Development of a mycobacterial enzymatic lysis reagent

Joseph Manion

A thesis submitted for the degree of Doctor of Philosophy

2020 – 2024

Biosciences Institute

Framlington Place

Newcastle University

Abstract

Mycobacteria are responsible for significant mortality and morbidity worldwide. They have a complex, multi-layered cell wall including peptidoglycan, arabinogalactan and mycolic acid which contributes to their pathogenesis and antibiotic resistance. Diagnosis of mycobacterial infection and resistance profiles can be slow and inaccurate, especially in lower and middle-income countries where species like *Mycobacterium tuberculosis* are endemic. However lysis of these cells for DNA extraction which are crucial for diagnostics and tracking of these diseases often involve long incubation times, harsh mechanical lysis, or toxic chemicals such as phenol:chloroform. The aim of this thesis is to develop a low-cost, rapid, enzymatic lysis reagent.

In this investigation we characterise a founding member of a new family of glycoside hydrolases, DG02470, from the gut Bacteroidota species *Dysgonomonas gadei* capable of hydrolysing β-D-arabinofuranose linkages within mycobacterial arabinogalactan. Through biochemical and structural analysis, we have characterised the kinetic activity and structure of DG02470.

Using DG02470 in combination with enzymes previously identified by the lab which target arabinogalactan, we have a developed a DNA extraction reagent using a modified GenElute™ Bacterial Genomic DNA Kit protocol. Enzymatic lysis extraction showed a significant increase in yield of gDNA when the protocol was used against *M. smegmatis* and *M. abscessus* when compared to extraction without the novel enzymes which target arabinogalactan and was comparable with bead beating extraction, which is not compatible with long read sequencing with Oxford Nanopore technology . We also show a significant increase in yield when the lysis reagent is utilised for genomic extraction from *M. avium* subsp. *paratuberculosis*, *M*. *bovis* BCG and *Nocardia farcinica.* Finally, we have shown we can obtain gDNA from the extracted DNA using the lysis reagent, which can be used for long read sequencing with Oxford Nanopore technology. Enzymatic lysis enabled complete genome assembly with fewer contigs when compared to bead beating. The enzymatic lysis reagent presented in this investigation has the potential to allow for a more accessible method of DNA extraction from mycobacteria, aiding in future research and diagnostics.

Acknowledgements

I would like to first thank Dr Elisabeth Lowe for the opportunity to do this PhD and the support provided throughout my PhD, I am eternally grateful for this. From aiding my understanding of glycobiology to creating a highly enjoyable environment to study a PhD.

I am grateful to each member of the M.2035 of the Cookson building lab past and present Carl Morland for provided guidance and help, particularly in the initial stages of my PhD with teaching methods used for the protein extraction and purification used throughout my PhD. Dr Dave Bolam for his aid in understanding biochemical principles as well singing. A special thanks to Dr Jenn Ross whose knowledge and understanding has been amazing as well as her guidance and help. Additionally, all of PhD students including Omar Al-Jourani, Cosette Hinkley, Diana Githwe, Kaspar Garnham, Natalia Los who have made the entire experience extremely enjoyable.

A thanks to other members of Newcastle University who have made this possible. Dr Arnaud Basle for aid in the structural biology aspects of my PhD including training in setting up of crystal trays, solving of the structure of DG02470 and help in understanding of the principles of structural biology. Dr Jon Marles-Wright who was instrumental in gaining an understanding of Linux and was always available for troubleshooting. A thank you to Prof. Jeremy Lakey and Dr Henrik Strahl for providing me with ideas and insight during my annual panels.

A massive thank you to those of the Moynihan Lab, including Dr. Patrick Moynihan, Amar Gudka and Abie Layton, as well as the Quick lab: Dr. Josh Quick and Natalie Sparks at Birmingham University who made me feel extremely welcome during my stay in Birmingham as well as being amazing collaborators and support throughout.

Finally, my family who have always been there for me, my brother Christopher Manion and his wife Ester Tripoldi for being gracious hosts on my escapes to Italy. My mum and dad Maria and Stephen Manion for constant support throughout my academic career.

Contents

List of Tables

List of Figures

Abbreviation	Meaning
AS	alignment score
HPAEC-PAD	High-performance anion exchanged chromatography with Pulsed Amperometric Detection
GND	Genome neighbourhood diagrams
GH	Glycoside Hydrolase
CG AG	*C. glutamicum* arabinogalactan
TLC	Thin-Layer Chromatography
LB	Luria-Bertani Broth
BHI	Brain-Heart infusion
TSB	Tryptic Soy Broth
PUL	Polysaccharide utilisation loci
Man-LAM	mannose capped lipoarabinomannans
AG	Arabinoglaatan
GPL	Glycopeptidolipids
LAM	lipoarabinomannans
LM	Lipomannans
MAC	*Mycobacterium avium complex*
MAC-PD	Mycobacterium avium complex pulmonary disease
mAGP	Mycolyl-arabinogalactan-peptidoglycan
Mtb	*Mycobactrium tuberculosis*
MTBC	*M. tuberculosis* complex
NTM	Non-tuberculous mycobacteria
NTM-PD	Non-tuberculous mycobacteria pulmonary disease
PG	Peptidoglycan
PIM	Phosphatidylinositol mannosides
RGM	Rapid growing mycobacteria
SGM	Slow growing mycobacteria
WGS	Whole Genome sequencing

Chapter 1. Introduction

1. General introduction.

Humans and bacteria have existed side by side since humans evolved. Whether that be cooperatively or pathogenically (Aujoulat et al., 2012). As a species we have used bacteria unknowingly for fermentation and other processes for a large portion of our history (Ozen & Dinleyici, 2015). Beneficially, bacteria play a major role in the human gut microbiota allowing for access to nutrients that are not innately accessible to humans, as well as playing an important role in preventing colonisation by pathogenic bacteria. The two dominant phyla in the human gut being Bacteroidota and Firmicutes (Hou et al., 2022).

However, the relationship between humans and bacteria is not always beneficial, pathogenic bacteria have evolved alongside humans adapting to our changing lifestyles and societal structures (Gluckman et al., 2016). One such genus mycobacteria, has evolved alongside humanity for thousands of years with the most well studied and dominant of these bacteria being *Mycobactrium tuberculosis* (Mtb) (BañUls et al., 2015; Gagneux, 2018). Tuberculosis (Tb) is still the largest killer of any infectious disease with 1.3 million deaths attributed to Tb in 2023 (WHO, 2023a). However, less reported mycobacterial infections are those cause by non-tuberculous mycobacteria (NTMs), comprising mycobacterium not belonging to the *M. tuberculosis* complex (MTBC), which include *M. tuberculosis* and *M. bovis* or *M. leprae.* NTMs such as the *M. avium* complex and the *M. abscessus* complex

have been on the rise since the 1990s, especially among the immunocompromised community (Dahl et al., 2022; Johansen et al., 2020; Nasiri et al., 2018; Ratnatunga et al., 2020). *M. abscessus* has become a greater problem in the cystic fibrosis community with the infection proving highly contagious and difficult to treat, leading to lung transplants being excluded as a treatment as an option for those infected due to the poor treatment outcomes, even after antibiotic treatment (Bryant et al., 2021).

As the burden of mycobacterial disease increases, adaptations to current molecular diagnostics are needed, which while currently effective, have limited depth, being tailored towards identification of known genes and identifiers. Advancements in Whole Genome Sequencing (WGS) provide a more accessible and cheaper route to high powered diagnostics. Companies such as Oxford Nanopore have provided a route to whole genome sequencing technology that is both affordable and portable, allowing for quicker sequencing and more accurate tracking of outbreaks. However, mycobacteria do pose a challenge due to their multilayered cell wall consisting of peptidoglycan, arabinogalactan and mycolic acids, making the extraction of gDNA from mycobacteria difficult using conventional lysozyme extractions that are common among commercial kits. Most current protocols for the extraction involve the use of chloroform or phenol, hazardous chemicals that require training to use as well as more specialist disposal, or the use of a bead beater, which are initially expensive and the mechanical nature of the lysis from bead-beating also causes shearing of the DNA affecting its usefulness in downstream analysis (Bouso & Planet, 2019; Epperson & Strong, 2020; Kumar et al., 2016). Although phenol:chloroform extraction is effective it requires >18 hours for classic protocols and equipment, such as fume hood for safe use of phenol or -70^{o}C freezers for ethanol precipitation which is often not available to a field lab or rural medical facility (Käser et al., 2009; Warren et al., 2006). In wealthier countries, in which NTMs are more commonly reported, diagnosis is done in specialist medical hospitals, often after a consultations and referrals (Haworth et al., 2017a). The lack of availability of these facilities across much of the world, especially in areas with high Tb burden, including sub-Saharan African countries and east Asian countries, often leads to the misdiagnosis of NTMs as Tb due to the overlapping symptoms (Prevots & Marras, 2015). While Tb diagnosis has a fast reliable and portable diagnosis tool in the form of GeneXpert MTB/RIF, a rapid NTM diagnosis system is currently not available.

The gut microbiota, with the composition consisting primarily of Firmicutes and Bacteroidota (Rinninella et al., 2019), has shown to be a great resource for the discovery of novel enzymes aimed at a variety of both industrial and clinical applications (Jia et al., 2022). One such classification of enzymes, Carbohydrate active enzymes (CAZymes), which catalyse the modification, creation and breakdown of carbohydrates, have been

shown to have great biopharmaceuticals applications in a wide range of fields from diagnoses to cancer treatment (Wardman et al., 2022). Glycoside hydrolases are one class of CAZymes. A recent study by Al-Jourani et al. (2023) identified glycoside hydrolases active upon α-D-arabinofuranose linkages that degrade the arabinose domain of arabinogalactan, into arabinose monosaccharides from gut bacteria *D. gadei.* This discovery in conjunction with previously identified β-D-Galactofuranosidases from *Bacteroides finegoldii* may potentially provide a route to an enzymatic lysis protocol which can be performed with commercial DNA extractions kits, that can allow for adequate DNA yields at high enough quality for WGS.

1.1 Acid-Fast bacteria

Acid-fast bacteria are a group of bacteria that are resistant to decolouration by acids during staining due to the high mycolic acid content of their cell walls (Reynolds et al., 2009). These bacteria include the genera *Mycobacterium* and *Nocardia.* Both are identifiable via Ziehl-Neelsen stain (McMurray, 1996). These are characterised by a thin layer of peptidoglycan, a layer of arabinogalactan and an outer membrane containing mycolic acid and overlaid with a variety of polypeptides and glycolipids, they are found spread throughout the world inhabiting a vast array of environmental niches.

1.1.1 Mycobacteria

Mycobacteria are rod-shaped, non-spore forming, aerobic, acid fast non-motile bacteria, first described in 1882 by Robert Koch, then known as *Tubercle bacillus*. The characteristic waxy morphology of mycobacterium is provided via their unique cell wall. The genus is large and varied including strict pathogens, opportunistic pathogens and non-pathogenic species (Forbes et al., 2018).

The classification of mycobacteria into three groups is shown in **Table 1.1**, MTBC, NTMs and *M. leprae,* the latter of these will not be discussed here as it is genotypically and

phenotypically distinct from any other identified *Mycobacterium* having a reduced genome and is represented in a separate clade (Cole et al., 2001).

Table 1.1.1 Categorisation of mycobacteria. Table of the three groups of mycobacteria: NTMs, *M. tuberculosis* complex and *M. leprae* with examples of species present within each**.**

Rapid growing mycobacteria	**Slow growing mycobacteria**		
Non-tuberculous mycobacteria		*M. tuberculosis* complex	*M. leprae*
M. abscessus complex	*M. avium* complex		
M. fortutium	*M. hemophilium*		
M. smegmatis	*M. xenopi*		
M. vaccae	*M. kansaii*		
	M. simiae		
	M. Terrae complex		
	M. gordonae		
	M. marinum		

1.1.1.1 Non-tuberculous mycobacterium

There are over 200 identified species of NTMs which have currently been identified to date (Matsumoto et al., 2019). NTMs are classified via the Runyon classification, which classifies NTMs based on their growth rate; slow growing mycobacteria (SGM) are classified by taking longer than 7 days to grow, and rapid growing mycobacteria (RGM) under 7 days. SGMs can further be characterised by their production of yellow pigment, in the form of carotenoids, before or after exposure to light. There are 4 classifications, with three dividing SGMs and one for all RGM these are: Runyon I - Photochromogens which are slow grown a produce the yellow pigment when exposed to light; Runyon II - Scotochromogens which are slow growing and produce pigments regardless of exposure to light; Runyon III - Nonchromogens which are slow growing and do not produce pigment; and finally, Runyon IV - Rapid Growers which are all RGMs and do not produce pigment (Runyon, 1959).

While NTMs are mainly non-pathogenic environmental organisms inhabiting environmental niches such as soil and water, species such as *M. avium* complex, *M. kansasii* and *M. abscessus* are classed as opportunistic pathogens, causing infections predominantly in those with underlying medical conditions, these infections are becoming an increasing threat to global human health (Ratnatunga et al., 2020).

1.1.2 Nocardia

Nocardia are filamentous environmental bacteria named after their discoverer French veterinarian Edmond Nocard in 1988 (Fatahi-Bafghi, 2018). There are roughly 100 species of *Nocardia* at the time of writing, with about 50 identified as human or animal pathogens. They can cause cutaneous, subcutaneous and disseminated infection, although cases are rare (Lafont et al., 2020a; Mehta & Shamoo, 2020) .

