supported with the results from genome mapping **(Figure 3.7)** showing poor recovery (~12% breadth, 0.9x mean depth). This suggested that the non-specific amplification favored either host DNA or components of the broader microbiome, effectively outcompeting low-abundance TPA sequences in the mix. Unfortunately, host DNA content (as measured by BGB) was not assessed post-WGA, but the absence of TPA reads support a predominance of background amplification.

In contrast, SWGA, which uses primers computationally designed with a high affinity to TPA and minimal binding to host DNA, dramatically improved performance. Using the SWGA-Pal 12 primer set from Thurlow et al. (2022), TPA enrichment was observed in nearly all samples, with up to a 12-cycle Ct decrease in *Tp47* and copy number increases ranging from 10^3 to over 10^6 **(Figure 3.9)** (97). Unlike rpWGA, SWGA did not amplify host DNA, as shown by stable or increased BGB Ct values post-amplification **(Table 3.3).** This improved specificity provided sufficient TPA read recovery to achieve 100% genome breadth and >1,000X coverage in several samples, allowing for successful long-read assemblies directly from clinical material. These results support the original findings of Thurlow et al. (2022), confirming that SWGA can effectively enrich T. pallidum DNA even in low-input, high-background clinical swab extracts.

One notable exception was WGA-sample 22, which showed a marked decrease in amplification efficiency **(Figure 3.9).** This was likely due to DNA loss during post-amplification bead cleanup rather than a failure of the SWGA process itself. Ampure XP bead purifications can introduce variability and risk of sample loss if handling steps are suboptimal. Moreover, despite its advantages, SWGA was not without limitations. Coverage across the TPA genome was highly uneven, with some regions amplified to >30,000x while others had little to no supporting reads **(Figure 3.8).** This variability severely impacted de novo assembly: subsampling-based tools failed to retain low-coverage regions, and assemblers using full datasets produced highly fragmented contigs **(Results 3.6.2)**. As a result, reference-guided assembly had to be adopted as the method to produce genomes, further elaborated on in **Discussion 4.6.2**.

4.4. Development of a Long-Read Sequencing Protocol for TPA

Long-read sequencing of TPA from clinical specimens presents a unique set of challenges due to the low abundance of bacterial DNA, the fragile nature of the pathogen, and the overwhelming presence of host nucleic acids. Despite improvements in TPA DNA concentrations using SWGA, amplification bias was present. This meant that further protocol optimization was required throughout the sequencing library preparation workflow.

4.4.1. DNA Requirements and Sequencing Strategy

Incorporating the larger reaction volumes described in the ONT Whole-genome amplification protocol with the multiplexing method described in Native Barcoding

expansion protocol provided enough DNA to duplicate sequencing libraries. This enabled a flow cell reload strategy that substantially improved sequencing performance on the ONT platform. The first sequencing attempt was performed without a reload, pore activity dropped below 10% within the first 18–20 hours and total yield plateaued at 2.4 Gb, far below what was needed for assembly **(Figure 3.11).** In contrast, performing a flow cell wash and reload extended productive sequencing time to 40–50 hours and resulted in a 2- to 3-fold increase in total data output. The early decline in pore activity occurred despite loading the ONT recommended amount DNA (50 fmol), suggesting that sequencing depletion typically occurs within ~20 hours even under optimized conditions.

Another reason behind this drop in activity could be due to clogged or inactive pores. As pores become idle or obstructed with DNA fragments, overall throughput diminishes. Performing a flow cell wash removes buffer contaminants and clears the blocked pores. Reloading a duplicate library then replenishes available DNA, allowing pores to re-engage in sequencing. Notably, the restoration of pore activity after reload was 40–50% and sequencing again tapered off after 20–30 hours, suggesting that pore degradation or DNA exhaustion still limits the sequencing duration even with washing **(Figure 3.11)**. Despite this, the ability to double sequencing time and yield from a single flow cell represents a substantial efficiency gain.

Even with this improvement in data generation, rpWGA-enriched samples were still insufficient to recover full TPA genomes. Subsequent advances in amplification strategy (via SWGA) and multiplex optimization were ultimately necessary to enable successful

genome assembly **(Table 3.4).** However, the wash and reload strategy remained a key component of the workflow, as additional data generation continued to be important even under improved conditions.

4.4.2. Multiplexing samples with variable coverage

Our results demonstrate the trade-offs between cost efficiency and genome recovery when multiplexing metagenomic TPA samples on the ONT MinION platform. While the theoretical capacity of a MinION flow cell (~50 Gb) suggests that up to 96 TPA isolates could be sequenced at high depth in a single run, this assumes ideal conditions where the entire sequencing yield consists of target DNA (38,166). In reality, clinical samples contain substantial amounts of non-target human and microbial DNA, even after targeted enrichment.

This metagenomic background imposes a practical limit on the number of samples that can be multiplexed per run. Our data show that the total number of sequencing reads per flow cell remains relatively constant (~5 million reads), regardless of the number of barcoded samples **(Table 3.4)**. Therefore, increasing library size does not increase sequencing output but instead divides the available reads among more samples. This directly impacts the ability to achieve sufficient depth for high-quality genome assembly. Multiplexing 4-6 samples per flow cell consistently yielded the highest number of near-complete genomes. Conversely, as library size increases (N > 8) more samples failed to generate assemblies due to insufficient per-sample read depth. This highlights the

importance of matching a multiplexing strategy to both sequencing output and genome size.

These results are also consistent with the empirical throughput limitations of MinION sequencing. Although the theoretical output over a 72-hour run is up to 50 Gb, our observed yields were closer to 10 Gb per flow cell, aligning with reports from other ONT metagenomic studies (138,166). With an average read length of 2 kb, this corresponds to the 5 million reads per run shown in **Table 3.4**. These results suggest that the early decline in pore activity may reflect the complete consumption of loadable DNA within the first 18–20 hours of sequencing. In other words, most sequence-able material is processed early in the run, and pores become idle not because of technical failures, but due to exhaustion of available DNA. This is further supported by the successful recovery of sequencing activity following a library reload, emphasizing that throughput is governed by available DNA rather than run duration.

While increasing multiplexing may seem efficient from a cost-per-run standpoint, it can be counterproductive in low-yield metagenomic contexts where depth of coverage is critical. This is further exacerbated by the uneven amplification seen in SWGA (**Figure 3.8)** which requires extended sequencing to obtain reads for low coverage regions. Improvements in WGA that can provide more uniform coverage would allow for larger library sizes. Until then, low abundance specimens, such as TPA, require conservative multiplexing to increase the likelihood of complete genome recovery.

4.5. Standard vs. Adaptive Nanopore Sequencing

AS offers a novel approach to target enrichment during ONT sequencing. By exploiting ONT's real-time basecalling, the flow cell can reject DNA strands that do not align to the TPA SS14 reference genome (NC_021580.1). This prevents the pore from spending sequencing time on host or microbial DNA. In our experiments, sequencing runs typically generated 10 Gb of data per flow cell, as discussed in the previous section**.** Without enrichment, a larger proportion of this yield was comprised of non-target human DNA. AS allowed us to preserve pore availability for TPA DNA, thereby increasing the number of useful reads and improving read depth across all samples.

Increased read counts translated directly into improved genome coverage and sequencing depth (**Figure 3.12**). Improvements in read depth were substantial, even in high-quality samples with near-complete coverage (>99%). Median read depth increased 1.5 fold (1.13 to 2.74) across all samples and in some cases more than doubled. For example, Sample 61 rose from 474x with SS to 1301x with AS.

Out of the 20 samples, 19 achieved greater or equal coverage through AS, particularly in low-coverage specimens such as Sample 57 (20.4% to 38.3%). Sample 52 showed a negligible difference (99.9% vs. 99.8%) in coverage, as it had already achieved near complete genome recovery under both conditions. This result, paired with the samples that obtained marginally better coverage with AS, suggests that when sufficient TPA DNA is present any further enrichment yields diminishing returns. Conversely, in samples with very low DNA input, AS improved coverage but could not fully rescue them to the threshold required for whole-genome assembly. This demonstrates that there is a DNA

concentration limit where some target DNA is present, but not enough for confident genome recovery without AS enrichment.

Deeper sequencing not only increases the breadth of genome coverage but also improves basecalling accuracy and reduces the likelihood of errors in consensus assembly. The increased depth from AS improved downstream assembly metrics, as shown by enhanced BUSCO gene recovery **(Figure 3.18)** and has implications for further analysis such as variant calling were accuracy and uniform coverage. Our findings support previous studies, including those by Ong et al. (2021), which demonstrated that AS increased microbial read proportions without reducing throughput in host-dominated bovine vaginal samples (140).

However, AS is not without limitations. Reported by Martin et al. (2022), enrichment efficiency depends on the similarity between the target DNA and the reference used during adaptive sampling (139). If there is significant divergence between the sample and the reference genome, enrichment may fail to recognize variant sequences. In using AS there is a risk of introducing bias by enriching reads similar to the reference and discarding divergent regions, potentially masking diversity. Therefore, AS should be used with caution in studies exploring strain variation or discovering novel lineages.

Through these results, adaptive sampling provides a practical advantage in the long read metagenomic sequencing of TPA. It enhances sequencing efficiency and improves read depth, though the reliance on a reference genome may limit its use in high-diversity contexts.

4.6. Bioinformatics Analysis

4.6.1. Basecalling and Read Quality

In Fall 2022, ONT released the R10.4 flow cells, announcing R9.4.1 flow cells would only be available for SARS-CoV-2 surveillance. This shift to the updated chemistry ultimately proved advantageous as the R10 flow cells produced higher quality reads. Early sequencing runs with R9.4.1 flow cells produced reads with a mean Q-score of 11.62 **(Figure 3.14),** whereas the transition to R10.4.1 flow cells, coupled with ONT's updated SUP basecalling models, led to a significant improvement in read quality. The average Q-score increased to 18.58, corresponding to an approximate 10-fold reduction in per-base error rate from 7.6% to 1.4%. These comparisons are consistent with recent benchmarking data, which demonstrate that R10.4.1 flow cells, in combination with SUP basecalling, deliver significantly higher read accuracy than earlier chemistries (175). While still below the Q40 values typical of Illumina platforms, reaching Q20 with ONT sequencing represents a significant step forward for long-read applications.

Despite using both selective whole genome amplification (SWGA) and adaptive sampling (AS) for targeted enrichment, a portion of the sequenced data consisted of host and microbiome DNA. This reflects the biological complexity of clinical swabs and the limitations of current strategies. Using Kraken2 to classify raw reads and filter out non-TPA DNA, ensured that downstream analysis was focused solely on TPA reads. A beneficial byproduct of Kraken2 filtering was that it reduced the file sizes by up to 50%, improving both storage and computational efficiency.

From the Kraken2 filtering the results showed that successful genome recovery was a result of the absolute number of TPA reads rather than their relative abundance, further supporting the findings of **Results 3.5** and the decision to incorporate AS. Several samples with less than 30% TPA-classified reads still had high read counts (as50, as52, as58) and achieved near-complete assemblies (**Table 3.6).** While others (as57, as59) with very small read counts (<10,000) yielded only partial coverage (Table 3.6). These findings, together with **Discussion 4.5,** indicate that future enrichment efforts should prioritize maximizing absolute TPA read yield rather than relative enrichment alone.

4.6.2. Genome assembly

De novo assembly remains the gold standard for reconstructing microbial genomes without introducing reference bias and this was a goal for the genome assembly portion of this study. Efforts to identify a *de novo* pipeline were conducted with the twelve highest-coverage TPA samples. However, the use of SWGA, while necessary to generate sufficient input DNA, introduced several challenges. The data generated for each sample showed variable read depths across the genome. To obtain coverage in low depth regions, sequencing runs had to be extended leading to FASTQ files far larger than typically expected for a 1.14 Mb genome. This created computational bottlenecks and hindered the performance of assembly programs.

Initial attempts using Flye, Raven, and Unicycler were unsuccessful in producing complete assemblies. Raven and Unicycler failed outright on the full datasets due to memory constraints, and although Flye completed assembly, the results were highly

fragmented, with some samples producing over 250 contigs. Random down sampling was considered as a way to reduce file size, but this approach removed reads from already underrepresented regions, exacerbating the unevenness introduced by SWGA and reducing genome completeness further.

To address this, targeted down-sampling was implemented using the subsample_bam utility from Pomoxis **(Results 3.6.2).** This tool trims aligned reads to a defined maximum depth (50x or 100x), retaining low-coverage regions while removing the excess read depth that overwhelms computational memory. With the Pomoxis-normalized datasets, all three assemblers ran to completion. Assemblies generated from 50x or 100x datasets did not differ significantly **(Figure 3.15).** However, both Raven and Unicycler generated assemblies with significantly fewer contigs than Flye. Despite these improvements, none of the de novo assemblies produced a single contig or circularized genome. All methods remained partially fragmented, indicating that uneven depth still posed a barrier to complete reconstruction. Given these limitations, a reference-guided approach had to be adopted.