1.1.3 Diseases

Acid-fast bacteria are responsible for a wide range of disease from Nocardiosis caused predominantly by *N. asteroides* to Tb caused by *M. tuberculosis.*

The primary method of transmission for both *Nocardia* and mycobacteria is inhalation, for both *Nocardia* and mycobacterial infections this predominantly happens from environmental reservoirs. Both are often associated with immunocompromised individuals such as those with HIV or on antibiotic treatments. Due to the nature of the cell wall (**Section 1.4**) both have a high level of intrinsic resistance to antimicrobial agents. However, the antibiotic resistance suite of the bacteria varies greatly making identification down to species and strain level increasingly important (Duggal & Chugh, 2020; Saxena et al., 2021).

1.1.1.2 Mycobacterial diseases

Mycobacteria cause a wide variety of diseases with the most well documented being Tb, with roughly 7.5 million people developing the disease and a reported 1.3 million dying in 2023. Rates are back on the decline after the rise partially blamed on the effects of COVID-19 (WHO, 2023a). An estimated one third of the world population are infected with latent Tb, with the active form of the disease in 5% of cases (Lee, 2016). Spread through the air, the infection particles are small enough to reach into the lower airways. Although most bacilli are killed by alveolar macrophages, some are able to survive, infecting alveoli epithelial cells. This subsequently leads to an immune response with the recruitment of T and B cells, and the formation of a granuloma in which Mtb can replicate (Frieden et al., 2003).

Tb is not the only disease caused by mycobacterial species, they also cause a wide variety of diseases outside of this including the often forgotten about disease leprosy, caused by *M. leprae*, over 200 000 new cases reported each year (WHO, 2023b). Finally, there are NTMs which are considered mainly opportunistic pathogens, predominantly affecting those of the immunocompromised community such as cystic fibrosis patients and patients with immune cell abnormalities (Sexton & Harrison, 2008). Most reports show a global increase in the prevalence of NTM infection and disease with *M. avium* complex (MAC) and *M. abscessus* complex being those most common (Dahl et al., 2022; Johansen et al., 2020; Ratnatunga et al., 2020).

NTM diseases most commonly manifest as pulmonary disease (PNTM). These can occur in one of three forms: hypersensitivity pneumonitis, cavity tuberculosis-like disease or nodular bronchiectasis (Ratnatunga et al., 2020). Hypersensitivity pneumonitis is associated with prolonged exposure to the NTM, it is a granulomatous inflammatory lung disease caused by constant inhalation of pathogenic agents from areas such as hot tubs and particulates from farming activities (Daito et al., 2011). Tuberculosis-like pulmonary disease is most associated with underlying health conditions such as COPD or prior

tuberculosis. This form of pathogenesis is most common with RGMs, MAC and *M. kansaii*. Finally, nodular bronchitis is most common in older women with no history of smoking, with symptoms being a non-specific chronic cough with or without sputum (Weiss & Glassroth, 2012). Pulmonary NTMs cause a wide range of clinical symptoms, the most notable being a chronic cough that often produces mucus. The non-specific nature of these symptoms makes of NTM diagnosis particularly difficult and time consuming (Glassroth, 2008; Griffith et al., 2007).

Of particular concern is the prevalence of NTMs in the cystic fibrosis community with rates of infection of 7.9% among patients (Prieto et al., 2023). Although these patients may not develop PNTM the presence of the bacteria is enough to prevent lung transplants due to the poor treatment outcome (Aliberti et al., 2020). The global burden and prevalence of each NTM varies greatly from region to region however an overall increasing trend is prevalent (Dahl et al., 2022), with a lack of global reporting leading to difficulty in understanding the extent of the burden of NTM. Many published studies are case studies or regional analysis, or misdiagnosis of NTMs as Tb (Maiga et al., 2012).

1.2.3.1.1 M. abscessus

First isolated in 1951 by Moore & Frerichs, (1953) it was only in 1992 that *M. abscessus* was separated from *M. chelonae*. Since its discovery *M. abscessus* complex have become important clinical pathogens to human health. The *M*. *abscessus* complex is divided into 3 subspecies *M. abscessus* subsp. *abscessus*, *M. abscessus* subsp*. massiliense* and *M. abscessus* subsp*. bolletii*, with the difference being in the erm(41) gene patterns, which confers macrolide resistance. *M. abscessus* subsp. *abscessus* and *M. abscessus* subsp*. bolletii* both have full length erm(41) while *M. abscessus* subsp*. massiliense* has a truncated erm(41) gene with a 379 bp deletion (Brown-Elliott et al., 2015; Kim et al., 2010). Like other mycobacteria, *M. abscessus* complex is highly drug and disinfectant resistant, leading to it being a causative agent of post-surgical infections (Moreno-Izquierdo et al., 2020).

Like other NTMs *M. abscessus* has two morphologies; rough and smooth dependant on the abundance of glycopeptidolipids (GPL) on the cell wall, with smooth *M. abscessus* having GPLs and rough having an absence of GPLs either though transcriptional downregulation or mutation causing loss of function (Rüger et al., 2014). GPLs consist of 3 moieties a peptide, lipid and carbohydrate moieties. They can be either polar or non-polar dependent upon if the carbohydrate is triglycosylated or diglycosylated (Gutiérrez et al., 2018).

GPL is composed of a core peptide structure comprised of a 3-hydroxy or 3-methoxy C_{26}-C_{33} fatty acid chain N-linked to a long chain fatty acyl residue making up the lipid moiety (Daher et al., 2020, 2022; Gutiérrez et al., 2018). Finally, the carbohydrate domain 6-deoxytalosyl (dTal) unit linked to the allo-Thr residue and by an O-methylated rhamnosyl unit linked to the terminal alaninol residue (**Figure 1.1)** (Lopez-Marin et al., 1994).

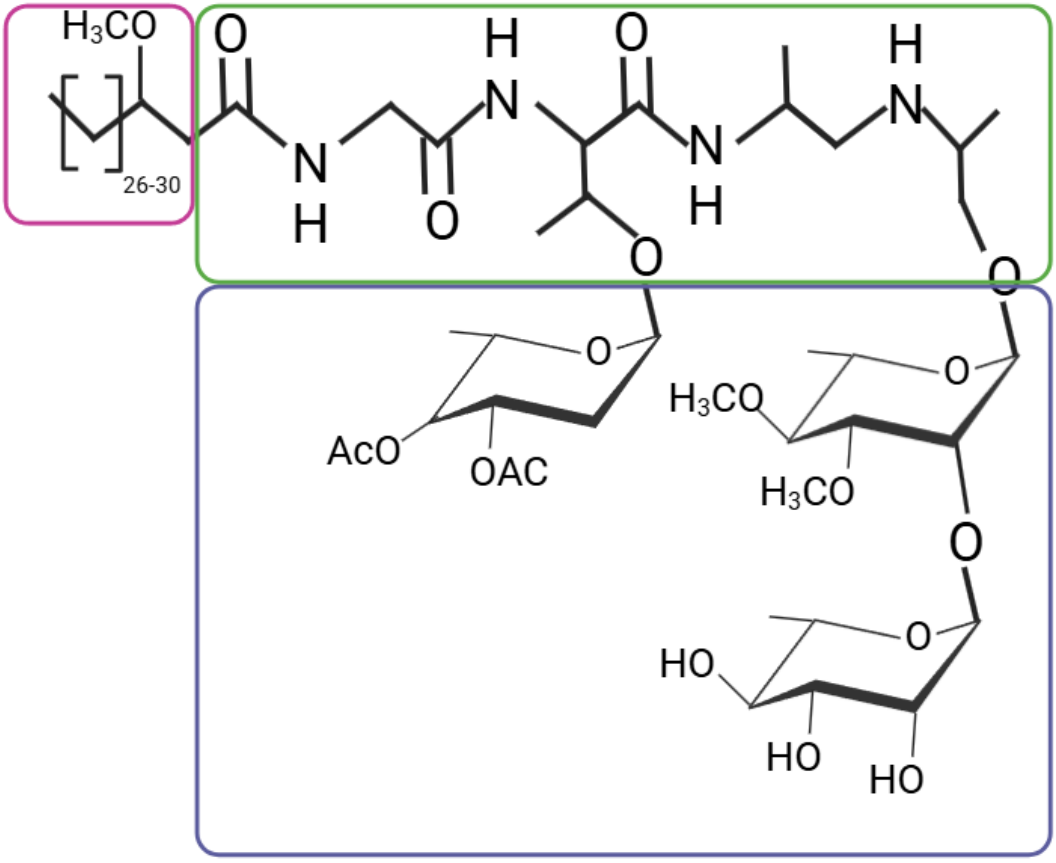


Figure 1.1 *M. abcessus* GPL. Schematic of *M. abcsessus* GPL with the lipid moiety in pink, fatty acid moity in green and carbohydrate moiety in blue.

Single smooth bacilli are taken up by macrophages, while the lack of Glycopeptidolipids (GLPs) in rough morphotypes allows aggregation of the bacilli as they remain adhered to the phagocytic cups on the macrophage periphery (Rüger et al., 2014).

These two morphologies also confer a difference in infection, with each causing distinct clinical outcomes. Smooth variants can survive for long durations within the phagosome (Kam et al., 2022). These cells are able to undergo irreversible genetic changes to rough morphology. The transition of the cells from smooth to rough causes rapid granuloma break down and massive cord-like structures causing severe and rapid loss of pulmonary function (Roux et al., 2016). There is still a large gap in knowledge surrounding the factors which cause this change. The infection process is reviewed thoroughly by (Johansen et al., 2020).

1.1.4 Nocardial diseases

The primary disease caused by *Nocardia* sp. is Nocardiosis, with the primaries species responsible being *N. asteroies*, which can affect the brain, lungs and skin. As stated before, it is most common among those with pre-existing conditions, although still rare, most commonly effecting the lungs however it can also affect the central nervous system (Mehta & Shamoo, 2020). The high prevalence among those who are already immunocompromised is due to immunity to Nocardiosis being mainly mediated by T-cells, therefore those suffering from AIDS and other immune diseases are more susceptible, due to their lower T-cell counts (Lafont et al., 2020b; Uttamchandani et al., 1994).

The clinical manifestations of Nocardiosis are subtle and nonspecific including, fever, weight loss, night sweats, cough, chest pains and pneumonia when infecting the lungs. If the infection has spread to the CNS the symptoms can also include headaches, weakness, confusion and seizures (CDC, 2016).

1.1.1.3 Diagnosis and treatment.

Both bacterial infections are difficult to diagnose both in part due to the duration of time it takes for positive identification as well as the nonspecific nature of the symptoms. This can lead to an increased likelihood of incorrect diagnosis.

In recent times there has been a push towards not just identifying the infection for treatment purposes but also the species, which can be difficult with this order due to their time-consuming diagnosis techniques and overlapping symptoms. This move towards higher level of specificity in diagnosis has been aided in part by the commercialisation and increase in availability of whole genome sequencing (Köser et al., 2014; Purushothaman et al., 2022).

Due to the non-specific symptoms, laboratory diagnosis for Nocardiosis is essential. The diagnosis is often carried out via the collection of bronchioalveolar lavage fluids, sputum samples or tissue biopsies (Duggal & Chugh, 2020). Diagnosis based on direct sampling is done via PCR of *Nocardia* specific genes, this however this can be of low specificity with Rouzaud et al. (2018) finding a sensitivity of 88% and specificity of 74% for this PCR. Due to the slow growing nature of *Nocardia* involving 14 to 21 days of culturing is needed before biochemical can be undertook. These techniques are often superseded by the results of PCR due to the time scale (Lafont et al., 2020).

The treatment of Nocardiosis is via antibiotic therapy, which is dictated by the site of infection as well as the species, there are 5 main classes of antibiotics that are prescribed for Nocardiosis: carbapenems, cotrimoxazole, linezolid, amikacin and cephalosporins. The treatment regime is for a minimum of 6 month with one-month extensions based on subsequent testing (Lafont et al., 2020; Margalit et al., 2021).

Being a well-documented human pathogen, in most countries, Tb has a set treatment regime and plan, with the two first line anti-tuberculous drugs being Rifampicin or Isoniazid which are available globally. Rifampicin acts by attaching itself to RNA polymerase and therefore inhibiting the elongation of RNA. Isoniazid inhibits cell wall synthesis when activated by the bacterial catalase-peroxidase enzyme KatG (Khawbung et al., 2021). These treatments can last for 4-9 months based on the regimes, however for multi-drug resistant tuberculosis (MDR-TB), treatments such as bedaquiline and

fluoroquinolones are required, which can be as long as 20 months with varying levels of antibiotics based on the resistance suite of the strain.

Conversely, there is little global coordination on the diagnosis and treatment of NTM diseases. What limited coordination that there is, is often limited to countries with more resources (Van Ingen et al., 2018). The current UK guidelines for the diagnosis of PNTM are outline by the British Thoracic society **Figure 1.2** (Haworth et al., 2017a). However much of the diagnosis and treatment pathways are still up to the discretion of the doctor. Although the diagnosis of NTM diseases have improved via the use of molecular diagnostics including ribosomal 16S sequencing. There is still difficulty in the discrimination between closely related mycobacterium species due to its low taxonomical resolution (Rizal et al., 2020). Nontuberculous mycobacterial pulmonary disease (NTM-PD) must be treated in a variety of ways, dependent upon the causative agent. Mycobacterium avium complex pulmonary disease (MAC-PD) is often treated with azalide, clarithromycin, ethambutol, rifampicin.