Using the TPA SS14 genome (NC_021508.1) as a scaffold, we aligned reads and generated consensus sequences for each sample. Unlike the de novo approach, this method required no subsampling and successfully produced full-length (1.14 Mb), single-contig genomes in every case. Even low-coverage samples yielded assemblies, although regions with poor read support contained high densities of misassembled regions or

strings of ambiguous bases (Ns)**.** This method permitted full genome recovery and enabled further analysis.

Despite the success of this assembly approach, limitations remain. The SS14 reference genome was generated with short-read sequencing. Structural variation, repetitive regions and novel insertions may be missed or misrepresented due to alignment bias. As such, while reference-guided reconstruction was practical and reliable given the constraints of our dataset, it may obscure biologically relevant variation among circulating strains in Manitoba.

De novo assembly using long-read Nanopore data from clinical specimens is feasible but remains challenging due to sample quality and SWGA-related coverage variability. Targeted normalization with Pomoxis improved performance but was ultimately unable to overcome the highly variable depth. Reference-guided assembly proved to be a more robust and scalable strategy within this assembly pipeline, providing high-quality genome reconstructions suitable for phylogenetic analysis.

4.6.3. Genome Polishing

Although the base accuracy in ONT sequencing is improving **(Figure 3.14),** it still lags behind Illumina at the single-base level, particularly in homopolymer-rich regions (176). This makes genome polishing an important step for correcting base-level errors and reducing indels that may impact downstream analyses.

In this study, we compared polishing tools to improve reference-guided TPA assemblies **(Figure 3.16).** While Racon is commonly used in Illumina or hybrid assembly pipelines, it

has also been applied to long-read data (151). In our analysis, Racon effectively removed ambiguous bases but introduced excessive indels and was therefore excluded from the final workflow. This observation is consistent with other reports highlighting Racon's limitations in ONT-only datasets (176)

Medaka demonstrated more consistent performance. As an ONT-supported tool, Medaka benefits from models trained on specific ONT flow cell chemistries, which enhances its ability to correct typical long-read sequencing errors. One or two rounds of Medaka polishing reduced both ambiguous base content and indel rates across all test samples. A third round did not improve results and occasionally reintroduced indel errors, so it was removed from the workflow.

To further reduce the indel burden, particularly in homopolymer regions, a final step using Homopolish was implemented. When applied to assemblies previously polished with Medaka, this combination yielded the lowest overall indel rates.

Although effective, it is worth noting that Homopolish relies on comparisons to bacterial reference genomes. In cases where the target genome diverges from reference models, it may introduce incorrect corrections (176). Despite this limitation, the final polishing workflow, two rounds of Medaka followed by one round of Homopolish, proved to be a reliable and effective strategy for improving ONT-only TPA assemblies.

4.6.4. Assembly assessment

BUSCO evaluates the presence of conserved genes across a given taxonomic group, for this project it was the Order Spirochaetales. Interestingly, even complete TPA reference

genomes consistently contain only 318 of the 345 expected BUSCO genes. This reflects the organism's genome reduction and supports the use of 318 as a practical maximum for completeness (38). The ONT assemblies mirrored this pattern, with high-quality genomes attaining no more than 318 complete BUSCOs.

When BUSCO completeness was plotted against either the number of ambiguous bases or genome fraction, strong correlations were observed **(Figure 3.17).** Assemblies with higher BUSCO scores had fewer ambiguous bases and covered a larger percentage of the genome. These results are expected as higher-quality assemblies should be less error-prone and have higher gene identity.

These relationships also explain why assemblies have inflated branch lengths in our phylogenetic analysis, sometimes extending beyond what is seen in the TPE (Yaws) outgroup **(Figure 3.21).** Branch elongation is unlikely to reflect true biological divergence but rather due to higher accumulations of assembly errors. This interpretation is consistent with prior work showing that *T. pallidum* genomes are highly clonal with limited genomic diversity (96,103,177).

With a clearer understanding of assembly quality, we next compared AS and SS assemblies. Using BUSCO, we found that AS produced more high-quality genomes with a completeness of 318 **(Figure 3.18).** This trend aligns with the depth of coverage results described in **Results 3.5**, where AS yielded higher median read depth and broader coverage. Greater read depth likely contributed to improved consensus accuracy, reducing ambiguity and enabling more genes to be confidently called by BUSCO. These

findings provide additional support for incorporating AS into sequencing protocols for TPA, particularly in metagenomic contexts where DNA quantity and purity are limiting.

4.7. Phylogenetic Analysis

4.7.1. Multiple Sequence Alignments

For clonal organisms such as TPA, phylogenetic reconstruction depends on accurate alignments, since subtle differences among sequences can provide a signal of diversity. The TPA genome is highly conserved but duplicated operons and variable repeat regions can introduce technical challenges during alignment (103). In this study, MAFFT-based alignment of complete genomes introduced large artificial gaps stemming from misalignment of the two rRNA operons **(Figure 3.19).** These operons are identical in sequence but separated by ~50 kbp on the chromosome. MAFFT frequently misaligned rRNA copy 1 from one genome to copy 2 on another. As a result, regions that should have 100% sequence identity displayed artificial gaps. This creates distorted tree topologies and inflated branch lengths.

To address this, each genome was split immediately after the first rRNA operon based on Prokka annotated coordinates, allowing homologous sections to be aligned separately. The aligned segments were then merged to reconstruct full-length genomes. This approach resolved the operon mismatch, eliminating the non-biological gaps and restoring alignment accuracy **(Figure 3.20).** Specifically, the alignment showed high sequence similarity in the rRNA regions between 230,000 to 280,000 bp.

After correcting for the rRNA misalignment, four distinct regions of high variability remained in the MSA **(Figure 3.20).** The first spanned 329,000 to 340,000 bp, which included *tprE* and *tprG.* In all ONT assemblies, this region consistently had lowest read depth, which reduced basecalling accuracy and consensus reliability. The apparent variation in the MSA is therefore more likely a technical limitation of the SWGA rather than true biological diversity.

The remaining three regions reflect genuine biological variation. The *arp* gene (466,000 bp) contains 60-bp tandem repeats that vary in copy number between strains. This variation produces substantial length heterogeneity, making *arp* a useful molecular subtyping marker. However, the repeats can also pose a challenge for genome assemblies, particularly with short-reads (76). Similarly, the *tp0470* gene (510,000 bp) contains 24-bp repeat units, with repeat counts ranging from 4 to 29 in TPA and even more in TPE. There is clade-associated variation in the repeat number as well, which makes *tp0470* relevant for molecular typing (74). Finally, the *tprK* gene (989,000 bp) is a well-known site of antigenic variation and undergoes extensive sequence diversification through recombination and insertion events. These three loci: *arp*, *tp0470*, and *tprK* exhibited alignment gaps due to true biological diversity, and as such required no correction.

These observations emphasize the importance of distinguishing between alignment artifacts and authentic sequence polymorphisms in whole-genome comparisons. While regions like the rRNA operons required correction, biologically variable loci should be

preserved in an alignment to reflect a true signal. Although some comparative genomic studies choose to mask repetitive regions, doing so may obscure meaningful diversity, particularly in genes under selection or associated with antigenic variation (97,103,124).

Long-read sequencing offers improved resolution with these challenging regions. Previous work has shown that short-read datasets often fail to fully resolve repeat loci like *arp*, resulting in incomplete assemblies or requiring further Sanger sequencing (76). In contrast, Lieberman et al. (2022) successfully utilized ONT sequencing to accurately resolve repeat architecture in the *arp* and *tp0470* genes. Though not whole genome sequencing, this demonstrated the value of long-read technologies in characterizing complex genomic features (74).

Collectively, the alignment strategy and repeat-aware approach ensured that both conserved and variable regions were properly represented leading to accurate phylogenetic reconstruction.

4.7.2. Phylogenetic analysis

The results in **Section 3.6.4** supported the assembly threshold of requiring a minimum of 316 complete BUSCOs. However, the initial tree had substantial variability in branch length among ONT-derived samples **(Figure 3.21)**. Specifically, several assemblies exhibited terminal branches longer than that of the TPE (yaws) outgroup, despite being derived from clinical samples with laboratory and epidemiological profiles consistent with TPA. This suggests that the exaggerated divergence is a result of assembly artifacts,

rather than actual biological differences. This also indicates that using BUSCO completeness alone has limitations as a quality threshold.

To better understand these artifacts, QUAST assembly quality metrics were included in the tree as heatmaps. A threshold of less than 5 indels per 100kbp appeared useful for flagging higher quality genomes. However, this metric was not perfectly predictive. Some high-indel assemblies showed reasonable branch lengths, and conversely, some low-indel genomes had terminal branches approaching the TPE **(Table 3.7).** Other metrics such as ambiguous bases and mismatch rate similarly failed to clearly separate high from low quality genomes. Because QUAST metrics were also insufficient for identifying the highest-quality assemblies, further review of the MSA was necessary.

During the review of the alignment, some sequence variation was noted in the rRNA operons and they were confirmed using BLAST to find assemblies with less than 100% sequence identity *23S* genes **(Results 3.7.2).** rRNA regions are highly conserved in bacteria, so the presence of SNPs in the alignment is likely due to assembly artifacts and not novel genomic variation (167). The remaining 23 ONT assemblies were re-aligned with the reference genomes but variation in the operon remained **(Figure 3.22)**. Closer inspection showed regions of high SNP density were present in 10 assemblies. The previous BLAST search was only performed on the *23S* gene, explaining why those assemblies had not been filtered out and suggests that future work should include the whole rRNA operon.

Focusing on the 23s region did allow for the assessment of the macrolide resistance-associated A2058G mutation. Among the top 13 ONT assemblies, all had sufficient read depth required for variant calling and, with the exception of the NML Nichols strain (2058A), the macrolide resistant 2058G allele was present in all Manitoba assemblies **(Table 3.7).**

With the exclusion of low-quality assemblies, the remaining 13 ONT derived genomes were used for the final phylogenetic analysis (**Figure 3.23).** This subset represents the best assemblies generated during this project and it is worth noting that these span the entirety of the project. Although the majority of the 13 assemblies were sequenced with AS and R10 flow cells three originated from alternate workflows. These combinations were; AS with R9 flow cells (07-08_as-01 and 07-08_as-02), SS with R10 flow cells (10-14_std-22) and AS with R10 flow cells. Based on the results of this project, the reasoning for this is likely that these 3 samples had higher read depth obtained during those experiments as compared to their replicates, but analysis of the replicates has not been performed.

In **Figure 3.23,** the ONT assembled genomes represent the highest quality, and their placement in the phylogenetic tree is consistent with this interpretation. The terminal branch lengths are similar to the reference sequences, in contrast to the branches seen in the lower-quality assemblies of **Figure 3.21**. While branch length is not a formal quality metric, it supports the conclusion that these ONT genomes are biologically accurate.

In the phylogenetic analysis, limited but detectable regional clustering was observed among Manitoba samples. Though many genomes were interspersed through the broader SS-14 lineage, a small six-sample Manitoba subclade was identified, indicating a regional epidemiological link. However, available patient metadata did not reveal any direct epidemiological links between these cases. This pattern may reflect either multiple introductions of genetically similar strains or ongoing circulation of a well-mixed local population. It is also possible that this clustering is a result of the bioinformatics pipeline used, being biased through mapping against the SS-14 reference genome, which is known to introduce assembly bias (178). Importantly, all Manitoba ONT sequenced genomes fell within the SS-14 clade. This observation is consistent with a previous study finding the majority of British Columbia and Alberta also clustered within SS14 (103). However, given the small number of Manitoba genomes analyzed in this current project, conclusions about the dominant strain circulating in Manitoba cannot yet be made.

4.7.3. *In-silico* gene analysis

Although the majority of sequenced clinical samples did not yield high-quality, complete genome assemblies, the data generated still provides substantial value. In-silico analysis of individual genes remains feasible provided sufficient local sequencing depth is achieved. This demonstrates a key strength of WGS, where even if full genome reconstruction is not possible, targeted analysis of specific loci remains informative and actionable.

The *tp47* gene, used in our screening qPCR assay, serves as a prime example. Among the 132 assemblies generated, 61 had a minimum of 50x read depth across the gene, enabling reliable sequence comparison. Reviewing the MSA confirmed that the gene is highly conserved across Manitoba TPA genomes, with no variation observed in the primer or probe binding regions **(Figure 3.24)**. This supports its use as a diagnostic target and provides confidence in the assay. The only SNP detected, a T->G substitution at position 124, was confined to multiple replicates of the NML Nichols strain. This SNP was upstream of the PCR amplicon, indicating it will not impact diagnostic performance.