To understand the full burden of cost which is implicated in the treatment of NTM much more research is needed. Many existing studies are limit to individual countries or regions with large amounts of the world lacking data and even those with data often lacking standardised methods that can be implemented on a larger scale. Even with the lack of global coverage, it is clear that the cost of treatment is high, with Goring et al. (2018) reporting the costs per person per year for patients of £9,300, £15,264, £10,434 and £9,727 in Canada, France, Germany and the UK, respectively, in 2015. Few comparative studies of the kind exist therefore it is difficult to obtain a current cost per patient.

Although improvements in overall surveillance have occurred, it is still not fully understood as to the true extent of the burden of NTMs around the world. Future increased surveillance and tracking would be the ideal method however, this would involve large investment in countries which NTM disease are not a priority and without the funding to launch such large-scale surveillance.

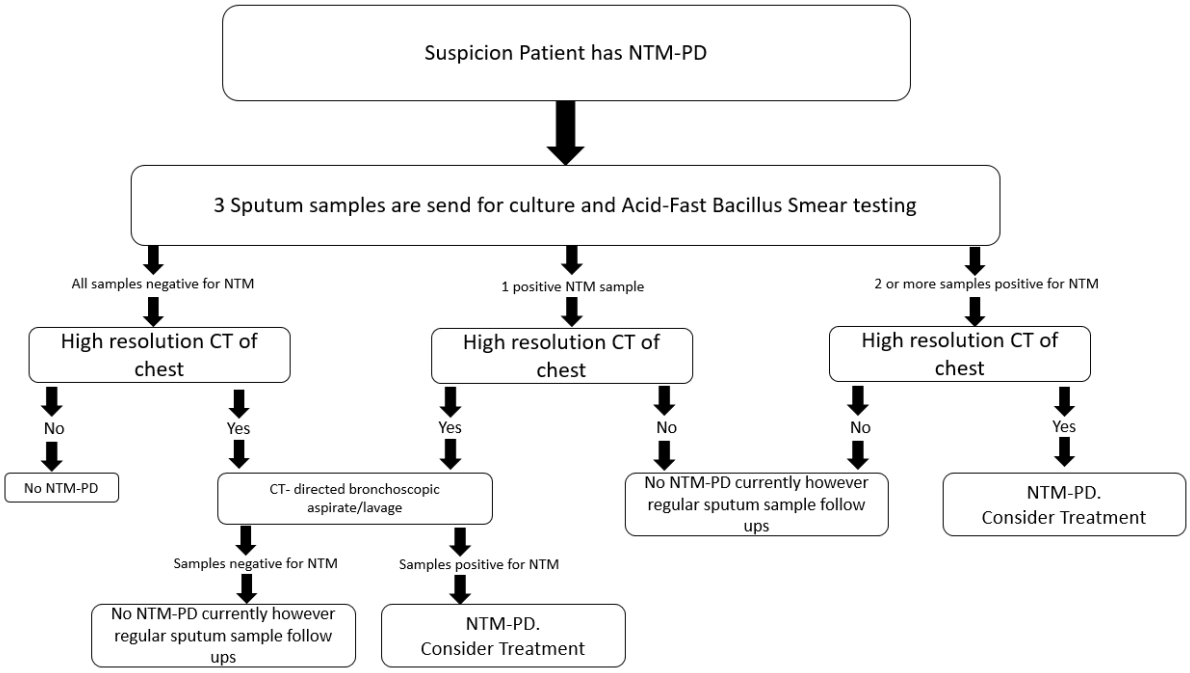


Figure 2 Flow chart of UK diagnosis of NTM. British Thoracic societies guidelines on the diagnosis and treatment of NTM infections adapted from Haworth et al. (2017).

1.3 Whole Genome sequencing.

As WGS sequencing has become more accessible and cheaper it is being proposed as the new gold standard for diagnosis, widely considered the future, especially in cancer and genetic research (Schwarze et al., 2019). The technology has potential for not just more accurate disease tracking, but also to improve treatments and allow for predictions of potentially medically important mutations (Comín et al., 2022; Ford et al., 2011; Leshchiner et al., 2012; Su et al., 2019).

1.3.1 Whole Genome sequencing mycobacteria

In the case of mycobacteria, the future is not just a more rapid identification at a species level but also the accurate identification of single nucleotide polymorphisms (SNPs) and prediction of mutations (Hall et al., 2023). As previously described, mycobacterial diseases have common symptoms and morphologies, however their repertoire of antimicrobial resistance can vary a lot leading to incorrect treatment and increased resistance down the line (Maiga et al., 2012). WGS allows a level of diagnosis that was not possible with previous molecular techniques. Although not a sole replacement for classical diagnosis it can be used in conjunction to provide the most optimal treatment for the patient and increased tracking accuracy.

The current most common method of WGS used by health services is Illumina, which is short read based, and though the individual base calling accuracy of Illumina is higher than that of Oxford Nanopore there are some disadvantages; these include the need for higher sample size for Illumina as well as lack of portability (Smith et al., 2021). Additionally in the case of mycobacteria, the high GC content (65%) and repetitive nature of the genomes can be an issue for sequencing technologies such as Illumina which due to their short reads have GC or AT rich bias meaning areas rich in GC or AT are under-sampled (Bainomugisa et al., 2018; Di Marco et al., 2023).

The long read sequencing platforms such as Oxford Nanopore (Oxford NanoPore Technologies) and PacBio single-molecule real-time (SMRT) sequencing (Pacific Biosciences Menlo Park, CA) allows bypassing some of the hurdles in genome sequencing mycobacteria such as the high GC content and repetitive nature of the genome. Although the functionality of Nanopore sequencing has been well documented for use in the detection of Mtb (Dippenaar et al., 2022; Hall et al., 2023; Meehan et al., 2019) much less has been done regarding its use for detection of NTMs outside of more research-based settings. In recent times it has been shown to have the potential to be used for the rapid identification of mycobacteria due to the ability to use the technology without the need for culturing (Hendrix et al., 2023; Xing et al., 2022), although many of these studies are limited, it shows progress and the potential for this technology to be used in the field of mycobacterial diagnostics.

1.4 Mycobacterial Cell wall

One of the key intrinsic resistance mechanisms and reason for difficulty extracting DNA is their highly impermeable multi-layered cell wall; a multilayer barrier composed of hydrophilic compounds and lipids **Figure 1.3**.

The mycobacterial cell wall plays a pivotal role in the innate resistance to antimicrobials, biofilm formation and modulation of host immune responses. The cell wall is comprised of 3 main regions together called the mycolyl-arabinogalactan-peptidoglycan (mAGP) complex these are: peptidoglycan, arabinogalactan and mycolic acids (Alderwick et al., 2015).

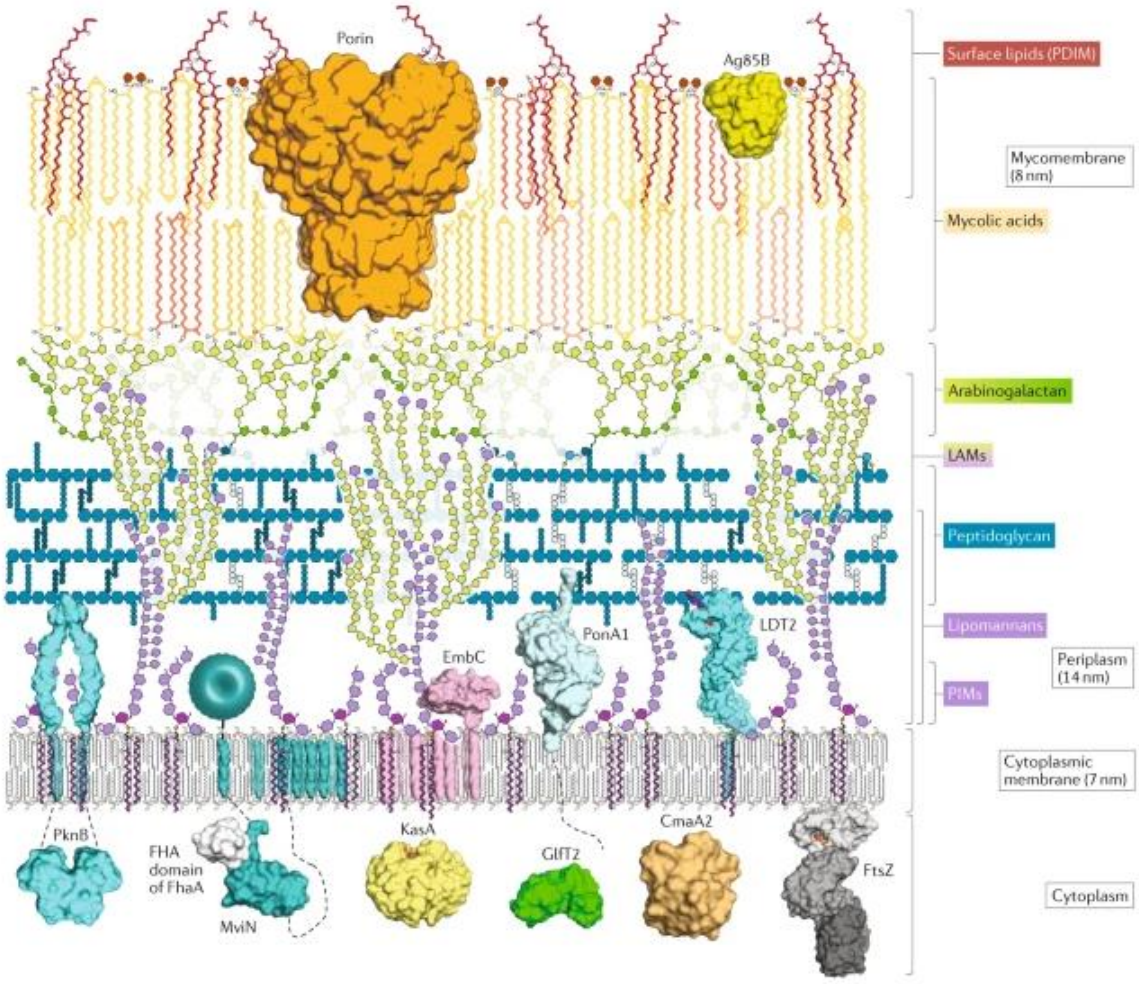


Figure 1.3 Mycobacterial cell wall. Representation of the various layers of the mycobacterial cell wall from the cytoplasm to the surface lipids. Peptidoglycan (blue) made up of *N*-acetylglucosamine and modified muramic acid is covalently linked to an arabinogalactan domain esterified with mycolic acids (Adapted from Dulberger et al., 2020).

1.4.1 Peptidoglycan

Peptidoglycan (PG) is ubiquitous among bacteria. Existing outside of the plasma membrane it is a mesh-like structure made up of glycan strands crosslinked by short peptides, that provides support and rigidity to the cell (Vollmer et al., 2008), allowing the cell to withstand osmotic stress while maintaining the integrity of the cell. Without PG the cell would lyse (Garde et al., 2021).

Mycobacterial PG (**Figure 1.4**) is formed of short peptide and glycan strands which are alternating *N*-acetylglucosamine and modified muramic acid linked via glycosidic bonds in a β(1-4) configuration. The modifications to muramic acids are either N-glycolyl and N-acetyl modifications (Raymond et al., 2005). Tetrapeptide side chains consisting of L-alaninyl-D-*iso*gluteminyl-*meso*-diaminopimelyl-D-alanine cross-linked with identical short peptides of neighbouring glycan chains. These cross-links include the expected 3-4 *meso*-diaminopimelic acid (Dap) and D-Ala bond that is common to most prokaryotes, as well as 3-3 *m*Dap-*m*Dap (Alderwick et al., 2015).