These results support the continued use of *tp47* as a stable diagnostic marker for TPA. It also illustrates how even sub-optimal genomic data has value. As sequencing technologies become more deeply integrated into clinical workflows, routine *in-silico* monitoring of diagnostic targets may serve as a valuable complement to traditional wet-lab validation.

Similarly, analysis of the *arp* gene revealed a uniform repeat copy number of 14 tandem 60-bp units across all 34 samples with sufficient sequencing depth (≥5x). Several additional samples with lower *arp* depth (3-4x) showed the same result when analyzed with Tandem Repeat Finder, suggesting that this repeat profile may be accurate in those cases as well, despite falling below our depth threshold of 5x.

This finding is consistent with earlier surveillance data from Manitoba. Shuel et al. (2018) analyzed TPA-positive clinical specimens collected between 2012 and 2016 and found that most typeable samples (N = 55) carried 14 *arp* repeats. They also identified two

samples showing 9 repeats and 17 were indeterminant (179). Although the ONT data is limited, the agreement between studies suggests that the 14 repeat *arp* gene is common in the region. While a single repeat pattern for *arp* provides limited resolution with-in Manitoba, its ability to detect divergent or imported profiles still offers valuable epidemiological insight.

4.8. Future Directions

4.8.1. Improving whole genome amplifications

Improving WGA represents the next logical objective for this project. Although SWGA significantly increased TPA DNA, the amplification varied dramatically across the genome, creating challenges from library preparation to genome assembly. One particular region, from 330,000-340,000 bp, had low coverage across all assemblies. Further primer optimization, including adjusting primer ratios or designing supplemental primers for underrepresented regions, could provide more uniform coverage. The goal would be to reduce the amount of sequencing time, allow for higher multiplexing in sample libraries, and permit *de novo* sample assembly possible for clinical specimens.

4.8.2. Finding the true DNA limit of flow cells

Another area for improvement is finding the true upper limit of DNA for loading flow cells. In this study, the maximum DNA concentrations recommended by ONT were followed through each step of the library preparation and flow cell loading. However, post-barcoding yielded sufficient DNA to allow for a duplicate library and flow cell re-load. Taken together with the results showing sequencing activity dropped below 10% within 20 hours raises the question of whether library splitting could be avoided by loading

significantly more DNA than ONT suggests. If pore clogging does not become a limiting factor, this would simplify the sequencing workflow by removing the steps of washing and reloading. Future protocols should test whether “over-loading” the flow cell provides sequencing advantages for sequencing performance and wet lab time.

4.8.3. Producing high quality reference genome

There has yet to be an established reference genome for TPA. Both NC_021508.1, the SS-14 strain, and CP004010.1, the Nichols strain, have been the more commonly used reference sequences but establishing a consensus reference would be beneficial for genome assembly and comparisons. A recent publication reported two Canadian TPA genomes, one from Manitoba and one from Saskatchewan (180). The Saskatchewan genome was assembled with a hybrid approach using ONT long reads and Illumina short reads, while the Manitoba genome was based on short reads alone. Building on that study, the workflow developed for this project demonstrates that long-read sequencing is feasible, and a hybrid assembly incorporating short-read sequencing could produce an in-house high-quality reference genome.

4.8.4. Improving Phylogenetic analysis

Future work should also focus on improving the precision of phylogenetic analysis. One approach would be to incorporate root-to-tip analysis with QC metrics to better identify indicators of quality assemblies. In this project, BUSCO and QUAST were primarily used for assembly assessment, but there were no definitive indicators found requiring the need for manual reviews of MSAs. Using the root-to-tip results provides an objective metric

when comparing to the outgroup (YAWS), as opposed to stating branches appear shorter or longer. Identifying correlations between root-to-tip results and one or more assembly metrics could improve assembly assessment and reduce reliance on manual inspections.

4.8.5. Expanding regions of interest analysis

During this project 132 samples were sequenced however only 13 proved good enough for phylogenetic analysis. Despite that, sufficient data was produced to analyze relevant genes such as *tp47* and *arp* in more than just the top assemblies. This could be expanded to other regions, provided appropriate read depth was present, including MLST genes (*tp0470* and *tp0705*), vaccine targets (*tp0751*) and the *tpr* family (*tprK, tprD,* etc.).

4.9. Conclusion

This project established a workflow for sequencing *Treponema pallidum* subspecies *pallidum* from clinical swabs in Manitoba. The initial *tp47* screening PCR and subsequent implementation of the Lesion panel assay enabled identification of specimens with higher bacterial loads and made improvements to the province's diagnostic procedures. With a sufficient amount of positive material, extraction methods were tested for downstream sequencing suitability. Selective extractions for microbiome enrichment or host depletion failed to capture enough TPA DNA, leaving total nucleic acid extraction as the more reliable choice. Among the available platforms, the Kingfisher offered the best balance of yield, fragment size, and throughput. However, the amount of TPA DNA recovered was insufficient for ONT library prep requiring the need for whole genome amplification.

Random primed WGA was ineffective, but SWGA significantly enriched TPA DNA, allowing for successful ONT library preparation.

Sequencing strategies incorporated aspects of multiple protocols focusing on larger volumes, sample multiplexing, and flow cell re-loading. These optimizations, coupled with adaptive sampling, provided improved read depth when compared to standard sequencing methods resulting in better quality assemblies, although also potentially introducing reference bias. Uneven sequencing coverage prevented successful *De* novo assembly, so full genomes were only possible using a reference-based assembly method. The bioinformatic pipeline development involved TPA read classification, read filtering, reference mapping to the SS-14 genome and assembly polishing. These steps produced a practical path for recovering thirteen high-quality TPA assemblies from clinical samples.

The process of differentiating high- and low-quality assemblies requires multiple QC steps. BUSCO completeness and QUAST provided initial quality thresholds but were not predictive on their own. Careful review and split multiple sequence alignments of the duplicate rRNA operons improved phylogenetic analysis and aided in identifying thirteen high quality Manitoba genomes. These clustered within the SS-14 clade showing limited regional structuring, including a small Manitoba-specific sub-clade, although no direct epidemiological links between cases could be identified. Given the limited sample size, it is not possible to determine the dominant TPA lineage in Manitoba.

Finally, in-silico analysis demonstrated that incomplete datasets were informative. The Diagnostic PCR target, *tp47,* was stable, and *arp* repeat counts were homogenous, matching earlier provincial surveillance.

In summary, this work demonstrates that Nanopore sequencing of TPA from clinical swabs is feasible with careful optimization and meaningful insights into circulating TPA can be made even when complete genomes are not recovered.

Appendix 1.1: Reduction of ambiguous bases (Ns per 100 kbp) after polishing with the best performing polisher.

Sample	Best Method	Unpolished Ns/100 kbp	Polished Ns/100 kbp	Absolute Reduction	Percent Reduction (%)
Sample 1	Racon ×1	187.4	0.0	187.4	100.0
Sample 2	Medaka ×1	16.4	0.0	16.4	100.0
Sample 3	Medaka ×1	41.7	0.0	41.7	100.0
Sample 4	Medaka ×1	22.9	0.0	22.9	100.0

Appendix 1.2: Reference Genomes Included in Phylogenetic Analyses of *Treponema pallidum*

ID	Cluster	Country	Collection Date	A2058G	Publication	DOI
CP016055.1	SS14	Portugal	2013	G (resistant)	Pinto_2016	10.1038/nmicrobiol.2016.190
CP016057.1	SS14	Portugal	2013	G (resistant)	Pinto_2016	10.1038/nmicrobiol.2016.190
CP016058.1	SS14	Portugal	2013	A	Pinto_2016	10.1038/nmicrobiol.2016.190
CP016063.1	SS14	Portugal	2014	G (resistant)	Pinto_2016	10.1038/nmicrobiol.2016.190
CP016064.1	SS14	Portugal	2014	G (resistant)	Pinto_2016	10.1038/nmicrobiol.2016.190
CP016065.1	SS14	Portugal	2014	G (resistant)	Pinto_2016	10.1038/nmicrobiol.2016.190
CP016066.1	SS14	Portugal	2014	G (resistant)	Pinto_2016	10.1038/nmicrobiol.2016.190
CP016067.1	SS14	Portugal	2014	G (resistant)	Pinto_2016	10.1038/nmicrobiol.2016.190
CP016068.1	SS14	Portugal	2014	G (resistant)	Pinto_2016	10.1038/nmicrobiol.2016.190
CP016069.1	SS14	Portugal	2014	G (resistant)	Pinto_2016	10.1038/nmicrobiol.2016.190
CP028438.1	SS14	Czechia	2014	G (resistant)	Grillová_2018	10.1371/journal.pone.0202619
CP034912.1	SS14	Czechia	2017	G (resistant)	Grillová_2019	10.3389/fmicb.2019.01691
CP034915.1	SS14	France	2016	G (resistant)	Grillová_2019	10.3389/fmicb.2019.01691
CP034916.1	SS14	France	2015	G (resistant)	Grillová_2019	10.3389/fmicb.2019.01691
CP034917.1	Nichols	Cuba	2015	A	Grillová_2019	10.3389/fmicb.2019.01691
CP034918.1	Nichols	Australia	2014	A	Grillová_2019	10.3389/fmicb.2019.01691
CP034919.1	Nichols	France	2012	A	Grillová_2019	10.3389/fmicb.2019.01691
CP034920.1	SS14	Cuba	2013	G (resistant)	Grillová_2019	10.3389/fmicb.2019.01691
CP034921.1	SS14	Czechia	2014	A	Grillová_2019	10.3389/fmicb.2019.01691
CP034972.1	Nichols	Cuba	2016	A	Grillová_2019	10.3389/fmicb.2019.01691
CP040555.1	SS14	China	2017	G (resistant)	Liu_2019	10.1101/2019.12.16.877886
CP045004.1	SS14	USA	2009	G (resistant)	Addetia_2020	10.1371/journal.pntd.0007921
CP045005.1	SS14	USA	2003	G (resistant)	Addetia_2020	10.1371/journal.pntd.0007921
CP073397.2	Nichols	Africa	2007	A	Lieberman_2021	10.1371/journal.pntd.0010063
CP073399.2	Nichols	Africa	2007	A	Lieberman_2021	10.1371/journal.pntd.0010063
CP073401.2	Nichols	Africa	2007	A	Lieberman_2021	10.1371/journal.pntd.0010063
CP073402.2	Nichols	Africa	2006	A	Lieberman_2021	10.1371/journal.pntd.0010063
CP073403.2	Nichols	Africa	2006	A	Lieberman_2021	10.1371/journal.pntd.0010063
CP073404.2	Nichols	Africa	2006	A	Lieberman_2021	10.1371/journal.pntd.0010063
CP073405.2	Nichols	Africa	2006	A	Lieberman_2021	10.1371/journal.pntd.0010063
CP073468.2	SS14	Peru	2019	G (resistant)	Lieberman_2021	10.1371/journal.pntd.0010063

CP073470.2	SS14	Peru	2019	G (resistant)	Lieberman_2021	10.1371/journal.pntd.0010063
CP073471.2	SS14	Peru	2019	G (resistant)	Lieberman_2021	10.1371/journal.pntd.0010063
CP073472.2	SS14	Peru	2018	G (resistant)	Lieberman_2021	10.1371/journal.pntd.0010063
CP073473.2	SS14	Peru	2018	G (resistant)	Lieberman_2021	10.1371/journal.pntd.0010063
CP073474.2	SS14	Peru	2018	A	Lieberman_2021	10.1371/journal.pntd.0010063
CP073477.2	SS14	USA	2001	A	Lieberman_2021	10.1371/journal.pntd.0010063
CP073479.2	SS14	USA	2001	A	Lieberman_2021	10.1371/journal.pntd.0010063
CP073480.2	SS14	USA	2000	A	Lieberman_2021	10.1371/journal.pntd.0010063
CP073490.2	SS14	USA	2002	A	Lieberman_2021	10.1371/journal.pntd.0010063
CP073493.2	SS14	Japan	2020	G (resistant)	Lieberman_2021	10.1371/journal.pntd.0010063
CP073494.2	SS14	Japan	2020	G (resistant)	Lieberman_2021	10.1371/journal.pntd.0010063
CP073495.2	SS14	Japan	2020	G (resistant)	Lieberman_2021	10.1371/journal.pntd.0010063
CP073497.2	SS14	Japan	2020	G (resistant)	Lieberman_2021	10.1371/journal.pntd.0010063
CP073498.2	SS14	Japan	2019	G (resistant)	Lieberman_2021	10.1371/journal.pntd.0010063
CP073499.2	SS14	Japan	2019	G (resistant)	Lieberman_2021	10.1371/journal.pntd.0010063
CP073500.2	SS14	Japan	2019	G (resistant)	Lieberman_2021	10.1371/journal.pntd.0010063
CP073501.2	SS14	Japan	2019	G (resistant)	Lieberman_2021	10.1371/journal.pntd.0010063
CP073502.2	SS14	Japan	2019	G (resistant)	Lieberman_2021	10.1371/journal.pntd.0010063
CP073547.2	SS14	Italy	2017	G (resistant)	Lieberman_2021	10.1371/journal.pntd.0010063
CP073548.2	SS14	Italy	2017	G (resistant)	Lieberman_2021	10.1371/journal.pntd.0010063
CP073549.2	SS14	Italy	2017	G (resistant)	Lieberman_2021	10.1371/journal.pntd.0010063
CP073550.2	SS14	Italy	2017	G (resistant)	Lieberman_2021	10.1371/journal.pntd.0010063
CP073551.2	SS14	Italy	2017	G (resistant)	Lieberman_2021	10.1371/journal.pntd.0010063
CP073552.2	SS14	Italy	2017	G (resistant)	Lieberman_2021	10.1371/journal.pntd.0010063
CP073553.2	Nichols	Italy	2017	A	Lieberman_2021	10.1371/journal.pntd.0010063
CP073554.2	SS14	Italy	2017	G (resistant)	Lieberman_2021	10.1371/journal.pntd.0010063
CP073555.2	SS14	Italy	2017	G (resistant)	Lieberman_2021	10.1371/journal.pntd.0010063
CP073556.2	SS14	Ireland	2002	G (resistant)	Lieberman_2021	10.1371/journal.pntd.0010063
CP073558.2	SS14	Ireland	2002	G (resistant)	Lieberman_2021	10.1371/journal.pntd.0010063
CP073559.2	SS14	Ireland	2002	G (resistant)	Lieberman_2021	10.1371/journal.pntd.0010063
CP073560.2	SS14	Ireland	2002	G (resistant)	Lieberman_2021	10.1371/journal.pntd.0010063
CP073561.2	SS14	Ireland	2002	G (resistant)	Lieberman_2021	10.1371/journal.pntd.0010063
CP073562.2	SS14	Ireland	2002	G (resistant)	Lieberman_2021	10.1371/journal.pntd.0010063
CP073564.2	SS14	Ireland	2002	G (resistant)	Lieberman_2021	10.1371/journal.pntd.0010063
CP073565.2	SS14	Ireland	2002	G (resistant)	Lieberman_2021	10.1371/journal.pntd.0010063
CP073567.2	SS14	China	2018	G (resistant)	Lieberman_2021	10.1371/journal.pntd.0010063

CP073568.2	SS14	China	2018	G (resistant)	Lieberman_2021	10.1371/journal.pntd.0010063
CP073569.2	SS14	China	2018	G (resistant)	Lieberman_2021	10.1371/journal.pntd.0010063
CP073571.2	SS14	China	2018	G (resistant)	Lieberman_2021	10.1371/journal.pntd.0010063
CP073572.2	SS14	China	2018	G (resistant)	Lieberman_2021	10.1371/journal.pntd.0010063
CP073573.2	SS14	China	2018	G (resistant)	Lieberman_2021	10.1371/journal.pntd.0010063
CP073574.2	SS14	China	2018	G (resistant)	Lieberman_2021	10.1371/journal.pntd.0010063
CP073575.2	SS14	Japan	2019	G (resistant)	Lieberman_2021	10.1371/journal.pntd.0010063
CP073576.2	Nichols	Italy	2017	G (resistant)	Lieberman_2021	10.1371/journal.pntd.0010063
CP104704.1	SS14	China	2018	G (resistant)	Lieberman_2021	10.1371/journal.pntd.0010063
CP104707.1	Nichols	China	2019	NA	Lieberman_2021	10.1371/journal.pntd.0010063
CP104708.1	SS14	China	2019	G (resistant)	Lieberman_2021	10.1371/journal.pntd.0010063
CP112930.1	SS14	Peru	2020	G (resistant)	Lieberman_2021	10.1371/journal.pntd.0010063
CP112931.1	SS14	Peru	2020	G (resistant)	Lieberman_2021	10.1371/journal.pntd.0010063
CP128599.1	SS14	USA	2021	G (resistant)	Unpublished	BioProject: PRJNA974070
NC_016843.1	Yaws	Africa	1960	A	Unpublished	BioProject: PRJNA224116
NC_021508.1	SS14	USA	1977	G (resistant)	Unpublished	10.1186/1471-2180-8-76
CP073566.2	SS14	Ireland	2002	G (resistant)	Lieberman_2021	10.1371/journal.pntd.0010063
CP073478.2	SS14	USA	2000	A	Lieberman_2021	10.1371/journal.pntd.0010063
CP073392.2	Nichols	Africa	2007	A	Lieberman_2021	10.1371/journal.pntd.0010063
CP073394.2	Nichols	Africa	2007	A	Lieberman_2021	10.1371/journal.pntd.0010063
CP073398.2	Nichols	Africa	2007	A	Lieberman_2021	10.1371/journal.pntd.0010063
AE000520.1	Nichols	USA	1912	A	Fraser 1998	10.1126/science.281.5375.375

Chapter 5: References

1. Tampa M, Sarbu I, Matei C, Benea V, Georgescu SR. Brief history of syphilis. J Medicine Life. 2014;7(1):4–10.

2. Flint VIJ. Christopher Columbus [Internet]. 2021 [cited 2021 Nov 13]. Available from: https://www.britannica.com/biography/Christopher-Columbus/The-first-voyage

3. Harper KN, Zuckerman MK, Harper ML, Kingston JD, Armelagos GJ. The origin and antiquity of syphilis revisited: An Appraisal of Old World pre-Columbian evidence for treponemal infection. Am J Phys Anthropol. 2011;146(S53):99–133.

4. HACKETT CJ. ON THE ORIGIN OF THE HUMAN TREPONEMATOSES (PINTA, YAWS, ENDEMIC SYPHILIS AND VENEREAL SYPHILIS). B World Health Organ. 1963;29:7–41.

5. Stamm LV. Pinta: Latin America's Forgotten Disease? Am Soc Trop Med Hyg. 2015;93(5):901–3.

6. Rothschild BM. History of Syphilis. Clin Infect Dis. 2005;40(10):1454–63.

7. Baker BJ, Armelagos GJ, Becker MJ, Brothwell D, Drusini A, Geise MC, et al. The Origin and Antiquity of Syphilis: Paleopathological Diagnosis and Interpretation [and Comments and Reply]. Curr Anthropol. 1988;29(5):703–37.

8. Quetel C. History of Syphilis. Baltimore, Maryland: The Johns Hopkin's University Press; 1990. 327 p.

9. McGough LJ. Syphilis in History: A Response to 2 Articles. Clin Infect Dis. 2005;41(4):573–5.

10. Zhou Y, Gao G, Zhang X, Gao B, Duan C, Zhu H, et al. Identifying treponemal disease in early East Asia. Am J Biol Anthr. 2022;178(3):530–43.

11. Melo FL de, Mello JCM de, Fraga AM, Nunes K, Eggers S. Syphilis at the Crossroad of Phylogenetics and Paleopathology. Plos Neglect Trop D. 2010;4(1):e575.

12. EKSELIUS L, GERDIN B, VAHLQUIST A. The Syphilis Pandemic Prior to Penicillin: Origin, Health Issues, Cultural Representation and Ethical Challenges. Acta Derm-Venereol. 2024;104:34879.

13. Szreter S, Siena K. The pox in Boswell's London: an estimate of the extent of syphilis infection in the metropolis in the 1770s†. Econ Hist Rev. 2021;74(2):372–99.

14. WOO EJ, KIM JH, LEE WJ, CHO H, PAK S. Syphilitic infection in a pre-modern population from South Korea (19th century AD). Anthr Sci. 2019;127(1):55–63.

15. Douglas JM. Penicillin Treatment of Syphilis. Jama. 2009;301(7):769–71.

16. Forrai J. Syphilis - Recognition, Description and Diagnosis. 2011;

17. Mahoney JF, Arnold RC, Harris A. Penicillin Treatment of Early Syphilis—A Preliminary Report. Am J Public Health N. 1943;33(12):1387–91.

18. Gelpi A, Tucker JD. A cure at last? Penicillin's unintended consequences on syphilis control, 1944–1964. Sex Transm Infect. 2015;91(1):70.

19. Epidemiology B of CD. Sexually Transmitted Disease in Canada. Canada Disease Weekly Reprot [Internet]. 1989 Apr 1; Available from: https://publications.gc.ca/collections/collection_2016/aspc-phac/H12-21-1-15-S2-eng.pdf

20. LEE CB, BRUNHAM RC, SHERMAN E, HARDING GKM. EPIDEMIOLOGY OF AN OUTBREAK OF INFECTIOUS SYPHILIS IN MANITOBA. Am J Epidemiol. 1987;125(2):277–83.

21. Canada PHA of. REPORT ON SEXUALLY TRANSMITTED INFECTIONS IN CANADA: 2011. Centre for Communicable Diseases and Infection Control, Infectious Disease Prevention and Control Branch, Public Health Agency of Canada.

22. Totten S, MacLean R, Payne E. Infectious syphilis in Canada: 2003-2012. Can Commun Dis Rep. 2015;41(2):30–4.

23. Canada PHA of. CCDR 49-10. Canada Communicable Disease Report. 2023;

24. Manitoba G of. Sexually Transmitted and Blood-Borne Infections (STBBI) Surveillance Report [Internet]. 2022 [cited 2023 Feb 3]. Available from: https://www.gov.mb.ca/health/publichealth/surveillance/stbbi/index.html

25. Canada PHA of. SYPHILIS IN CANADA: TECHNICAL REPORT ON EPIDEMIOLOGICAL TRENDS, DETERMINANTS AND INTERVENTIONS [Internet]. Ottawa: Government of Canada / PHAC; 2020 [cited 2022 Mar 9]. Available from: https://www.canada.ca/en/services/health/publications/diseases-conditions/syphilis-epidemiological-report.html#5

26. Beattie S, Ellis J, Pylypjuk C, Liu XQ, Poliquin V. Retrospective Cohort Study of Syphilis-Related Stillbirths in Winnipeg, Manitoba From 2017–2020. J Obstet Gynaecol Can. 2024;46(5):102356.

27. Manitoba G of. Epidemiology and Surveillance. 2025 [cited 2025 Aug 31]. Sexually Transmitted and Blood-Borne Infections (STBBI) Surveillance Report. Available from: https://www.gov.mb.ca/health/publichealth/surveillance/stbbi/index.html

28. Lopez A, Lee SJ, Bullard J. Syphilis in pregnancy and infant outcomes in Manitoba. Paediatr Child Heal. 2022;27(3):183–9.

29. LaFond RE, Lukehart SA. Biological Basis for Syphilis. Clin Microbiol Rev. 2006;19(1):29–49.

30. Radolf JD, Tramont EC, Salazar JC. Mandell, Douglas, and Bennett's Principles and Practice of Infectious Diseases (Eighth Edition) [Internet]. 8th ed. Bennett JE, Dolin R, Blaser MJ, editors. 2015. 2684–2709 p. Available from: https://doi.org/10.1016/B978-1-4557-4801-3.00239-3.

31. Boodman C, Bullard J, Stein DR, Lee S, Poliquin V, Caeseele PV. Expanded prenatal syphilis screening in Manitoba, Canada: a direct short-term cost-avoidance analysis in an outbreak context. Can J Public Heal. 2023;114(2):287–94.

32. Peeling RW, Mabey D, Kamb ML, Chen XS, Radolf JD, Benzaken AS. Syphilis. Nat Rev Dis Primers. 2017;3(1):17073.

33. Radolf JD, Kumar S. Spirochete Biology: The Post Genomic Era. Curr Top Microbiol. 2017;1–38.

34. Norris SJ, Cox DL, Weinstock GM. Biology of Treponema pallidum: correlation of functional activities with genome sequence data. J Mol Microbiol Biotechnol. 2001;3(1):37–62.

35. Walker EM, Zampighi GA, Blanco DR, Miller JN, Lovett MA. Demonstration of rare protein in the outer membrane of Treponema pallidum subsp. pallidum by freeze-fracture analysis. J Bacteriol. 1989;171(9):5005–11.

36. Cox DL, Chang P, McDowall AW, Radolf JD. The outer membrane, not a coat of host proteins, limits antigenicity of virulent Treponema pallidum. Infect Immun. 1992;60(3):1076–83.

37. Canale-Parola E. Motility and Chemotaxis of Spirochetes. Annu Rev Microbiol. 1978;32(1):69–97.

38. Fraser CM, Norris SJ, Weinstock GM, White O, Sutton GG, Dodson R, et al. Complete Genome Sequence of Treponema pallidum, the Syphilis Spirochete. Science. 1998;281(5375):375–88.

39. Walker EM, Arnett JK, Heath JD, Norris SJ. Treponema pallidum subsp. pallidum has a single, circular chromosome with a size of approximately 900 kilobase pairs. Infect Immun. 1991;59(7):2476–9.