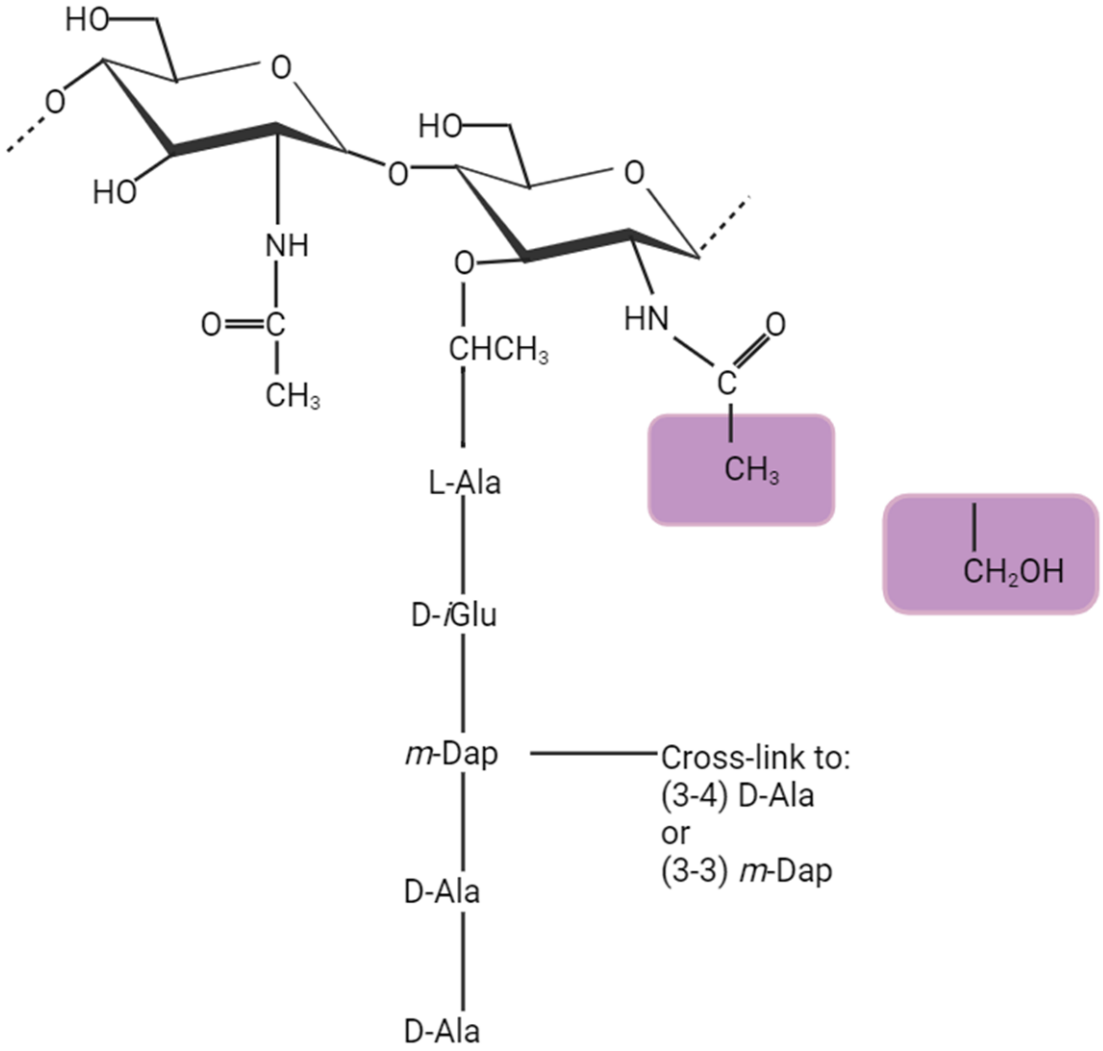


Figure 1.4 Structure of mycobacterial PG. Representation of the structure of mycobacterial PG with the repeating *N*-GlcNac 1-4 liked to modified Muramic acid connected to the lipid tetrapeptide side chain of L-alaninyl-D-isogluteminyl-meso-diaminopimelyl-D-alanine. Oxidation to the MurNac forming MurNgly is shown in the purple box. Muramic acid can be present in both the N-acetyl and N-glycolyl derivatives are both present in mycobacterial PG.

1.4.2 Arabinogalactan

AG is not an uncommon polysaccharide found in plant cell walls in its L configuration. Within plants it can be associated with proteins; arabinogalactan proteins, which are highly diverse and play a variety of roles within the plant cell (Ellis et al., 2010; Silva et al., 2020). Microbial AG is much rarer having been exclusively described in actinobacteria and is a major constituent of the cell wall of acid-fast bacteria (Alderwick et al., 2015). A key difference between these two forms of AG is that arabinofuranose is in plants it is present in its L isomer (L-Ara*f*) while microbial AG is its D isomer (D-Ara*f*).

Connected to the PG domain of the cell wall via a phosphodiester bond, AG makes up 30% of the total mass of the cell envelope, an essential component of the mycobacterial cell wall. Arabinogalactan is made up of D-galactofuranosyl and D-arabinofuranosyl residues **Figure 1.5**. These glycans are arranged with a Gal*f* back bone consisting of approximately 30 alternating β(1-5) and β(1-6) Gal*f* residues (McNeil et al., 1987). Approximately 23 Ara*f* residues are affixed to the c-5 hydroxyl of the 8^{th}, 10^{th} and 12^{th} Gal*f* residues, creating 3 branch points. The arabinan domain is highly branched; consisting of an α-1,5 back bone, with α-1,3 branch points. Attached to this are further α(1-5) linked Ara*f* residues. Branches of the arabinose domain are capped with a β-1,2 Ara*f* residue (Daffe et al., 1990).

Interestingly much is known about the synthesis of AG by mycobacteria, however very little is known about its breakdown and recycling. The synthesis of AG has been studied in depth and involves the transfer of GlcNAC-1-phosphate form UDP-GlcNAc to prenyl phosphate followed by an addition of rhamnose from dTDP-Rha forming the linker unit between Ag and PG by Mikušová et al. (1996). UDPGal-*f* is the donor for the Gal*f* units that make up the backbone. EmbA and EmbB are responsible for catalysing the linkage of the (1,5)Ara*f.* α(1,3)Ara*f* branch points are attached via AftC and AftD (Birch et al., 2008; Škovierová et al., 2009). The final β-Ara*f* cap on the arabinan domain is attached via AftB (Seidel et al., 2007). The arabinan branches are attached to the Gal*f* domain by AftA (Batt et al., 2020; Seidel et al., 2007).

Evidence as to the modes of recycling for both the arabinose and galactose domains has been lacking although Shen et al. (2020) have identified GlfH1 (Rv3096) which was shown

to exhibit exo-β-D-galactofuranose hydrolase activity cleaving the β-(1,5) and β-(1,6)-Gal*f* linkages, though very slowly. Very little is still known about breakdown of this domain however this displays the first steps towards the understanding of AG recycling.

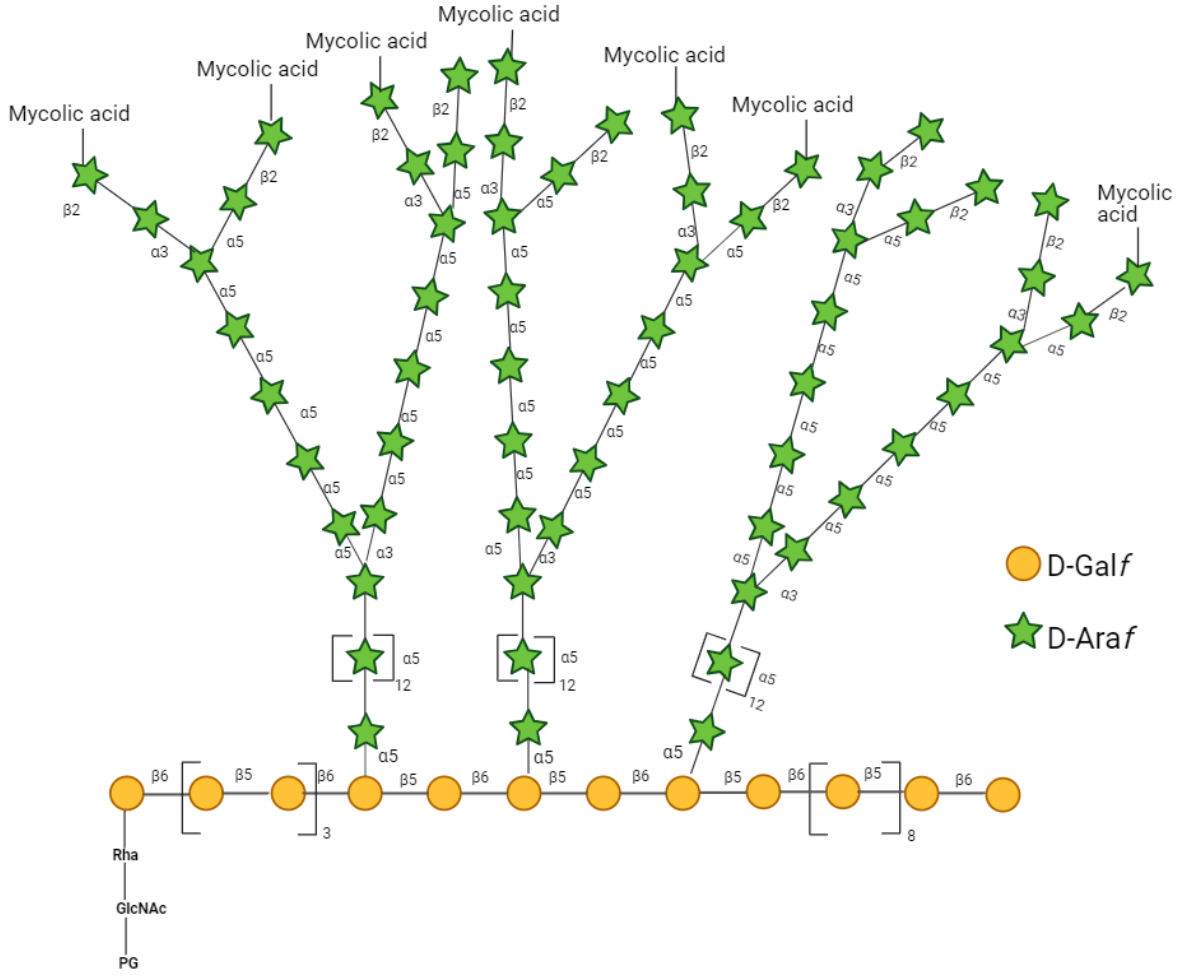


Figure 5 Schematic diagram representing the Structure of mycobacterial AG. Schematic representation of the AG domain of mycobacterial arabinogalactan, the galactan domain consists of approx. 30 repeating β(1-6)(1-5)Gal*f*. With α-1,5Ara*f* branch points at the 8, 10, 12 Gal*f* residues. The arabinan domains consists of α(1-5) chains with α(1,3)Araf branch points. The arabinose domains is capped by a β(1,2)Ara*f*. One third of the terminal arabinose moieties are esterified by mycolic acids.

1.4.3 Mycolic acids

Mycolic acids (**Figure 1.6)** are found ubiquitously on the cell envelope of mycobacteria and give mycobacteria their characteristic hydrophobicity. They are predominantly found covalently bound with approximately one third of the terminal β(1,2) Ara*f* residues. The essential nature of mycolic acids in mycobacterium makes them and their synthesis pathway a common target for antimicrobials such as isoniazid (Nataraj et al., 2015). Mycolic acids are made up of α-alkyl-β-hydroxyl long chain fatty acids which have an akyl side chain and a hydroxyl group making up 40-60% of the dry weight of the cell wall (Brennan & Nikaido, 1995; Korf et al., 2005). Present across much of the Mycobacteriales order, the length of the fatty acid changed considerable depending on genus and species, 22-38 carbons long in *Corynebacterium* to up to 100 in *Segniliparus.* With mycobacterium MA having 60-90 and *Nocardia* 46-60 (Burkovski, 2013; Marrakchi et al., 2014).

One of the keyways in which mycolic acids are involved in infection is by manipulation of the host immune system, by inhibiting the host immune response which include the cessation of phagosome-lysosome fusion and causing tissue damage, by the production of high levels of proinflammatory cytokines (Barkan et al., 2012; Korf et al., 2005; Marrakchi et al., 2014). As well as infection, they play an important role in the formation of biofilms which enable in the survival of mycobacterial cells when exposed to disinfectant chemicals in water ways and increased drug resistance in the body (Dokic et al., 2021; Ojha et al., 2008).

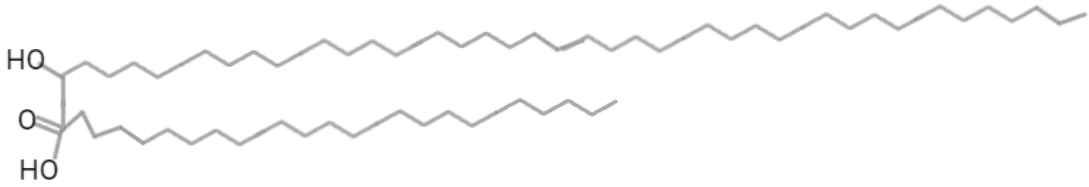


Figure 1.6 Schematic of mycolic acids. Schematic structure of mycobacterial mycolic acids.

1.4.4 Lipoglycans

Phosphatidylinositol mannosides (PIM) Lipomannans (LM), lipoarabinomannans (LAM) and mannose capped LAM (ManLAM) are all glycoconjugates present in the plasma membrane, non-covalently linked to inner and outer membrane of the cell wall via the Phosphatidylinositol (PI) unit (Abrahams & Besra, 2018).

These lipopolysaccharides are located in the plasma membrane, the presence of these is growth stage dependant. PIMs which were first described by Ballou et al. (1963), exclusively found in actinomycetes, are precursors to LAM and LM. PIMs are glycolipids comprised 3 domains; the phosph-*myo*-inositols attached to 1-6 mannose residues, PIM_2 and PIM_6 being the most abundant in the cell wall (Sancho-Vaello et al., 2017).

Both LM and LAM are derivatives of PIMs and share a common LM back bone (**Figure 1.7**). LAM is comprised of three distinct domains: (i) the PI linker domain, (ii) a mannan core which is an α(1-6)-mannan chain approximately 21-34 residues in length decorated with 5-10 units of α(1-2)man*p*, these two domains make up LM, and (iii) in addition to this, LM backbone LAM is further decorated with 55-70 Ara*f* (Mishra et al., 2011). The arabinan domain consists of α(1-5)-linked arabinosyl residues with some α(1-3)-branching, similar to the AG structure. The arabinan domain is capped by a β(1-2)-Ara unit. Additionally, a mannose cap can be present, forming ManLAM although not present across all mycobacterial species it is present in all Mtb complex strains as well as other pathogenic mycobacteria, such as *M. avium* and *M. marinum* among others (Turner & Torrelles, 2018).