40. Grillová L, Carrami EM, Roberts-Sengier W, Thomson NR. High quality transcriptome profiling confirms the transcriptional landscape of Treponema pallidum subsp. pallidum. Sci Rep. 2025;15(1):23272.

41. Radolf JD, Deka RK, Anand A, Šmajs D, Norgard MV, Yang XF. Treponema pallidum, the syphilis spirochete: making a living as a stealth pathogen. Nat Rev Microbiol. 2016;14(12):744–59.

42. Kent ME, Romanelli F. Reexamining Syphilis: An Update on Epidemiology, Clinical Manifestations, and Management. Ann Pharmacother. 2008;42(2):226–36.

43. Mikalová L, Strouhal M, Čejková D, Zobaníková M, Pospíšilová P, Norris SJ, et al. Genome Analysis of Treponema pallidum subsp. pallidum and subsp. pertenue Strains: Most of the Genetic Differences Are Localized in Six Regions. PLoS ONE. 2010;5(12):e15713.

44. Štaudová B, Strouhal M, Zobaníková M, Čejková D, Fulton LL, Chen L, et al. Whole Genome Sequence of the Treponema pallidum subsp. endemicum Strain Bosnia A: The

Genome Is Related to Yaws Treponemes but Contains Few Loci Similar to Syphilis Treponemes. PLoS Neglected Trop Dis. 2014;8(11):e3261.

45. Šmajs D, Norris SJ, Weinstock GM. Genetic diversity in Treponema pallidum: Implications for pathogenesis, evolution and molecular diagnostics of syphilis and yaws. Infect, Genet Evol. 2012;12(2):191–202.

46. Hackett CJ. On the epidemiology of yaws in African miners (1942). Trans R Soc Trop Med Hyg. 1984;78(4):536–8.

47. Willcox RR. The Evolutionary Cycle of the Treponematoses*. Br J Vener Dis. 1960;36(2):78.

48. Giacani L, Lukehart SA. The Endemic Treponematoses. Clin Microbiol Rev. 2014;27(1):89–115.

49. Karem KL, Pillay A. On the Origin of Syphilis and Contemporary Views of Disease Dynamics. J Anc Dis Prev Remedies. 2014;2014(03).

50. Nyatsanza F, Tipple C. Syphilis: presentations in general medicine. Clin Med. 2016;16(2):184–8.

51. Peeling RW, Mabey D, Chen XS, Garcia PJ. Syphilis. Lancet. 2023;402(10398):336–46.

52. Whiting C, Schwartzman G, Khachemoune A. Syphilis in Dermatology: Recognition and Management. Am J Clin Dermatol. 2023;24(2):287–97.

53. CHAPEL TA. The Signs and Symptoms of Secondary Syphilis. Sex Transm Dis. 1980;7(4):161–4.

54. Cerqueira LRP de, Monteiro DLM, Taquette SR, Rodrigues NCP, Trajano AJB, Souza FM de, et al. The magnitude of syphilis: from prevalence to vertical transmission. Rev Inst Med Trop São Paulo. 2017;59(0):e78.

55. Stoltey JE, Cohen SE. Syphilis transmission: a review of the current evidence. Sex Heal. 2015;12(2):103–9.

56. David M, Hcini N, Mandelbrot L, Sibiude J, Picone O. Fetal and neonatal abnormalities due to congenital syphilis: A literature review. Prenat Diagn. 2022;42(5):643–55.

57. O'Shea JG. 'Two Minutes with Venus, Two Years with Mercury'-Mercury as an Antisyphilitic Chemotherapeutic Agent. J R Soc Med. 1989;83(6):392–5.

58. Harrison LW. Ehrlich versus Syphilis. Br J Vener Dis. 1954;30(1):2.

59. Frith J. Syphilis - Its Early History and Treatment Until Penicillin, and the Debate on its Origins. Journal of Military and Veterans' Health [Internet]. 2012;20(4):49–58. Available from: https://jmvh.org/article/syphilis-its-early-history-and-treatment-until-penicillin-and-the-debate-on-its-origins/

60. Vernon G. Syphilis and Salvarsan. Br J Gen Pr. 2019;69(682):246–246.

61. HOOK EW, MARTIN DH, STEPHENS J, SMITH BS, SMITH K. A Randomized, Comparative Pilot Study of Azithromycin Versus Benzathine Penicillin G for Treatment of Early Syphilis. Sex Transm Dis. 2002;29(8):486–90.

62. Rekart ML, Patrick DM, Chakraborty B, Maginley JJ, Jones H, Bajdik CD, et al. Targeted mass treatment for syphilis with oral azithromycin. Lancet. 2003;361(9354):313–4.

63. Lukehart SA, Charmie G, J. MB, Patricia S, Susan H, Fiona M, et al. Macrolide Resistance in Treponema pallidum in the United States and Ireland. N Engl J Med. 2004;351(2):154–8.

64. Grant JR, Enns E, Marinier E, Mandal A, Herman EK, Chen C yu, et al. Proksee: in-depth characterization and visualization of bacterial genomes. Nucleic Acids Res. 2023;51(W1):W484–92.

65. Liu H, Rodes B, Chen CY, Steiner B. New Tests for Syphilis: Rational Design of a PCR Method for Detection of Treponema pallidum in Clinical Specimens Using Unique Regions of the DNA Polymerase I Gene. J Clin Microbiol. 2001;39(5):1941–6.

66. Orle KA, Gates CA, Martin DH, Body BA, Weiss JB. Simultaneous PCR detection of Haemophilus ducreyi, Treponema pallidum, and herpes simplex virus types 1 and 2 from genital ulcers. J Clin Microbiol. 1996;34(1):49–54.

67. Gayet-Ageron A, Combescure C, Lautenschlager S, Ninet B, Perneger TV. Comparison of Diagnostic Accuracy of PCR Targeting the 47-Kilodalton Protein

Membrane Gene of Treponema pallidum and PCR Targeting the DNA Polymerase I Gene: Systematic Review and Meta-analysis. J Clin Microbiol. 2015;53(11):3522–9.

68. Hedley A, Bullard J, Caeseele PV, Shaw S, Tsang R, Alexander DC, et al. A case for implementing an HSV1/2, VZV, and syphilis lesion panel in Manitoba, Canada. Microbiol Spectr. 2024;12(8):e00600-24.

69. Deka RK, Machius M, Norgard MV, Tomchick DR. Crystal Structure of the 47-kDa Lipoprotein of Treponema pallidum Reveals a Novel Penicillin-binding Protein*. J Biol Chem. 2002;277(44):41857–64.

70. Gayet-Ageron A, Ninet B, Toutous-Trellu L, Lautenschlager S, Furrer H, Piguet V, et al. Assessment of a real-time PCR test to diagnose syphilis from diverse biological samples. Sex Transm Infect. 2009;85(4):264.

71. Cha JY, Ishiwata A, Mobashery S. A Novel β-Lactamase Activity from a Penicillin-binding Protein of Treponema pallidum and Why Syphilis Is Still Treatable with Penicillin*. J Biol Chem. 2004;279(15):14917–21.

72. Liu H, Rodes B, George R, Steiner B. Molecular characterization and analysis of a gene encoding the acidic repeat protein (Arp) of Treponema pallidum. J Méd Microbiol. 2007;56(6):715–21.

73. Harper KN, Liu H, Ocampo PS, Steiner BM, Martin A, Levert K, et al. The sequence of the acidic repeat protein (arp) gene differentiates venereal from nonvenereal Treponema pallidum subspecies, and the gene has evolved under strong positive

selection in the subspecies that causes syphilis. FEMS Immunol Méd Microbiol. 2008;53(3):322–32.

74. Lieberman NAP, Armstrong TD, Chung B, Pfalmer D, Hennelly CM, Haynes A, et al. High-throughput nanopore sequencing of Treponema pallidum tandem repeat genes arp and tp0470 reveals clade-specific patterns and recapitulates global whole genome phylogeny. Front Microbiol. 2022;13:1007056.

75. PILLAY A, LIU H, CHEN CY, HOLLOWAY B, STURM WA, STEINER B, et al. Molecular Subtyping of Treponema pallidum Subspecies pallidum. Sex Transm Dis. 1998;25(8):408–14.

76. Pinto M, Borges V, Antelo M, Pinheiro M, Nunes A, Azevedo J, et al. Genome-scale analysis of the non-cultivable Treponema pallidum reveals extensive within-patient genetic variation. Nat Microbiol. 2016;2(1):16190.

77. Matějková P, Strouhal M, Šmajs D, Norris SJ, Palzkill T, Petrosino JF, et al. Complete genome sequence of Treponema pallidum ssp. pallidum strain SS14 determined with oligonucleotide arrays. Bmc Microbiol. 2008;8(1):76–76.

78. Romeis E, Lieberman NAP, Molini B, Tantalo LC, Chung B, Phung Q, et al. Treponema pallidum subsp. pallidum with an Artificially impaired TprK antigenic variation system is attenuated in the Rabbit model of syphilis. PLOS Pathog. 2023;19(3):e1011259.

79. Giacani L, Brandt SL, Puray-Chavez M, Reid TB, Godornes C, Molini BJ, et al. Comparative Investigation of the Genomic Regions Involved in Antigenic Variation of the

TprK Antigen among Treponemal Species, Subspecies, and Strains. J Bacteriol. 2012;194(16):4208–25.

80. Centurion-Lara A, Giacani L, Godornes C, Molini BJ, Reid TB, Lukehart SA. Fine Analysis of Genetic Diversity of the tpr Gene Family among Treponemal Species, Subspecies and Strains. Plos Neglect Trop D. 2013;7(5):e2222.

81. Giacani L, Godornes C, Puray-Chavez M, Guerra-Giraldez C, Tompa M, Lukehart SA, et al. TP0262 is a modulator of promoter activity of tpr Subfamily II genes of Treponema pallidum ssp. pallidum. Mol Microbiol. 2009;72(5):1087–99.

82. Burstain JM, Grimprel E, Lukehart SA, Norgard MV, Radolf JD. Sensitive detection of Treponema pallidum by using the polymerase chain reaction. J Clin Microbiol. 1991;29(1):62–9.

83. Noordhoek GT, Wolters EC, Jonge ME de, Embden JD van. Detection by polymerase chain reaction of Treponema pallidum DNA in cerebrospinal fluid from neurosyphilis patients before and after antibiotic treatment. J Clin Microbiol. 1991;29(9):1976–84.

84. Liu H, Rodes B, Chen CY, Steiner B. New Tests for Syphilis: Rational Design of a PCR Method for Detection of Treponema pallidum in Clinical Specimens Using Unique Regions of the DNA Polymerase I Gene. J Clin Microbiol. 2001;39(5):1941–6.

85. Theel ES, Katz SS, Pillay A. Molecular and Direct Detection Tests for Treponema pallidum Subspecies pallidum: A Review of the Literature, 1964–2017. Clin Infect Dis. 2020;71(Supplement_1):S4–12.

86. Suntoke TR, Hardick A, Tobian AAR, Mpoza B, Laeyendecker O, Serwadda D, et al. Evaluation of multiplex real-time PCR for detection of Haemophilus ducreyi, Treponema pallidum, herpes simplex virus type 1 and 2 in the diagnosis of genital ulcer disease in the Rakai District, Uganda. Sex Transm Infect. 2009;85(2):97.

87. Chi KH, Danavall D, Taleo F, Pillay A, Ye T, Nachamkin E, et al. Molecular Differentiation of Treponema pallidum Subspecies in Skin Ulceration Clinically Suspected as Yaws in Vanuatu Using Real-Time Multiplex PCR and Serological Methods. Am Soc Trop Med Hyg. 2015;92(1):134–8.

88. Peng RR, Wang AL, Li J, Tucker JD, Yin YP, Chen XS. Molecular Typing of Treponema pallidum: A Systematic Review and Meta-Analysis. PLoS Neglected Trop Dis. 2011;5(11):e1273.

89. Mikalová L, Pospíšilová P, Woznicová V, Kuklová I, Zákoucká H, Šmajs D. Comparison of CDC and sequence-based molecular typing of syphilis treponemes: tpr and arp loci are variable in multiple samples from the same patient. BMC Microbiol. 2013;13(1):178.

90. Grillová L, Pětrošová H, Mikalová L, Strnadel R, Dastychová E, Kuklová I, et al. Molecular Typing of Treponema pallidum in the Czech Republic during 2011 to 2013: Increased Prevalence of Identified Genotypes and of Isolates with Macrolide Resistance. J Clin Microbiol. 2014;52(10):3693–700.

91. Grange PA, Allix-Beguec C, Chanal J, Benhaddou N, Gerhardt P, Morini JP, et al. Molecular Subtyping of Treponema pallidum in Paris, France. Sex Transm Dis. 2013;40(8):641–4.

92. Grange PA, Mikalová L, Gaudin C, Strouhal M, Janier M, Benhaddou N, et al. Treponema pallidum 11qj Subtype May Correspond to a Treponema pallidum Subsp. Endemicum Strain. Sex Transm Dis. 2016;43(8):517–8.

93. Noda AA, Grillová L, Lienhard R, Blanco O, Rodríguez I, Šmajs D. Bejel in Cuba: molecular identification of Treponema pallidum subsp. endemicum in patients diagnosed with venereal syphilis. Clin Microbiol Infec. 2018;24(11):1210.e1-1210.e5.

94. Grillová L, Bawa T, Mikalová L, Gayet-Ageron A, Nieselt K, Strouhal M, et al. Molecular characterization of Treponema pallidum subsp. pallidum in Switzerland and France with a new multilocus sequence typing scheme. Plos One. 2018;13(7):e0200773.

95. Grillová L, Giacani L, Mikalová L, Strouhal M, Strnadel R, Marra C, et al. Sequencing of Treponema pallidum subsp. pallidum from isolate UZ1974 using Anti-Treponemal Antibodies Enrichment: First complete whole genome sequence obtained directly from human clinical material. Plos One. 2018;13(8):e0202619.

96. Grillová L, Oppelt J, Mikalová L, Nováková M, Giacani L, Niesnerová A, et al. Directly Sequenced Genomes of Contemporary Strains of Syphilis Reveal Recombination-Driven Diversity in Genes Encoding Predicted Surface-Exposed Antigens. Front Microbiol. 2019;10:1691.

97. Thurlow CM, Joseph SJ, Ganova-Raeva L, Katz SS, Pereira L, Chen C, et al. Selective Whole-Genome Amplification as a Tool to Enrich Specimens with Low Treponema pallidum Genomic DNA Copies for Whole-Genome Sequencing. Msphere. 2022;e00009-22.

98. Arora N, Schuenemann VJ, Jäger G, Peltzer A, Seitz A, Herbig A, et al. Origin of modern syphilis and emergence of a pandemic Treponema pallidum cluster. Nat Microbiol. 2016;2(1):16245.

99. Chen W, Šmajs D, Hu Y, Ke W, Pospíšilová P, Hawley KL, et al. Analysis of Treponema pallidum Strains From China Using Improved Methods for Whole-Genome Sequencing From Primary Syphilis Chancres. J Infect Dis. 2020;223(5):848–53.

100. Strouhal M, Oppelt J, Mikalová L, Arora N, Nieselt K, González-Candelas F, et al. Reanalysis of Chinese Treponema pallidum samples: all Chinese samples cluster with SS14-like group of syphilis-causing treponemes. Bmc Res Notes. 2018;11(1):16.

101. Stamm LV, Stapleton JT, Bassford PJ. In vitro assay to demonstrate high-level erythromycin resistance of a clinical isolate of Treponema pallidum. Antimicrob Agents Chemother. 1988;32(2):164–9.

102. Nechvátal L, Pětrošová H, Grillová L, Pospíšilová P, Mikalová L, Strnadel R, et al. Syphilis-causing strains belong to separate SS14-like or Nichols-like groups as defined by multilocus analysis of 19 Treponema pallidum strains. Int J Méd Microbiol. 2014;304(5–6):645–53.

103. Beale MA, Marks M, Cole MJ, Lee MK, Pitt R, Ruis C, et al. Global phylogeny of Treponema pallidum lineages reveals recent expansion and spread of contemporary syphilis. Nat Microbiol. 2021;6(12):1549–60.

104. Taouk ML, Taiaroa G, Pasricha S, Herman S, Chow EPF, Azzatto F, et al. Characterisation of Treponema pallidum lineages within the contemporary syphilis outbreak in Australia: a genomic epidemiological analysis. Lancet Microbe. 2022;3(6):e417–26.

105. Nishiki S, Lee K, Kanai M, Nakayama S ichi, Ohnishi M. Phylogenetic and genetic characterization of Treponema pallidum strains from syphilis patients in Japan by whole-genome sequence analysis from global perspectives. Sci Rep. 2021;11(1):3154.

106. Morando N, Vrbová E, Melgar A, Rabinovich RD, Šmajs D, Pando MA. High frequency of Nichols-like strains and increased levels of macrolide resistance in Treponema pallidum in clinical samples from Buenos Aires, Argentina. Sci Rep-uk. 2022;12(1):16339.

107. Lieberman NAP, Lin MJ, Xie H, Shrestha L, Nguyen T, Huang ML, et al. Treponema pallidum genome sequencing from six continents reveals variability in vaccine candidate genes and dominance of Nichols clade strains in Madagascar. Plos Neglect Trop D. 2021;15(12):e0010063.

108. Miller JN. Immunity in Experimental Syphilis. J Immunol. 1973;110(5):1206–15.

109. Liu A, Giacani L, Hawley KL, Cameron CE, Seña AC, Konda KA, et al. New Pathways in Syphilis Vaccine Development. Sex Transm Dis. 2024;51(11):e49–53.

110. Sun ES, Molini BJ, Barrett LK, Centurion-Lara A, Lukehart SA, Voorhis WCV. Subfamily I Treponema pallidum repeat protein family: sequence variation and immunity. Microbes Infect. 2004;6(8):725–37.

111. Centurion-Lara A, Castro C, Barrett L, Cameron C, Mostowfi M, Voorhis WCV, et al. Treponema pallidum Major Sheath Protein Homologue Tpr K Is a Target of Opsonic Antibody and the Protective Immune Response. J Exp Med. 1999;189(4):647–56.

112. Parker ML, Houston S, Pětrošová H, Lithgow KV, Hof R, Wetherell C, et al. The Structure of Treponema pallidum Tp0751 (Pallilysin) Reveals a Non-canonical Lipocalin Fold That Mediates Adhesion to Extracellular Matrix Components and Interactions with Host Cells. PLoS Pathog. 2016;12(9):e1005919.

113. Houston S, Hof R, Francescutti T, Hawkes A, Boulanger MJ, Cameron CE. Bifunctional Role of the Treponema pallidum Extracellular Matrix Binding Adhesin Tp0751. Infect Immun. 2010;79(3):1386–98.

114. Lithgow KV, Hof R, Wetherell C, Phillips D, Houston S, Cameron CE. A defined syphilis vaccine candidate inhibits dissemination of Treponema pallidum subspecies pallidum. Nat Commun. 2017;8(1):14273.

115. Lukehart SA, Molini B, Gomez A, Godornes C, Hof R, Fernandez MC, et al. Immunization with a tri-antigen syphilis vaccine significantly attenuates chancre

development, reduces bacterial load, and inhibits dissemination of Treponema pallidum. Vaccine. 2022;40(52):7676–92.

116. Desrosiers DC, Anand A, Luthra A, Dunham-Ems SM, LeDoyt M, Cummings MAD, et al. TP0326, a Treponema pallidum β-barrel assembly machinery A (BamA) orthologue and rare outer membrane protein. Mol Microbiol. 2011;80(6):1496–515.

117. Cameron CE, Lukehart SA, Castro C, Molini B, Godornes C, Voorhis WCV. Opsonic Potential, Protective Capacity, and Sequence Conservation of the Treponema pallidum subspecies pallidum Tp92. J Infect Dis. 2000;181(4):1401–13.

118. Fieldsteel AH, Cox DL, Moeckli RA. Cultivation of Virulent Treponema pallidum in Tissue Culture. Infect Immun. 1981;32(2):908–15.

119. Edmondson DG, Hu B, Norris SJ. Long-Term In Vitro Culture of the Syphilis Spirochete Treponema pallidum subsp. pallidum. Mbio. 2018;9(3):e01153-18.

120. Turner TB, Hollander DH, Organization WH. Biology of the Treponematoses. World Health Organization monograph series [Internet]. 1957;278. Available from: https://iris.who.int/handle/10665/41677

121. Edmondson DG, Norris SJ. In Vitro Cultivation of the Syphilis Spirochete Treponema pallidum. Curr Protoc. 2021;1(2):e44.

122. Hay PE, Clarke JR, Strugnell RA, Taylor-Robinson D, Goldmeier D. Use of the polymerase chain reaction to detect DNA sequences specific to pathogenic treponemes in cerebrospinal fluid. FEMS Microbiol Lett. 1990;68(3):233–8.

123. Arora N, Schuenemann VJ, Jäger G, Peltzer A, Seitz A, Herbig A, et al. Origin of modern syphilis and emergence of a pandemic Treponema pallidum cluster. Nat Microbiol. 2016;2(1):16245.

124. Beale MA, Thorn L, Cole MJ, Pitt R, Charles H, Ewens M, et al. Genomic epidemiology of syphilis in England: a population-based study. Lancet Microbe. 2023;

125. Thoendel M, Jeraldo PR, Greenwood-Quaintance KE, Yao JZ, Chia N, Hanssen AD, et al. Comparison of microbial DNA enrichment tools for metagenomic whole genome sequencing. J Microbiol Meth. 2016;127:141–5.

126. Heravi FS, Zakrzewski M, Vickery K, Hu H. Host DNA depletion efficiency of microbiome DNA enrichment methods in infected tissue samples. J Microbiol Meth. 2020;170:105856.

127. Charalampous T, Kay GL, Richardson H, Aydin A, Baldan R, Jeanes C, et al. Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. Nat Biotechnol. 2019;37(7):783–92.

128. Radolf JD, Chamberlain NR, Clausell A, Norgard MV. Identification and localization of integral membrane proteins of virulent Treponema pallidum subsp. pallidum by phase partitioning with the nonionic detergent triton X-114. Infect Immun. 1988;56(2):490–8.

129. Majander K, Pfrengle S, Kocher A, Neukamm J, Plessis L du, Pla-Díaz M, et al. Ancient Bacterial Genomes Reveal a High Diversity of Treponema pallidum Strains in Early Modern Europe. Curr Biol. 2020;30(19):3788-3803.e10.

130. Munnink BBO, Nieuwenhuijse DF, Stein M, O'Toole Á, Haverkate M, Mollers M, et al. Rapid SARS-CoV-2 whole-genome sequencing and analysis for informed public health decision-making in the Netherlands. Nat Med. 2020;26(9):1405–10.

131. Golparian D, Donà V, Sánchez-Busó L, Foerster S, Harris S, Endimiani A, et al. Antimicrobial resistance prediction and phylogenetic analysis of Neisseria gonorrhoeae isolates using the Oxford Nanopore MinION sequencer. Sci Rep. 2018;8(1):17596.

132. Ferreira FA, Helmersen K, Visnovska T, Jørgensen SB, Aamot HV. Rapid nanopore-based DNA sequencing protocol of antibiotic-resistant bacteria for use in surveillance and outbreak investigation. Microb Genom. 2021;7(4):000557.

133. Dai T, Qu R, Liu J, Zhou P, Wang Q. Efficacy of Doxycycline in the Treatment of Syphilis. Antimicrob Agents Chemother. 2016;61(1):10.1128/aac.01092-16.

134. Luetkemeyer AF, Donnell D, Dombrowski JC, Cohen S, Grabow C, Brown CE, et al. Postexposure Doxycycline to Prevent Bacterial Sexually Transmitted Infections. N Engl J Med. 2023;388(14):1296–306.

135. Hu M, Nandi S, Davies C, Nicholas RA. High-Level Chromosomally Mediated Tetracycline Resistance in Neisseria gonorrhoeae Results from a Point Mutation in the

rpsJ Gene Encoding Ribosomal Protein S10 in Combination with the mtrR and penB Resistance Determinants. Antimicrob Agents Chemother. 2005;49(10):4327–34.

136. Wizemann TM, Heinrichs JH, Adamou JE, Erwin AL, Kunsch C, Choi GH, et al. Use of a Whole Genome Approach To Identify Vaccine Molecules Affording Protection against Streptococcus pneumoniae Infection. Infect Immun. 2001;69(3):1593–8.

137. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet. 2020;395(10224):565–74.

138. Wang Y, Zhao Y, Bollas A, Wang Y, Au KF. Nanopore sequencing technology, bioinformatics and applications. Nat Biotechnol. 2021;39(11):1348–65.

139. Martin S, Heavens D, Lan Y, Horsfield S, Clark MD, Leggett RM. Nanopore adaptive sampling: a tool for enrichment of low abundance species in metagenomic samples. Genome Biol. 2022;23(1):11.

140. Ong CT, Ross EM, Boe-Hansen GB, Turni C, Hayes BJ, Tabor AE. Technical note: overcoming host contamination in bovine vaginal metagenomic samples with nanopore adaptive sequencing. J Anim Sci. 2021;100(1):skab344.

141. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021;10(2):giab008.

142. Coster WD, D'Hert S, Schultz DT, Cruts M, Broeckhoven CV. NanoPack: visualizing and processing long-read sequencing data. Bioinformatics. 2018;34(15):2666–9.

143. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019;20(1):257.