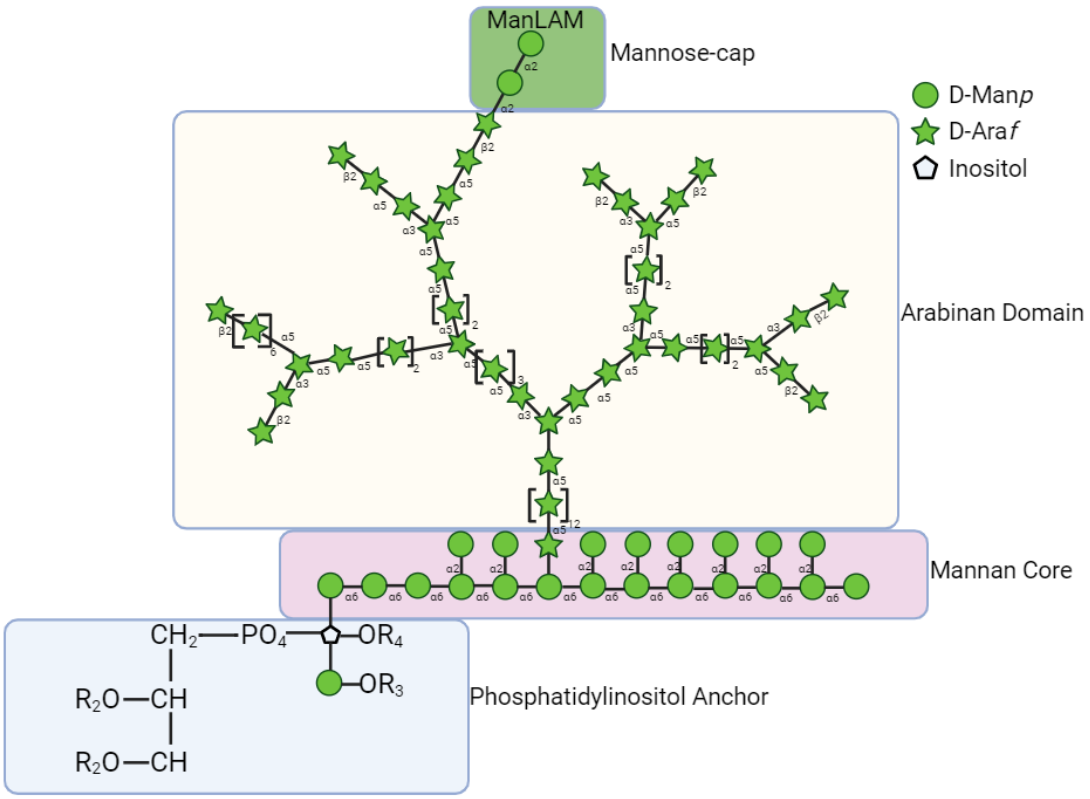


Figure 1.7 Structure of LM, LAM and ManLAM. Structure of LM, LAM and ManLAM including the separate domains highlighted in different boxes. LM consist of a PI domain (blue box) covalently bound to a mannan core (pink box). Arabinose domains can be found at branch points along the mannose core which form LAM these are α(1-5)Ara*f* with α(1-3)Ara*f* branch points. The arabinan domain is capped with a β(1-2)Ara*f* unit. There can be an additional mannose cap (green box).

1.5 Glycoside hydrolases

Carbohydrate active enzyme (CAZymes) are broadly categorised into five classes; these are: glycoside hydrolases (GH), glycosyltransferases, polysaccharide lyases, carbohydrate esterases, carbohydrate-binding modules (Davies & Sinnott, 2008)

1.5.1 Classification

First proposed by Henrissat in 1991, GHs were classified into families based on sequence similarity (Henrissat, 1991). Glycoside hydrolases are enzymes that catalyse the hydrolysis of glycosidic linkages. There are several methods of classification that can be used for

these such as endo or exo-acting, enzyme commission number, mechanistic classification and sequence-based classification (Henrissat & Davies, 1997). These are made into families, at of the time of writing there are 188 GH. These families can be further divided in subfamilies.

1.5.2 Catalytic mechanism

There are two main catalytic mechanisms. Inverting and retaining examples are shown in **Figure 1.8**; this is dependent upon the stereochemistry of the anomeric carbon after hydrolysis. In inverting mechanisms, the configuration of the chiral carbon is inverted (**Figure 1.8a**). On the other hand, in a retaining mechanism the anomeric carbon is retained in the same conformation (**Figure 1.8b**).

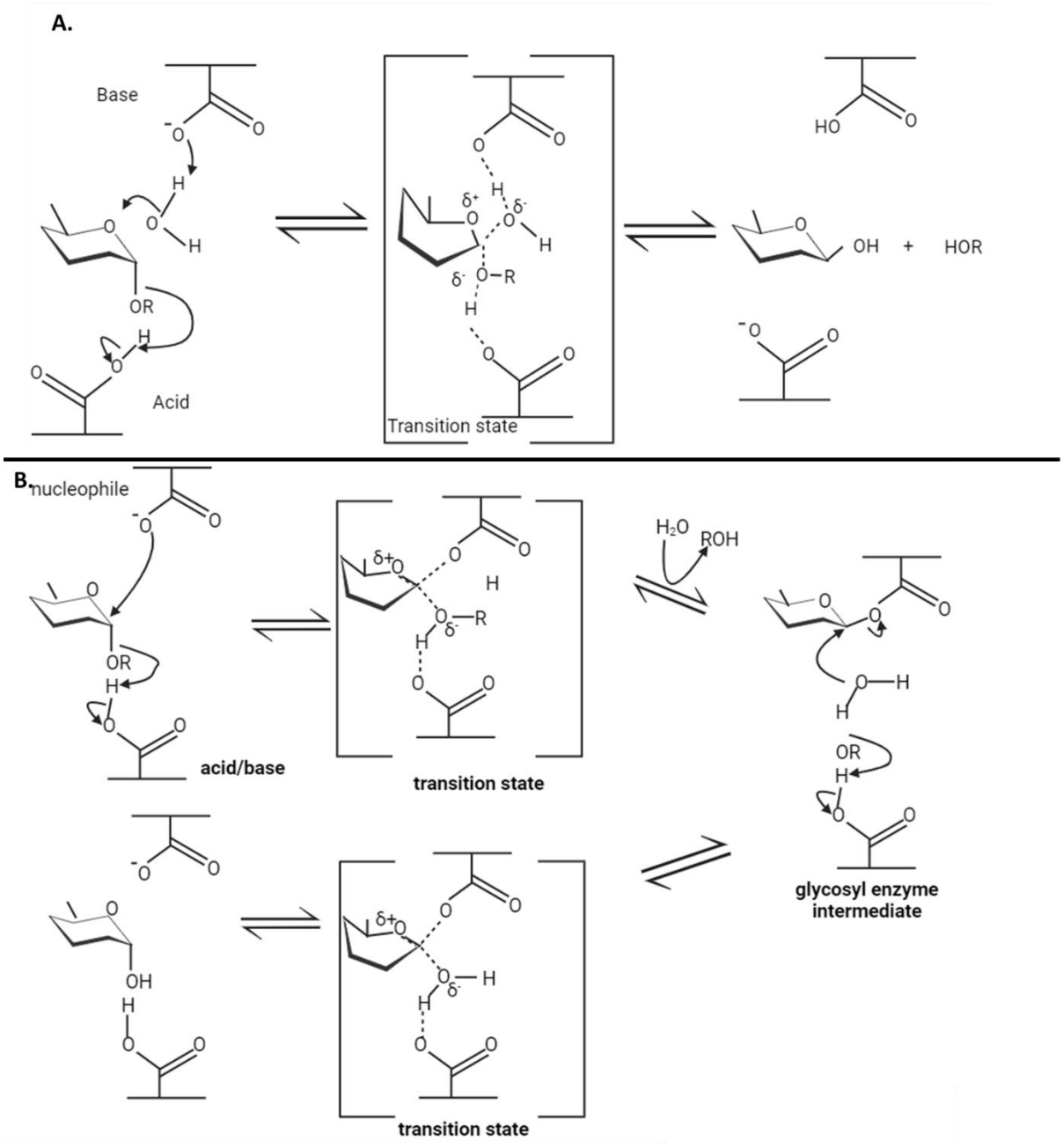


Figure 1.8 Inverting and retaining glycosidic mechanisms. a. Schematic diagram of the mode of hydrolysis of an inverting α-glycosidase, involving a general acid and a general base with the inversion of the chiral carbon. **b.** Schematic diagram of the mode of hydrolysis of a retaining α-glycosidase.

1.5.3 Glycoside hydrolase active site topology

GHs can be endo or exo-acting depending upon the point of cleavage along the glycan chain. Endo enzymes generally cleave the middle of a carbohydrate chain and produce a range of different sized oligosaccharides. Exo-acting enzymes generally recognise and cleave the non-reducing end of a polysaccharide chain and release monosaccharides. This activity can be conferred via the topology of the active site.

There are three main types of GH active site topology **Figure 1.9** these are: Pocket topology; optimal for the configuration for the recognition of monosaccharides **Figure 1.9a.** Open cleft, the open structure of this allows for the binding of internal regions of polysaccharides and is commonly found in endo-acting polysaccharides **Figure 1.9b**. Finally, Tunnel topology this is a characteristic of exo-acting enzymes with the polysaccharides moving through the tunnel with monosaccharides being released (G. Davies & Henrissat, 1995).

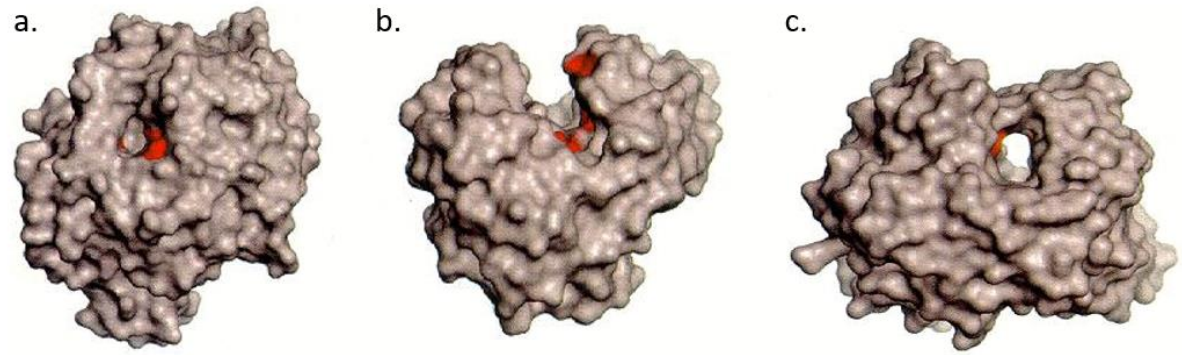


Figure 1.9 Three types of glycoside hydrolase topology. a. Pocket topology, **b.** Open cleft, **c.** tunnel topology *(G. Davies & Henrissat, 1995)*.

1.5.4 Sub-site nomenclature.

The substrate binding regions of GHs are divided into sub-sites which are characterised by the amino acids that are interacting with a single sugar shown in **Figure 1.10**. An enzyme sub-site is located at the scissile bond with sub-sites being numbered from –n (non-reducing end) to +n (reducing end) where –1 represents the active site residue and between -1 and +1 is the location of the scissile bond (Davies et al., 1997). The number of sub-sites and the topology varies from enzyme to enzyme. Understanding and characterisation of sub-sites is essential in the contribution of knowledge towards how enzymes can degrade complex branched glycans.

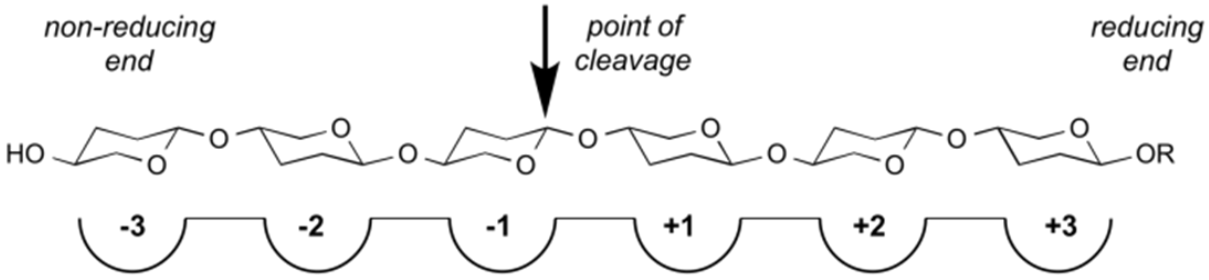


Figure 1.10 Schematic of subsite nomenclature. Subsite nomenclature from the point of cleavage towards the non-reducing end being indicated by '- values' sites decreasing from point of cleavage. Towards the reducing end indicated by '+ values' increasing from the point of cleavage (Davies et al., 1997).