144. Wick R. Filtlong [Internet]. 2017 [cited 2022 Sept 27]. Available from: https://github.com/rrwick/Filtlong

145. Coster WD, Rademakers R. NanoPack2: population-scale evaluation of long-read sequencing data. Bioinformatics. 2023;39(5):btad311.

146. Kolmogorov M, Bickhart DM, Behsaz B, Gurevich A, Rayko M, Shin SB, et al. metaFlye: scalable long-read metagenome assembly using repeat graphs. Nat Methods. 2020;17(11):1103–10.

147. Vaser R, Šikić M. Time- and memory-efficient genome assembly with Raven. Nat Comput Sci. 2021;1(5):332–6.

148. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol. 2017;13(6):e1005595.

149. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100.

150. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29(8):1072–5.

151. Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 2017;27(5):737–46.

152. Huang YT, Liu PY, Shih PW. Homopolish: a method for the removal of systematic errors in nanopore sequencing by homologous polishing. Genome Biol. 2021;22(1):95.

153. Manni M, Berkeley MR, Seppey M, Simao FA, Zdobnov EM. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. arXiv. 2021;

154. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30(14):2068–9.

155. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002;30(14):3059–66.

156. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Haeseler A von, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol. 2020;37(5):1530–4.

157. Kalyaanamoorthy S, Minh BQ, Wong TKF, Haeseler A von, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017;14(6):587–9.

158. Hoang DT, Chernomor O, Haeseler A von, Minh BQ, Vinh LS. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol Biol Evol. 2018;35(2):518–22.

159. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinform. 2009;10(1):421.

160. Chen CY, Chi KH, George RW, Cox DL, Srivastava A, Silva MR, et al. Diagnosis of Gastric Syphilis by Direct Immunofluorescence Staining and Real-Time PCR Testing. J Clin Microbiol. 2006;44(9):3452–6.

161. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27(2):573–80.

162. Consortium IHGS, Research: WI for BR Center for Genome, Lander ES, Linton LM, Birren B, Nusbaum C, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860–921.

163. Leichty AR, Brisson D. Selective Whole Genome Amplification for Resequencing Target Microbial Species from Complex Natural Samples. Genetics. 2014;198(2):473–81.

164. Sundararaman SA, Plenderleith LJ, Liu W, Loy DE, Learn GH, Li Y, et al. Genomes of cryptic chimpanzee Plasmodium species reveal key evolutionary events leading to human malaria. Nat Commun. 2016;7(1):11078.

165. Itsko M, Retchless AC, Joseph SJ, Turner AN, Bazan JA, Sadji AY, et al. Full Molecular Typing of Neisseria meningitidis Directly from Clinical Specimens for Outbreak Investigation. J Clin Microbiol. 2020;58(12).

166. Technologies ON. Flow Cells and Nanopores [Internet]. 2025 [cited 2025 Mar 13]. Available from: https://nanoporetech.com/platform/technology/flow-cells-and-nanopores

167. Ludwig W, Schleifer KH. Bacterial phylogeny based on 16S and 23S rRNA sequence analysis. FEMS Microbiol Rev. 1994;15(2-3):155–73.

168. Ma DY, Giacani L, Centurión-Lara A. The molecular epidemiology of Treponema pallidum subspecies pallidum. Sex Heal. 2015;12(2):141–7.

169. Tsang RS, Shuel M, Hoang W, Hayden K, Hink R, Bullard J, et al. Characteristics of polymerase chain reaction–positive syphilis cases in Manitoba, Canada, 2017 to 2020: Demographic analysis, specimen types, and Treponema pallidum gene targets. Off J Assoc Méd Microbiol Infect Dis Can. 2022;7(3):170–80.

170. Eagle SHC, Robertson J, Bastedo DP, Liu K, Nash JHE. Evaluation of five commercial DNA extraction kits using Salmonella as a model for implementation of rapid Nanopore sequencing in routine diagnostic laboratories. Access Microbiol. 2023;5(2):000468.v3.

171. Marotz CA, Sanders JG, Zuniga C, Zaramela LS, Knight R, Zengler K. Improving saliva shotgun metagenomics by chemical host DNA depletion. Microbiome. 2018;6(1):42.

172. Addetia A, Tantalo LC, Lin MJ, Xie H, Huang ML, Marra CM, et al. Comparative genomics and full-length Tprk profiling of Treponema pallidum subsp. pallidum reinfection. Plos Neglect Trop D. 2020;14(4):e0007921.

173. Beaudry MS, Bhuiyan MIU, Glenn TC. Enriching the future of public health microbiology with hybridization bait capture. Clin Microbiol Rev. 2024;37(4):e00068-22.

174. Freed NE, Vlková M, Faisal MB, Silander OK. Rapid and inexpensive whole-genome sequencing of SARS-CoV-2 using 1200 bp tiled amplicons and Oxford Nanopore Rapid Barcoding. Biology Methods Protoc. 2020;5(1):bpaa014.

175. Sereika M, Kirkegaard RH, Karst SM, Michaelsen TY, Sørensen EA, Wollenberg RD, et al. Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. Nat Methods. 2022;19(7):823–6.

176. Lee JY, Kong M, Oh J, Lim J, Chung SH, Kim JM, et al. Comparative evaluation of Nanopore polishing tools for microbial genome assembly and polishing strategies for downstream analysis. Sci Rep. 2021;11(1):20740.

177. Vrbová E, Noda AA, Grillová L, Rodríguez I, Forsyth A, Oppelt J, et al. Whole genome sequences of Treponema pallidum subsp. endemicum isolated from Cuban

patients: The non-clonal character of isolates suggests a persistent human infection rather than a single outbreak. PLoS Neglected Trop Dis. 2022;16(6):e0009900.

178. Pla-Díaz M, Akgül G, Molak M, Plessis L du, Panagiotopoulou H, Doan K, et al. Insights into Treponema pallidum genomics from modern and ancient genomes using a novel mapping strategy. BMC Biol. 2025;23(1):7.

179. Shuel M, Hayden K, Kadkhoda K, Tsang RSW. Molecular Typing and Macrolide Resistance of Syphilis Cases in Manitoba, Canada, From 2012 to 2016. Sex Transm Dis. 2018;45(4):233–6.

180. Singh N, Braukmann TWA, Neale M, Long GS, Stein D, Caeseele PV, et al. Complete genome sequences of two Treponema pallidum subsp. pallidum specimens from Canadian patients. Microbiol Resour Announc. 2025;14(10):e00641-25.

Supplementary Information:

Supplementary Information 1: split.fasta.py

```
import os from Bio import SeqIO
def find_first_5s_gene_end(gff_file):
"""Finds the end position of the first 5S gene near 236000 in the GFF file.""
with open(gff_file) as gff:
for line in gff:
# Skip comments and ensure line has enough fields if not line.startswith("#") and len(line.strip().split("\t")) >= 5:
fields = line.strip().split("\t")
feature_type = fields[2]
start_position = int(fields[3])

if feature_type == "rRNA" and "5S" in fields[-1] and 230000 <= start_position <= 240000:
end_position = int(fields[4])
return end_position
return None  # If no gene is found
def split_fasta(input_fasta, gff_file, output_dir):
"""Splits the FASTA file based on the end of the first 5S gene near 236000 and saves two parts."""
sample_id = os.path.splitext(os.path.basename(input_fasta))[0]
split_position = find_first_5s_gene_end(gff_file)
if split_position is None:
print(f"No 5S gene found near position 236000 in {gff_file}. Skipping {input_fasta}.")
 Return

# Output filenames for the two segments
output_fasta_1 = os.path.join(output_dir, f"{sample_id}_1.fasta")
output_fasta_2 = os.path.join(output_dir, f"{sample_id}_2.fasta")

with open(output_fasta_1, "w") as out1, open(output_fasta_2, "w") as out2: for record in SeqIO.parse(input_fasta, "fasta"):
part1_seq = record.seq[:split_position]
part2_seq = record.seq[split_position:]

SeqIO.write(SeqIO.SeqRecord(part1_seq, id=record.id, description=""), out1, "fasta")
SeqIO.write(SeqIO.SeqRecord(part2_seq, id=record.id, description=""), out2, "fasta")

print(f"Split {sample_id} at position {split_position} based on the first 5S gene.")
```

```
def process_all_samples(fasta_dir, gff_parent_dir, output_dir):
"""Processes all FASTA and GFF files in their respective directories."""
os.makedirs(output_dir, exist_ok=True)
for fasta_file in os.listdir(fasta_dir):
if fasta_file.endswith(".fasta"):
sample_id = os.path.splitext(fasta_file)[0]
gff_file = os.path.join(gff_parent_dir, sample_id, f"{sample_id}.gff")
fasta_path = os.path.join(fasta_dir, fasta_file)

if os.path.exists(gff_file):
split_fasta(fasta_path, gff_file, output_dir)
else:
print(f"GFF file for {sample_id} not found in {gff_parent_dir}/{sample_id}. Skipping.")

directory list
fasta_dir = "q-ref-fasta"
gff_parent_dir = "all/14-prokka-prot"
output_dir = "/split-fasta"
process_all_samples(fasta_dir, gff_parent_dir, output_dir)
```

Supplementary Information 2: merge-alignments.py

```
from Bio import SeqIO
Set Paths to aligned files
aligned_split1 = "split1.aligned.fasta"
aligned_split2 = "split2.aligned.fasta"
output_combined = "2025-01-27.aligned.fasta"

store the combined sequences
combined_sequences = {}

Load split 1 sequences
for record in SeqIO.parse(aligned_part1, "fasta"):
sample_id = record.id # Keep the full
sample ID combined_sequences[sample_id] = str(record.seq)

Load split 2 sequences and combine with split 1 sequences
for record in SeqIO.parse(aligned_part2, "fasta"):
sample_id = record.id
if sample_id in combined_sequences:
combined_sequences[sample_id] += str(record.seq)
else:
```

```
# contingency for when sample ID is not in split 1
combined_sequences[sample_id] = str(record.seq)

Write combined sequences to a new FASTA file
with open(output_combined, "w") as output_file:
for sample_id, sequence in combined_sequences.items():
output_file.write(f">{sample_id}\n{sequence}\n")

print(f"Combined aligned sequences saved to {output_combined}")
```