1.5.5 Polysaccharide utilisation loci

PULs are clusters of co-regulated genes that work together to sense and co-ordinate the degradation of a glycan (Bjursell et al., 2006). Characteristic of the Bacteroidota phylum, PULs allow bacteria to grow on a wide variety of glycans which enables survival in niches such as the human gut, which has a high proportion of complex glycans due to the human diet (Grondin et al., 2017). A PUL has at least one sequential pair of SusC, and SusD (saccharide uptake system) homologues (Martens et al., 2009). SusC are members of the TonB-dependant receptor family, acting as a membrane spanning transporter while SusD are binding proteins. SusC and SusD interact to form a SusCD complex that facilitates the uptake of large nutrients (Gray et al., 2021). PULs are generally specialised towards one

glycan, however a PUL can also have activity towards more than one glycan if they are chemically similar (Rogowski et al., 2015).

1.5.6 Identification of AG degrading enzymes.

Previous work by Al-Jourani et al. (2023) identified several members of the *Bacteroidales* order that were able to grow on the AG purified from *M. smegmatis* Mc2 155 (**Figure 1.4**). When analysed via HPAEC-PAD it was shown that *Dysgonomonas gadei* was the only species to release arabinose monosaccharide during growth, and the majority of the others were only degrading the galactan component.

During the profiling of AG breakdown to identify D-arabinan degrading enzymes, the presence of arabinose and galactose in the supernatant made the identification of the PUL responsible for the degradation of the arabinan domain complicated, to aid in this they developed a method to obtain purified D-arabinan lacking the galactan domain. Two enzymes that degraded the galactan were characterised from *B. finegoldii,* present in PUL39 and PUL47 (**Figure 1.11**). These two enzymes were β-D-galactofuranosidase GH43_31 (BACFIN_08810) and a new GH family GH182 (BACFIN_04787). When used in combination these two enzymes completely hydrolyse the galactan domain of arabinogalactan to galactose.

With the purified D-arabinan which was used as the sole carbon source for *D. gadei,* proteomics identified a PUL containing several proteins of unknown function. The enzymes belonging to the DUF2961 and DUF4185 superfamilies were prioritized as they have homology to theoretical proteins present in mycobacteria (**Figure 1.11**).

The enzymes containing DUF2961 from PUL 42 of *D. gadei* were characterized as a new GH family, GH172, which was shown to have α−D-arabinofuranosidase activity, this family was subdivided into: $DgGH_{172a}$, $DgGH_{172b}$ and $DgGH_{172c}$ which have belonging to them the genes: HMPREF9455_02467, HMPREF9455_02471 and HMPREF9455_02479 respectively, which encode separate glyscodife hydrolyase proteinsThe other superfamily prioritised

from PUL42, DUF4185, were designated GH183 (HMPREF9455_02480 and HMPREF9455_02481), these enzymes were shown to have endo α−D-arabinanase activity. Together then, the enzymes can degrade the majority of the arabinan domain of arabinogalactan into arabinose.

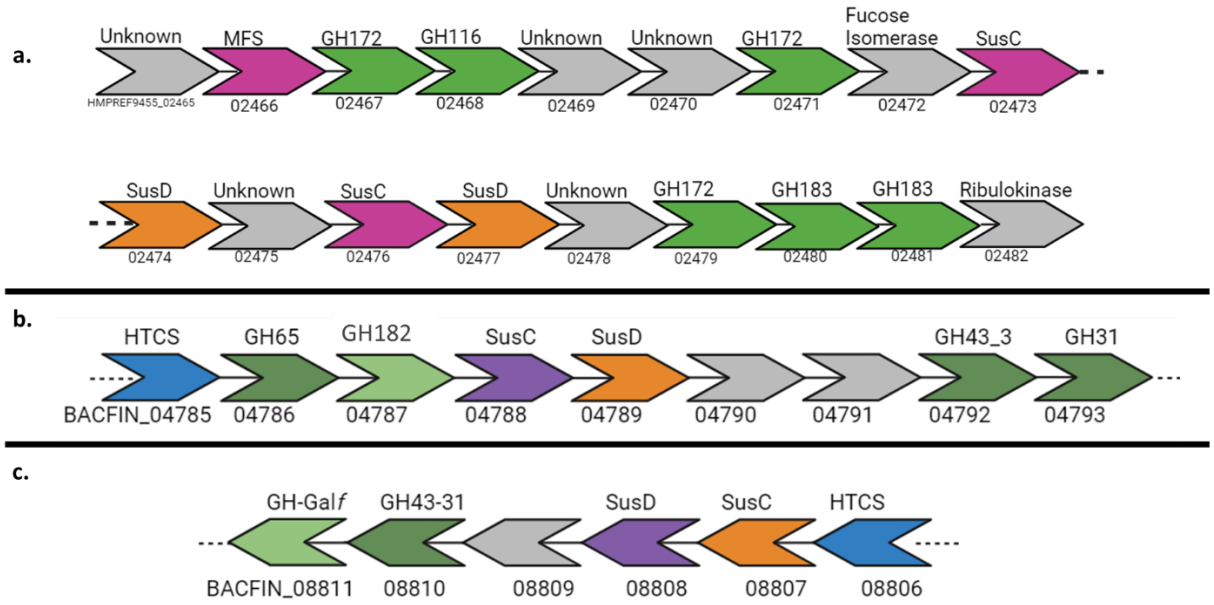


Figure 1.11 Schematic representation of PUL 42 of *D. gadei* and PUL47 and 37 of *B. finegoldii*. A. Schematic representation of PUL 42 of *D. gadei*. **B**. Schematic representation of PUL 47 of *B. finegoldii*. **C.** schematic representation of PUL 39 of *B. finegoldii.*

1.6 Aims and objectives.

Mycobacteria have a high resistance to lysozyme which has been documented (Gordon & Barnett, 1977; Kanetsuna, 1980) due to the *N*-glycolylation on the *N*-acetylmuramic acid (Raymond et al., 2005). Therefore, many current DNA extraction protocols, especially commercially available kits which rely on lysozyme for the lysis of the cells produce yield of DNA insufficient for downstream analysis. Many current lysis and DNA techniques involve harsh chemicals such as phenol-chloroform or mechanical lysis such as bead beating (Epperson & Strong, 2020). With the discovery of the recently described arabinanases described in **Section 1.6.6** and Gal*f* active enzymes identified from *B. finegoldii,* there are potential new tools for the degradation of mycobacterial cell wall to

extract nucleic acids. The major aim of this thesis therefore, is to investigate the possibility of developing an enzymatic lysis reagent, particularly for the extraction of gDNA for diagnostic sequencing. This will be explored through the following objectives:

1. Are there other new enzymes to be discovered in *D. gadei* PUL 42?
2. Can arabinogalactan degrading enzymes act as mycobacterial lysis reagent?
3. Can gDNA from enzymatic lysis be used to sequence whole mycobacterial genomes through long-read sequencing?

2. Chapter 2. Materials and methods

2.1 Molecular biology

2.1.1 Chemicals, commercial kits and water

All experiments were set up using 18.2 MΩ/cm H_2O produced by a Millipore Milli-RO 10 Plus Water Purification System. Water was used as a solvent unless stated otherwise. All the chemicals, enzymes and media used in molecular biology were obtained from Sigma Chemical Company, ThermoFisher or Melford unless stated otherwise.

2.1.2 Buffers

Buffers used are listed in Table 2.1.

Table 2.1 Buffers used in this study.

Buffer	Components	pH
Buffer A	20 mM HEPES 150 mM NaCl	7.5
Buffer B	20 mM HEPES 150 mM NaCl 10 mM imidazole	7.5
Buffer C	20 mM HEPES 150 mM NaCl 100 mM imidazole	7.5
Buffer D	20 mM Potassium Phosphate buffer	7.4

2.1.3 Bacterial strains and plasmids.

The different *Escherichia coli* and bacterial plasmids used during this study are listened in Table 2.2 and Table 2.3. Plasmid diagrams are shown in appendix **Figure1.**

Table 2.2 Bacterial strains used for cloning.

Strain	Genotype	Use	Reference
Tuner™ (DE3)	F- *omp*T *hsd*SB (r_B- m_B-) *gal dcm* *lac*Y1 (DE3)	Protein expression	Novagen™
One Shot™ TOP10	F- *mcrA* Δ(*mrr-hsd*RMS-*mcr*BC) Φ80*lac*ZΔM15 Δ *lac*X74 *rec*A1 *ara*D139 Δ(*araleu*)7697 *gal*U *gal*K *rps*L (StrR) *end*A1 *nup*G	DNA replication	Invitrogen™

Table 2.3 Plasmid vectors used for cloning.

Plasmid	Size (kb)	Features	Reference
pET21a	5.44	Amp^r, T7, *lac, laciq*	Novagen^{TM}
pET28b	5.37	kan^r, T7 *lac lacI^q*	Novagen^{TM}

2.1.4 Sterilisation

All solutions, growth media and glassware were sterilised through autoclaving either by the use of an Astell Hearson 200 series Autoclave or Prestige© Medical Series 2100 Clinical Autoclave at 121 °C or 32 lb/inch^{-2} for 20 minutes. Solutions for size-exclusion chromatography were filter-sterilised using sterile Millipore filter discs with a 0.22 µM pore size (Stupor® 3.2 Gelman Sciences) and a suitable sterile syringe (Plastipak®, Becton Dickinson).

2.1.5 Growth media and antibiotics

Media was created either as liquid broth or as solid plates which were made with the addition of 2% (w/v) Bacteriological agar (Ovoid^{TM}). Media used in this study is detailed in Table 2.4. For positive selection of transformant *E. coli* 100 µg/ml of ampicillin for pET21a and 50 µg/ml of kanamycin for pET28a was used. A stock of antibiotics was kept at 1000-fold higher than working concentration at -20 ^oC.

Table 2.2 Growth media. Growth media used through out the investigation for bacterial growths.

Media	Component	Amount per litre
Luria-Bertani Broth (LB)(Lennix)	LB powder (Sigma-Aldrich)	25 g
Brain-Heart infusion (BHI)	BHI powder (Sigma-Aldrich)	37.5 g

Tryptic Soy Broth (TSB)	TSB Powder (Sigma-Aldrich)	30 g
Middlebrook 7H9 Broth	Middlebrook 7H9 Broth Base	4.7 g
	OADC Enrichment	40 ml

2.1.6 Centrifugation

The centrifuge used was based on the volume of the solution. Volumes above 50 ml were centrifuged in a Beckman J2-21 centrifuge with a JA-10 rotor and 500 ml centrifuge tubes (Nalgene). Volumes between 25 and 50 ml were also centrifuged in the Beckman J2-21 however at 27,000 RCF for 30 minutes at 4 ^{o}C. Volumes ranging from 2 and 25 ml were centrifuged at 3,420 RCF in 25 ml universal (Sterilin) tubes in a Hettich Mikro 220R benchtop centrifuge with a fixed angle rotor.

Volumes under 2 ml were centrifuged using microcentrifuge (Eppendorf) tubes at 17,000 RCF using a Heraeus Instruments Biofuge pico benchtop centrifuge. For protein concentration, a Harrier 18/80R centrifuge with swing out rotor was used at 4000 RCF at 4 °C.

2.1.7 Storage of DNA and bacteria.

Short term storage of genomic DNA prior to analysis was at 4 ^{o}C for up to 2 weeks, long term storage of genomic DNA was at -80 ^{o}C. Stocks of DNA plasmid were store at -20 ^{o}C. Cryovials of bacterial strains as glycerol stocks (glycerol, 25% w/v) were stored at -80 ^{o}C.

2.1.8 Transformation of chemical competent E. coli

Competent *E. coli* were prepared by Mr. Carl Morland **(Appendix A)** and stored at -80 ^{o}C for up to one year. A 100 µl aliquot of cells was thawed and subsequently mixed with 1 µl (30 - 100 ng) of plasmid DNA and incubated on ice for 30 minutes. Following this, cells were heat-shocked at 42 °C for 90 seconds, then incubated on ice for a further 5 minutes. Cells were then mixed with 350 µl of LB media and incubated for 1 hour at 37 °C in a ThermoFisher MaxQ™ 8000 shaking incubator at 180 revolutions per minute (RPM). The 450 µl of cells were then plated out onto LB plates containing the appropriate antibiotic for the vector. Plates were then incubated in an inverted position at 37 °C for 16 hours using a static LEEC thermoregulated incubator.

2.1.9 Growth conditions for bacterial growth for DNA propagation

Stocks of plasmid DNA were made by transformation of a recombinant plasmid into TOP10. From transformation plates a single colony was picked and inoculated into 5 ml of LB and grown for 16 h at 37 ^{o}C, 180 RPM.

2.1.10 Plasmid DNA isolation

Plasmid DNA was purified from 5 ml cultures using the QiAspin Miniprep Kits (QIAGEN).

2.1.11 Determination of DNA concentration.

Determination of concentration of plasmid DNA was done using NanoDrop 2000 UV-Vis spectrophotometer (Thermo Fisher Scientific). Genomic DNA was quantified using qPCR (**Section 2.1.12**) or Qubit 4 Fluorometer (ThermoFisher) suing Qubit^{TM} 1x dsDNA HS Assay Kit.

2.1.12 Polymerase Chain Reaction (PCR)

Q5® High-Fidelity DNA Polymerase (New England BioLabs) was used throughout this study with the standard reaction being shown in appendix **Table 1**.

Table 2.5 Conditions and cycle of a standard PCR.