Supplementary Information 3: R Script to produce QC tree

```
library(ape) l
ibrary(ggtree)
library(readxl)
library(tidyverse)
library(ggnewscale)
#this script makes a tree with the green/red QC heat maps
----------------------------------------------------------------------
Set working directory and file paths
----------------------------------------------------------------------
setwd("Tree/prune tree")
tree_file <- "busco316.treefile"
metadata_file <- "tree_metadata.xlsx"
quast_study_path <- "/quast/transposed_report.tsv"
quast_ref_path <- "ref-quast/transposed_report.tsv"
----------------------------------------------------------------------
Read and prep tree + metadata
----------------------------------------------------------------------
tree <- ape::read.tree(tree_file)
tree <- ape::root(tree, outgroup = "NC_016843.1", resolve.root = TRUE)

metadata <- read_excel(metadata_file) %>% rename( sample_id = Header, cluster = Cluster, country = Pays, macrolide_snp = A2058G, sample_origin = Origin) %>% filter(sample_id %in% tree$tip.label)

metadata <- metadata %>% mutate(sample_origin = ifelse(sample_origin == "This Study", "Manitoba Specimens", sample_origin))
----------------------------------------------------------------------
Load and merge QUAST data
----------------------------------------------------------------------
```

```
quast_study <- read_tsv(quast_study_path, col_types = cols_only( Assembly = col_character(), # indels per 100 kbp = col_double(), # N's per 100 kbp = col_double(), # mismatches per 100 kbp = col_double(), Genome fraction (%) = col_double()))

quast_ref <- read_tsv(quast_ref_path, col_types = cols_only( Assembly = col_character(), # indels per 100 kbp = col_double(), # N's per 100 kbp = col_double(), # mismatches per 100 kbp = col_double(), Genome fraction (%) = col_double()))

quast_combined <- bind_rows(quast_study, quast_ref) %>% rename( sample_id = Assembly, indels_per_100kbp = # indels per 100 kbp, Ns_per_100kbp = # N's per 100 kbp, mismatches_per_100kbp = # mismatches per 100 kbp, genome_fraction = Genome fraction (%)) %>% mutate( low_indels = indels_per_100kbp < 5, low_Ns = Ns_per_100kbp <= 300, low_mismatches = mismatches_per_100kbp < 20, high_genome_frac = genome_fraction > 99)

metadata <- left_join(metadata, quast_combined, by = "sample_id")
-------------------------------------------------------------------
Collapse selected reference-heavy nodes
-------------------------------------------------------------------
tree_base <- ggtree(tree) %<+% metadata

nodes_to_collapse <- c()

tree_collapsed <- tree_base for (node in nodes_to_collapse) { tree_collapsed <- collapse(tree_collapsed, node = node)}

tree_labeled <- tree_collapsed + geom_point2(aes(subset = (node %in% nodes_to_collapse)), shape = 21, size = 3, fill = "steelblue") + geom_tiplab(aes(label = label), size = 2.5, offset = 0.0002, align = TRUE)
-------------------------------------------------------------------
Heatmap 1: Sample Origin
-------------------------------------------------------------------
sample_origin_data <- metadata %>% select(sample_id, sample_origin) %>% column_to_rownames("sample_id")

h1 <- gheatmap(tree_labeled, sample_origin_data, offset = 0.00095, width = 0.02, colnames = FALSE) + scale_fill_manual( name = "Sample Origin", values = c("Reference Sequence" = "#1b9e77", "Manitoba Specimens" = "#d95f02"), guide = guide_legend(order = 1))

#Heatmap 2: Indels
h2 <- h1 + new_scale_fill() indel_data <- metadata %>% select(sample_id, low_indels) %>% mutate(low_indels = ifelse(low_indels, "< 5", "≥ 5")) %>% column_to_rownames("sample_id")
```

```
h3 <- gheatmap(h2, indel_data, offset = 0.00115, width = 0.02, colnames = FALSE) + scale_fill_manual( name = "Indels per 100kbp", values = c("< 5" = "green", "≥ 5" = "red"), guide = guide_legend(order = 2)) + annotate("text", x = 0.01005, y = length(tree$tip.label) * 1.01, label = "Indels", angle = 60, hjust = 0)

#Heatmap 3: Ns
h4 <- h3 + new_scale_fill() ns_data <- metadata %>% select(sample_id, low_Ns) %>% mutate(low_Ns = ifelse(low_Ns, "≤ 300", "> 300")) %>% column_to_rownames("sample_id")

h5 <- gheatmap(h4, ns_data, offset = 0.00135, width = 0.02, colnames = FALSE) + scale_fill_manual( name = "N's per 100kbp", values = c("≤ 300" = "green", "> 300" = "red"), guide = guide_legend(order = 3)) + annotate("text", x = 0.01025, y = length(tree$tip.label) * 1.01, label = "Ns", angle = 60, hjust = 0)

#Heatmap 4: Mismatches
h6 <- h5 + new_scale_fill() mismatch_data <- metadata %>% select(sample_id, low_mismatches) %>% mutate(low_mismatches = ifelse(low_mismatches, "< 20", "≥ 20")) %>% column_to_rownames("sample_id")

h7 <- gheatmap(h6, mismatch_data, offset = 0.00155, width = 0.02, colnames = FALSE) + scale_fill_manual( name = "Mismatches per 100kbp", values = c("< 20" = "green", "≥ 20" = "red"), guide = guide_legend(order = 4)) + annotate("text", x = 0.01045, y = length(tree$tip.label) * 1.01, label = "Mismatches", angle = 60, hjust = 0)

#Heatmap 5: Genome Fraction
h8 <- h7 + new_scale_fill() genome_frac_data <- metadata %>% select(sample_id, high_genome_frac) %>% mutate(high_genome_frac = ifelse(high_genome_frac, "> 99", "≤ 99")) %>% column_to_rownames("sample_id")

final_plot <- gheatmap(h8, genome_frac_data, offset = 0.00175, width = 0.02, colnames = FALSE) + scale_fill_manual( name = "Genome Fraction", values = c("> 99" = "green", "≤ 99" = "red"), guide = guide_legend(order = 5)) + annotate("text", x = 0.01065, y = length(tree$tip.label) * 1.01, label = "Genome %", angle = 60, hjust = 0) + geom_tippoint( data = subset(tree_collapsed$data, label == "NC_016843.1"), aes(x = x, y = y), shape = 21, size = 6, fill = "red", color = "black") + theme_tree2() + xlab("") + xlim(0, 0.011) + ylim(0, 150) + ggtitle("") + theme( legend.position = "right", legend.title = element_text(size = 20), legend.text = element_text(size = 20), legend.box = "vertical", legend.margin = margin(), axis.line.x = element_blank(), axis.ticks.x = element_blank(), axis.text.x = element_blank(), axis.title.x = element_blank())

---------------------------------------------------------------------
Print final tree
```

```
---------------------------------------------------------------------
print(final_plot)
save image as 1500x1500 and crop
```

Supplementary Information 4: R Script to produce Final tree

```
library(ape)
library(ggtree)
library(tidyverse) l
ibrary(readxl)
library(ggnewscale)
library(ggimage)
---------------------------------------------------------------------
Set working directory and file paths
---------------------------------------------------------------------
setwd("thesis figures/Tree")
tree_file <- "thesis figures/Tree/2025-05_green/green.treefile"
metadata_file <- "tree_metadata.xlsx"
quast_study_path <- "thesis figures/assembly assessment/quast/transposed_report.tsv"
quast_ref_path <- "thesis figures/Tree/prune tree/ref-quast/transposed_report.tsv"

#Load tree and metadata
tree <- ape::read.tree(tree_file)
tree <- ape::root(tree, outgroup = "NC_016843.1", resolve.root = TRUE)

#Load metadata and filter to included tips
metadata <- read_excel(metadata_file) %>% rename( sample_id = Header, clade = Cluster, macrolide_snp = A2058G, sample_origin = Origin) %>% filter(sample_id %in% tree$tip.label)

#Rename "This Study" to "Manitoba Specimens" in the sample_origin column
metadata <- metadata %>% mutate(sample_origin = ifelse(sample_origin == "This Study", "Manitoba Specimens", sample_origin))

#Attach metadata to tree
tree_plot <- ggtree(tree) %<+% metadata

#Heatmap 1: Sample Origin
sample_origin_data <- metadata %>% select(sample_id, sample_origin) %>% column_to_rownames("sample_id")

h1 <- gheatmap( tree_plot, sample_origin_data, offset = 0.00064, width = 0.06, colnames = FALSE) + scale_fill_manual( name = "Sample Origin", values = c( "Reference
```

```
Sequence" = "#1b9e77", "Manitoba Specimens" = "#d95f02"), guide = guide_legend(order = 1))

#Heatmap 2: Clade
h2 <- h1 + new_scale_fill()
clade_data <- metadata %>% select(sample_id, clade) %>% column_to_rownames("sample_id")
h3 <- gheatmap( h2, clade_data, offset = 0.00078, width = 0.06, colnames = FALSE) + scale_fill_manual( name = "Clade", values = c( "SS14" = "#F2A900", "Nichols" = "#3C5E9B", "Yaws" = "#FF0000", "Manitoba" = "#710A1B"), guide = guide_legend(order = 2))

#Heatmap 3: Macrolide SNP
h4 <- h3 + new_scale_fill()

macrolide_snp_data <- metadata %>% select(sample_id, macrolide_snp) %>% mutate(macrolide_snp = case_when( macrolide_snp == "A" ~ "A (Susceptible)", macrolide_snp == "G (resistant)" ~ "G (Resistant)", TRUE ~ macrolide_snp)) %>% column_to_rownames("sample_id")

h5 <- gheatmap( h4, macrolide_snp_data, offset = 0.00092, width = 0.06, colnames = FALSE) + scale_fill_manual( name = "Macrolide SNP", values = c( "A (Susceptible)" = "#00d1b1", "G (Resistant)" = "purple"), guide = guide_legend(order = 3))
#Heatmap 4: Country

#Heatmap 4: Country with custom continent-tinted colors
h6 <- h5 + new_scale_fill()

#Clean and format country column
country_data <- metadata %>% select(sample_id, country = Pays) %>% mutate(country = ifelse(is.na(country) | country == "NA", "Unknown", country)) %>% column_to_rownames("sample_id")

#Manual assignment of continent-themed colors
country_to_color <- c(
#North America (red variants)
"Canada" = "#FF0000",
"USA" = "#67000d",
"Cuba" = "#FCA5A5",

#Europe (green variants)
"Portugal" = "#006D5B",
"France" = "#4dff4d",
"Italy" = "#94B94B",
```

```
"Czechia" = "#9FE1B0",
"Ireland" = "#019529",

#Asia
"China" = "#FDAE6B",
 "Japan" = "#cc4b00",

#Africa
"Africa" = "#ffff33",

#South America
"Peru" = "#377EB8",

#Oceania
"Australia" = "#D5D514",

#Fallback
"Unknown" = "#000")

#Generate heatmap
h7 <- gheatmap( h6, country_data, offset = 0.00106, width = 0.06, colnames = FALSE) + scale_fill_manual( name = "Country", values = country_to_color, guide = guide_legend(order = 4))

#Add heatmap labels
heatmap_labels <- list( annotate("text", x = 0.00312, y = length(tree$tip.label) * 1.015, label = "Sample Origin", angle = 60, hjust = 0), annotate("text", x = 0.00326, y = length(tree$tip.label) * 1.015, label = "Clade", angle = 60, hjust = 0), annotate("text", x = 0.00340, y = length(tree$tip.label) * 1.015, label = "Macrolide SNP", angle = 60, hjust = 0), annotate("text", x = 0.00354, y = length(tree$tip.label) * 1.015, label = "Country", angle = 60, hjust = 0))

#Final tree styling
final_tree <- h7 + heatmap_labels + geom_tiplab( aes(label = label), size = 2.5, offset = 0.0005, align = TRUE) + geom_tippoint( data = subset(tree_plot$data, label == "NC_016843.1"), aes(x = x, y = y), shape = 21, size = 3, fill = "red", color = "black") + theme_tree2() + xlim(0, 0.0037) + ylim(0, length(tree$tip.label) * 1.05) + coord_cartesian(clip = "off") + ggtitle("") + theme( legend.position = "right", legend.title = element_text(size = 20), legend.text = element_text(size = 20), legend.key.size = unit(1, "cm"), legend.box = "vertical", legend.margin = margin(t = 0, r = 5, b = 0, l = 0), axis.line.x = element_blank(), axis.ticks.x = element_blank(), axis.text.x = element_blank(), axis.title.x = element_blank())
---------------------------------------------------------------------
Print final tree
```

```
---------------------------------------------------------------------
print(final_tree)
#export tree as .png with 1500x1700 aspect ratio
```

Supplementary Information 5: Bash script to pull Tp47 genes with >50x depth

```
#!/bin/bash

#Directory list
parent_gff_dir="all/14-prokka-prot"
fasta_dir="/mnt/sata1/Tpal/all/11.1-fasta"
depth="/mnt/sata1/Tpal/all/121-tp47/tp47_gene_depth.csv"
output_dir="/mnt/sata1/Tpal/all/121-tp47/50x-tp47"

#Create output directory
mkdir -p "$output_dir"

#Read the depth into array
declare -A min_depths
while IFS=, read -r sample_id avg_depth min_depth; do
# Skip the header row
 if [[ "$sample_id" != "SampleID" ]]; then
# Remove ".sorted" suffix
normalized_id=$(echo "$sample_id" | sed 's/.sorted$//')
# output qc: Print each sample and its depth
echo "Read depth data: SampleID=$normalized_id MinDepth=$min_depth"
min_depths["$normalized_id"]=$min_depth
fi
done < "$depth_csv"

#Iterate over each Prokka result folder
for folder in "$parent_gff_dir"/; do
if [[ -d "$folder" ]]; then
# Locate the GFF
gff_file=$(find "$folder" -name ".gff")
sample_id=$(basename "$folder")
fasta_file="$fasta_dir/${sample_id}.fasta"

# Check if the sample exists
if [[ -v min_depths["$sample_id"] ]]; then
min_depth=${min_depths["$sample_id"]}

# Check sample min depth >= 50
```

```
if [[ $min_depth -ge 50 ]]; then
# Ensure both the GFF and FASTA exist
if [[ -f "$gff_file" && -f "$fasta_file" ]]; then
# Extract Tp47 coordinates using `product` or `locus_tag`
tp47_coordinates=$(grep -i "product=.*Tp47" "$gff_file" | awk '{print $1, $4, $5, $7}')

if [[ -n "$tp47_coordinates" ]]; then
chr=$(echo "$tp47_coordinates" | awk '{print $1}')
start=$(echo "$tp47_coordinates" | awk '{print $2}')
end=$(echo "$tp47_coordinates" | awk '{print $3}')
strand=$(echo "$tp47_coordinates" | awk '{print $4}')

# Define output file
tp47_output="${output_dir}/${sample_id}_Tp47.fasta"

# Extract the Tp47 sequence
samtools faidx "$fasta_file" "${chr}:${start}-${end}" > "$tp47_output"

echo "Extracted Tp47 for $sample_id into $tp47_output"
else
echo "Tp47 gene not found in $gff_file. Skipping..."
fi
else
echo "GFF or FASTA file missing for $sample_id. Skipping..."
fi
else
echo "Sample $sample_id does not meet the minimum read depth of 50. Skipping..."
fi
else
echo "Sample $sample_id not found in depth data. Skipping..."
fi
fi
done
```