Description	Temperature (°C)	Time	Number of cycles
Initial denaturation	95	2 minutes	1
Denaturation	95	30 seconds	35*
Annealing	55	1 min	
Elongation	68	1 min/kb fragment size	
Final elongation	68	10 minutes	1
Storage/hold	4	∞	∞

* For site-directed mutagenesis, PCR was performed with only 16 cycles

2.1.13 Quantitative PCR (qPCR)

Quantification of DNA was done using Roche Lightcycler 96 using Luna® Universal qPCR Master Mix (NEB). 10 µl reactions were set up as shown in Appendix **Table 3** and programme settings shown in **Table 2.6**. Results were then analysed using LightCycler® 96 (Roche) software. The primers used are shown in Table 2.7.

Table 2.6 qPCR programme settings.

Name	Ramp (°C/s)	Target (°C)	Duration (s)	Cycles
Preincubation	4.4	90	600	1
3 Step Amplification	4.4	95	10	40
	2.2	60	10	
	4.4	72	10	
Preincubation	4.4	72	600	1
High Resolution Melting	4.4	95	60	1

Table 2.7 qPCR primers used for *M. smegmatis* and *M. abscessus* subsp. abscessus DNA quantification.

Species (target)	Primers	
***M. smegamtis* (Msmeg_2793)**	Forward	GTG ATG ATC AAC CCG TTC CG
	Reverse	CCT TGG TGC GTT CCT GTT C

M. abscessus* subsp. *abscessus	Forward	GCC AGC TAC GAA GCA AAA C
	Reverse	ACC GG ACA CAT CCA GAA AC

2.1.14 Site Directed Mutagenesis (SDM).

Mutagenesis of single amino acids was carried out by site-directed mutagenesis (SDM). SDM utlilizes an appropriate double stranded plasmid and two synthetic oligonucleotides flanked by 10-15 nucleotides that fully complement the DNA template. The PCR reactions were made up at shown in **Table 2.3** and PCR conditions are shown in **Table 2.4**.

2.1.15 Agarose gel electrophoresis

Genomic DNA and PCR products were run on horizontal agarose gel according to Table 2.10.

Table 2.8 Agarose gel conditions. Agarose gel condition for DNA visualisation based on target DNA base pair range.

Bp range	% W/V agarose	Voltage (V)	Time (min)
>15000	0.6	40	90
500- 6000	0.8	70	50
500-100	2	50	120

2.1.16 Sequencing of recombinant plasmids.

Sequencing of recombinant plasmids was done using Eurofins Tubeseq service. Roughly 100 ng of DNA sent for sequencing using the standard sequencing primers T7 (TAATACGACTCACTATAGGG) and T7term (GCTAGTTATTGCTCAGCGG). Sequences were checked using Clustal OMEGA.

2.1.17 Overexpression of recombinant proteins in E. coli cells.

All recombinant protein expression was done in *E. coli* TUNER cells. After transformations the plates were incubated at 37 ^{o}C overnight. The following day plates were scraped with 5 ml of LB and inoculated into 1 L of LB containing the appropriate antibiotic. These were then grown at 37 ^{o}C until an OD_{600} 0.6 was reached, then 1 ml 0.2M isopropylthio-β-D-galactoside (IPTG) for a final concentration of 0.2 mM was added. The flasks are then incubated at 16 ^{o}C overnight.

2.1.18 Sonication

Sonication of suspended cell pellets was done using a B. Braun Labsonic U sonicator set at 45 watts 0.5 second cycling for 2 minutes.

2.1.19 Immobilised metal affinity chromatography (IMAC)

His-tagged proteins were purified from cell-free extract using IMAC. Purifications were carried out using Buffer A and cobalt-charged TALON^{TM} resin (Takara). A 2 cm^3 bed of resin was made and equilibrated with buffer A. Lysate was loaded on to the column through a filter. Non-specifically bound molecules were washed off with Buffer A. Buffer B and Buffer C were then added sequentially to elute the bound protein.

2.1.20 Sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE).

The eluted fractions from IMAC were ran on 12.5% SDS-PAGE gels made as shown in Table 2.11 at 35 mA per gel. Gels were developed using InstantBlue^{TM} stain (Expedeon) and viewed using a gel documentation system (Bio-Rad Gel Doc 1000, Molecular Analyst^{TM} PC Windows Software).

Table 2.9 Preparation of 12.5% SDS-PAGE gels.

Component	Reagent	Volume per gel
Resolving gel	0.75 M Tris-HCl, pH 8.8, 0.2% SDS	3.2 ml
	40% Acrylamide*	1.9 ml
	H_2O	1.1 ml
	10% (w/v) ammonium persulfate	30 µl
	TEMED	10 µl
Stacking gel	0.75 M Tris-HCl, pH 6.8, 0.2% SDS	1.25 ml
	40% Acrylamide*	0.25 ml
	H_2O	1 ml
	10% (w/v) ammonium persulfate	20 µl
	TEMED	7 µl

2.1.21 Protein quantification

Concentration of protein in samples was measured routinely, using a NanoDrop 2000 UV-Vis spectrophotometer (Thermo Fisher Scientific) and utilising the Beer-Lambert equation below:

$$A = \varepsilon C l$$

A = Absorbance at 280 nm

ε = Extinction coefficient (M^{-1} cm^{-1})

C = Concentration (Molar)

l = Light path (cm)

The ProtParam program (www.expasy.com) was used to calculate molar extinction coefficient of protein (Wilkins et al., 1999).

2.1.22 Concentration and buffer exchange

Fractions containing protein were pooled and placed into dialysis tubing (Medicell, 14 kDa molecular weight cut-off). 1 L of Buffer A was used as dialysis buffer per 10 ml of protein elution. Dialysis was left for 16 h at 4 ^{o}C. Proteins were concentrated in a Vivaspin 20 centrifugal concentrator with a 10kD molecular weight cut-off (SARTORIUS).

2.1.23 Size-Exclusion Chromatography

Concentrated samples were loaded onto a size exclusion chromatography column (HiLoadTM 16/600 SuperdexTM 200 Prep grade, GE Healthcare) which was equilibrated with Buffer A. Fractions containing the protein of interest were identified by SDS-PAGE and then pooled and concentrated as above.

2.1.24 Biomass prepation prior to genomic DNA preparation

All bacteria were cultured in appropriate media and in appropriate conditions until an OD_{600} of 0.7 was reached. Wet weights of bacterial pellets were calculated by weighing the 1.5 ml Eppendorf's adding 700 µl of culture. Then using a benchtop centrifuge, these were spun down for 3 minutes at 13000 rcf and the supernatant was discarded. The Eppendorf's were then reweighed to get the wet weight. All data was normalised to 50 mg unless stated otherwise.

2.1.24.1 Enzymatic methods for genomic DNA preparation

Proteins were prepared the previous day (**Section 2.1.17**) and stored at 4 ^{o}C. Using a spin concentrator, each enzyme was concentrated to 100 µM. The bacterial pellets were re-suspended in 200 µl Gram positive Lysis Buffer (45 mg ml^{-1} Lysozyme, 10 mg ml^{-1} Lipase). 20 µl of each 100 µM protein was added and vortexed to ensure homologous mixture. The reaction was then incubated in a thermal mixer (Eppendorf ThermoMixer® C) preheated to 37^{o}C for 60-120 minutes 2000RPM. Following the incubation, the GenElute™ Bacterial Genomic DNA Kit protocol was followed. Any variations are detailed in the results chapter 4.

2.1.24.2 Bead Beating

Bead beating was performed using a Bead Genie (Scientific Industries, Inc.) using 2 ml Micro tubes (Sarstedt) with 0.5 g of 0.1 mm silica disruption beads, at 4000 oscillations per minute, for 45 seconds. The samples were then stored on ice for 60 seconds before a second bead beat. The tubes were then centrifuged at 11000 x g for 3 minutes. The supernatant was then carefully moved into a new Eppendorf as to not disturb the beads. Following this the GenElute™ Bacterial Genomic DNA Kit protocol was followed.

2.1.25 Oxford Nanopore sequencing

Sequencing was all performed using GridION^{TM} (Oxford Nanopore Technologies) with a Flow Cell (R9.4.1).

Rapid sequencing kits using Kit 9 chemistry, transposome-based sequencing was used in this study. 9 µl of 50 ng of DNA was made up per sample. 1 µl of barcode was added per sample. The samples were then incubated at 30 °C for two minutes followed by 80 °C for 2 minutes. The samples were then placed on ice for 1 minute. This was to ensure the barcodes had bound to the DNA.

All of the samples were then pooled into a 1.5 ml Eppendorf. An equal measurement of AMPure SPRI Beads were added to the pooled samples. The samples were then incubated

at room temperate for 5 minutes the Eppendorf placed on a magnetic rack and left for 5 minutes for the beads to move towards the magnet. The supernatant was removed and the remaining beads were washed twice with750 µl 80% ethanol.

The DNA was then eluted from the beads using 10 µl of 10 mM Tris pH 8. The supernatant was moved to another Eppendorf and 1 µl of Rapid Adapter F (RAP-F) was added. The library was then prepared as shown in **Table 2.12**. Flow cell priming followed the Rapid Sequencing protocol (SQK-RBK110.96)(**Appendix A**).

Table 2.10 Settings for GridION sequencing run.

Setting	
Duration	24 hours
Voltage	-180 V
Base calling	Base calling and barcoding on
Protocol	96 Rapid Barcode it SQK-RBK110.96

2.2 Microbiology

Table 2.11 Growth conditions and times for microbial cultures. Growth conditions for all bacterial strains tested all bacterial cultures were grown at 37 ^{o}C.

Bacteria	Media	Incubation time (days)
M. smegmatis Mc2 155	TSB	2
M. abscessus subsp *abscessus* ATCC 19977	TSB	3
M. marinum ATCC927	TSB	2
M. avium subsp. *paratuberculosis* NCTC 8578	TSB	2
M. bovis BCG NCTC 5692	BHI	14
T. paurometabola DSM 20162	BHI	1
N. farcinica NCTC11134	BHI	1

2.3 Biochemical Methods

2.3.1 Enzymatic assays

All enzymatic assays were carried out at an enzyme concentration of 1 µM and substrate concentration of 2 mg/ml for 16 h at 37 ^{o}C unless stated otherwise. With data being analysed in GraphPad Prism 9.5.0.

2.3.2 Thin Layer Chromatography (TLC)

All TLCs were performed on TLC silica gel 60 W (Supelco). 6 µl sample were loaded onto the TLCs and dried. The running buffer used for all TLCs was comprised of butan-1-ol: acetic acid: water in a 2:1:1 ratio. Each TLC was run twice, once to 1 cm from the top it was then removed, dried and placed in and ran to the top. TLCs were stained using orcinol-sulfuric acid stain (sulphuric acid/ethanol/water 3:70:20 v/v, orcinol 1%) and developed at 100 ^{o}C.

2.3.3 High-performance anion exchanged chromatography with Pulsed Amperometric Detection (HPAEC-PAD).

Oligosaccharides from enzymatic polysaccharide digestion were analyzed using a CARBOPAC PA-300 anion exchange column (ThermoFisher) on an ICS-6000 system Assays analysed were performed in phosphate-citrate buffer pH 7.4. Detection enabled by PAD using a gold working electrode and a PdH reference electrode with standard Carbo Quad waveform. Buffer A – 100 mM NaOH, Buffer Bch – 100 mM NaOH, 0.5 M Na Acetate. Each sample was run at a constant flow of 0.25 ml·min^{-1} for 100 min using the following program after injection: 0-10 min; isocratic 100% buffer A, 10-70 min; linear gradient to 60% buffer B, 70-80 min; 100% buffer B. The column was then washed with 10 mins of 500 mM NaOH, then 10 min re-equilibration in 100% buffer A. L-arabino-oligosaccharides (DP = 2–9) obtained commercially (Megazyme) were used as standards at a concentration of 50 µM. Data were processed using ChromeleonTM Chromatography Management System V.6.8. Final graphs were created using GraphPad Prism 8.0.1. Data was collected on Chromeleon^{TM} Chromatography Management system (ThermoFisher).

2.3.4 Lyophilisation

All samples which were flash frozen in liquid nitrogen and Lyophilised in a Christ Alpha 1-2 freeze dryer at -40 ^{o}C for 3 days.

2.3.5 Porous Graphitic Carbon Chromatography

Hypersep Hypercarb PGC 50 mg (ThermoScientific) were used for all PGC chromatography. The sample was loaded onto the PGC. 5 column volumes of water were then used to wash any unbound substrates. A dilution series of 2 ml saturated butanol from 1:32 butanol: water to 100% butanol was used to elute any bound substrate. The fractions were then analysed on TLC. Those containing substrate of interest were diluted 1:3 with water and lyophilised as described in **Section 2.2.4**.

2.4 X-ray Crystallography.

2.4.1 Protein crystallisation screens.

Protein crystallisation was performed by a vapour diffusion sitting-drop method using an automated MosquitoR nanodrop dispensing system (TTP Labtech). Crystal trays were set up using MRC 96 well crystallisation plates (Molecular Dimensions) using 100:100 and 200:100 nanolitre ratios of protein sample to crystallisation condition. Various commercially available screening kits were used in initial screens: Index (Hampton Research) and PACT, Structure, Index, JCSG+ (Molecular Dimensions). Crystallisation trays were created at Newcastle Structural Biology Laboratory and left equilibrating at 20 °C for 6 weeks. Plates were examined weekly for crystals using a Leica MZ-6 crystallisation microscope (Leica Microsystems).

2.5 Bioinformatics and analysis

2.5.1 Bioinformatics

Amino acid sequence searches were carried out using basic Local Alignment Search Tool (BLAST). Identification of evolutionary amino acid positions was performed using ConSurf (Ashkenazy et al., 2016).

2.5.2 ONT sequencing analysis.

2.5.2.1 Initial sequencing quality analysis.

Initial read quality was analysed and initial statistics were generated using NanoStat (https://github.com/wdecoster/nanostat) (De Coster et al., 2018). Mean read length, N50 and longest read.

2.5.2.2 Contig construction.

De novo assembly of the reads was done using FLYE V2.9.2 (https://github.com/fenderglass/Flye) (Kolmogorov et al., 2020) which constructs repeat graph to generate high quality assemblies from reads to form continuous genomic segments The genome was then predicted by traversing the graph so each of the unique edges appear once (**Appendix A**).

2.5.2.3 Genome comparison

Comparison to a reference Genome to analyse completeness and quality of assembly was done using QUAST (Quality Assessment Tool) (https://github.com/ablab/quast) (Gurevich et al., 2013) **(Appendix A)**.

2.5.2.4 Genome annotation and analysis.

Assembled contigs were input into Bacterial and Viral Bioinformatics resource Center (BV-BRC) PAThosystems Resource Integration Center (PATRIC). The Comprehensive Genome Analysis tool was used to analysis completeness and contamination (Olson et al., 2023) as well as Minihash for genome identification (Ondov et al., 2016) **(Appendix A**).

2.5.2.5 Phytogenic tree construction.

Assembled contigs were input into Type (strain) Genome Server (TYGS) which allows for web-based genome-based taxonomy (Meier-Kolthoff & Göker, 2019).

Chapter 3. Characterisation of DG02470

3.1 Introduction

D. gadei is a gram-negative facultative anaerobe, a member of the Bacteroidetes phylum which makes up a large proportion of the healthy human gut microbiota. Aiding in facilitation of degradation of complex glycans which are not able to be processed directly by humans, *D. gadei* possess a rich variety of carbohydrate active enzymes (Kaoutari et al., 2013; Martens et al., 2011; Zafar & Saier, 2021). Until recently no β-D-arabinofuranosidase had been described in the literature, although after the completion of this work, a GH116 identified from *Microbacterium arabinogalactanolyticum* JCM 9171 ($ExoMA2_{GH116}$) by Shimokawa et al. (2023) was shown to be active upon AG purified from *M. smegmatis* specifically characterised as an exo-β-D-arabinofuranosidase.

As described previously (**Section 1.6.6**), PUL 42 (**Figure 3.1**) of *D. gadei* provided novel insight into D-arabinan degrading enzymes, resulting in the characterisation of exo-α-D-arabinofuranosidases into family GH172 and α-D-endo-arabinofuranosidases GH183 (Al-Jourani et al., 2023). Together, these allow the complete breakdown of the α-linked arabinan domain of AG into arabinose, though an enzyme in the PUL responsible for removing the β-linked caps has not yet been identified. There are still several predicted proteins within PUL42 which have not been characterised. The aim of this chapter is to investigate the hypothetical protein DG02470 and determine if it is involved in degradation.

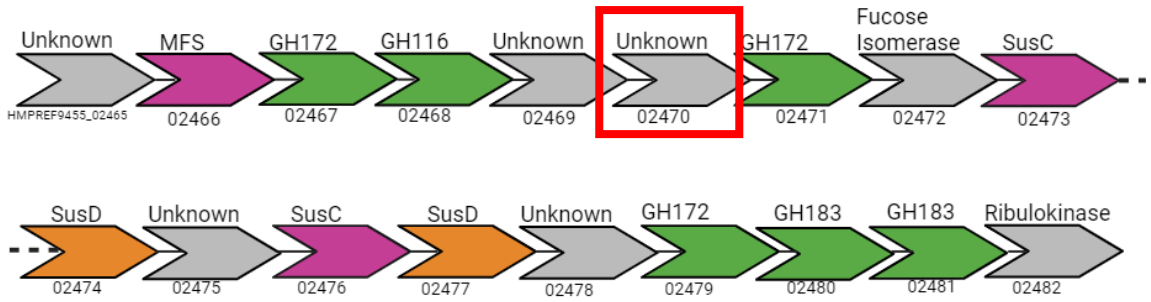


Figure 3.1 PUL42 of *D. gadei*. Schematic diagram of PUL42 of *D. gadei* with characterised proteins coloured green, SusCs and SusDs coloured magenta and orange respectively and uncharacterised proteins in grey. With DG02470 highlighted in the red box.

3.2 Results

3.2.1 Sequence Analysis

Using the SignalP 6.0 webserver (Teufel et al., 2022) it was determined that the protein was a periplasmic protein, although the start site was wrongly annotated. The annotated sequence has no clear periplasmic or lipoprotein signal peptides, but inspection of the upstream region of DNA allowed us to identify another start site 24bp upstream, which when translated gives a convincing periplasmic signal peptide (**Figure 3.2).** The gene encoding DG02470, without the signal sequence was cloned into pET28b, with a C-terminal His-tag, prior to the start of my PhD by Carl Morland.

```
M  L  V  I  S  F  L  S  L  F  S  S  L  L  F  P  V  R  V  L
ggc gaa gaa ctg ctg gaa att cgc aaa ctg ccg aaa gaa ttt cag gcg gtg att cat agc
 G  E  E  L  L  E  I  R  K  L  P  K  E  F  Q  A  V  I  H  S
att cgc ttt gat cgc ctg ggc agc att aaa ccg cag cag gaa att aaa aac att gat aaa
 I  R  F  D  R  L  G  S  I  K  P  Q  Q  E  I  K  N  I  D  K
ggc ctg gtg cat gtg cgc ctg att ttt gat ctg ccg gcg gaa att aaa cag gat gat tgg
 G  L  V  H  V  R  L  I  F  D  L  P  A  E  I  K  Q  D  D  W
cag ctg acc att cag ccg gcg ttt att ccg gat ttt aac tgg gcg ccg cat att acc ccg
 Q  L  T  I  Q  P  A  F  I  P  D  F  N  W  A  P  H  I  T  P
ggc gat atg aac att att gat cag cat gtg ttt cgc agc ccg gcg atg att gtg acc gat
 G  D  M  N  I  I  D  Q  H  V  F  R  S  P  A  M  I  V  T  D
aac aaa cag ggc ctg gcg gtg att ccg gat att aac att atg aaa aac ggc acc ccg acc
 N  K  Q  G  L  A  V  I  P  D  I  N  I  M  K  N  G  T  P  T
cgc tgg tat atg gat atg att gcg ccg gaa aac aaa ctg gtg ctg ggc atg agc gaa acc
 R  W  Y  M  D  M  I  A  P  E  N  K  L  V  L  G  M  S  E  T
ctg gtg agc ggc cat gtg att tat gaa cgc aaa gaa ggc gcg gtg tat ccg aaa ggc cgc
 L  V  S  G  H  V  I  Y  E  R  K  E  G  A  V  Y  P  K  G  R
gtg gaa ttt ggc ttt tat ctg atg aaa tat gcg gat aaa gaa gat ctg gaa aac ccg ttt
 V  E  F  G  F  Y  L  M  K  Y  A  D  K  E  D  L  E  N  P  F
cgc cgc ccg ctg aaa ttt ttt tgg acc aac tgg ggc gaa aaa ctg tat gaa gaa ggc aac
 R  R  P  L  K  F  F  W  T  N  W  G  E  K  L  Y  E  E  G  N
ccg att aaa ggc gat ctg gaa ccg tat gtg gat tat acc tat aac tgg gcg ttt aac acc
 P  I  K  G  D  L  E  P  Y  V  D  Y  T  Y  N  W  A  F  N  T
tgg aaa gat gcg gtg tgg cag gaa ttt atg ctg ggc gat aaa aaa gtg ggc gcg ccg gtg
 W  K  D  A  V  W  Q  E  F  M  L  G  D  K  K  V  G  A  P  V
ttt att gtg aac acc acc cag agc ccg aac tat ccg ggc aaa gtg aac gaa cgc gaa ttt
 F  I  V  N  T  T  Q  S  P  N  Y  P  G  K  V  N  E  R  E  F
ctg agc att tgg aac cag gcg tgg ttt agc agc ctg cgc agc gcg agc ggc ctg tat cgc
 L  S  I  W  N  Q  A  W  F  S  S  L  R  S  A  S  G  L  Y  R
tat gcg cgc cgc aaa ggc aac aaa ggc ctg ctg aac aaa gcg ctg ctg agc aaa gaa ctg
 Y  A  R  R  K  G  N  K  G  L  L  N  K  A  L  L  S  K  E  L
gcg ctg gcg gcg ccg atg aaa gaa ggc ttt ttt tat ggc ctg att ggc acc gaa atg gaa
 A  L  A  A  P  M  K  E  G  F  F  Y  G  L  I  G  T  E  M  E
acc gtg gaa acc gaa ggc aaa cag tat aac cgc agc aaa ggc tgg gat acc tat tat tgg
 T  V  E  T  E  G  K  Q  Y  N  R  S  K  G  W  D  T  Y  Y  W
ggc aac agc aac cgc aac ccg tat acc tgg gat gcg cgc aaa agc ccg ttt cat att ctg
 G  N  S  N  R  N  P  Y  T  W  D  A  R  K  S  P  F  H  I  L
gat atg agc tgg acc gcg att ctg atg ctg cgc tgg tat gaa gaa ctg gaa aaa gat gaa
 D  M  S  W  T  A  I  L  M  L  R  W  Y  E  E  L  E  K  D  E
cgc ctg ctg gaa tat gcg cgc aaa tat gcg gat gcg ctg ctg cag aaa cag gat gcg aaa
 R  L  L  E  Y  A  R  K  Y  A  D  A  L  L  Q  K  Q  D  A  K
ggc ttt ttt ccg ggc tgg ctg gat ctg gat acc ctg gaa ccg atg gaa tat ctg aac gat
 G  F  F  P  G  W  L  D  L  D  T  L  E  P  M  E  Y  L  N  D
agc ccg gaa acc agc atg agc gtg acc ttt ctg ctg aaa ctg tat gaa ctg acc aaa aac
 S  P  E  T  S  M  S  V  T  F  L  L  K  L  Y  E  L  T  K  N
gaa aaa tat cgc agc agc gcg ctg cgc gcg att gat gcg gtg tgc aaa gaa att gtg ccg
 E  K  Y  R  S  S  A  L  R  A  I  D  A  V  C  K  E  I  V  P
gtg ggc cgc tgg gaa gat ttt gaa acc tat tgg agc tgc tgc cgc ctg ggc acc ccg gaa
 V  G  R  W  E  D  F  E  T  Y  W  S  C  C  R  L  G  T  P  E
tgg att ggc aaa aaa att gaa cgc aac aac atg tat aaa cag tgc aac ttt agc att ttt
 W  I  G  K  K  I  E  R  N  N  M  Y  K  Q  C  N  F  S  I  F
tgg acc gcg gaa gcg ctg ctg gat agc tat cgc att acc aaa aac aaa gaa tat ctg aaa
 W  T  A  E  A  L  L  D  S  Y  R  I  T  K  N  K  E  Y  L  K
ctg ggc cag cgc acc ctg gat gaa atg ctg atg acc cag acc gtg tgg cag ccg ccg tat
 L  G  Q  R  T  L  D  E  M  L  M  T  Q  T  V  W  Q  P  P  Y
att tat gtg aac gcg gtg ggc ggc ttt ggc gtg atg aac gcg gat ggc gaa tgg aac gat
 I  Y  V  N  A  V  G  G  F  G  V  M  N  A  D  G  E  W  N  D
agc cgc cag agc ctg ttt gcg gaa acc att ctg cag tat ggc aaa ctg ctg aac aaa aaa
 S  R  Q  S  L  F  A  E  T  I  L  Q  Y  G  K  L  L  N  K  K
gaa tat aac gaa cgc ggc ctg agc gcg att cgc gcg agc ttt agc atg atg tat tgc ccg
 E  Y  N  E  R  G  L  S  A  I  R  A  S  F  S  M  M  Y  C  P
gaa aac ccg ctg gcg aaa gaa cag tgg gaa cgc gtg tgg ccg ttt ttt aac gaa aaa gat
 E  N  P  L  A  K  E  Q  W  E  R  V  W  P  F  F  N  E  K  D
tat ggc ttt acc atg gaa aac tat ggc cat gat ggc cgc acc agc aaa gat ggc att ggc
 Y  G  F  T  M  E  N  Y  G  H  D  G  R  T  S  K  D  G  I  G
att ggc gaa ttt acc att tat gat tgg ggc aac ggc gcg gcg gcg gaa gcg tat aac cgc
 I  G  E  F  T  I  Y  D  W  G  N  G  A  A  A  E  A  Y  N  R
att cgc gat aaa ttt ggc att gaa att ttt cgc
 I  R  D  K  F  G  I  E  I  F  R
```

Figure 3.2 Sequence of DG02470. With the Potential signal region in green.

3.2.2 Expression of 02470

The plasmid encoding DG02470 was transformed into TUNER cells, and protein production and purification was carried out as described in **Section 2.1.8**.

Fractions from IMAC purification were run on a 12.5% SDS-PAGE gel and visualised as shown in **Figure 3.3a.** The dominant band of protein is visible between the 66.2 kDa and 116 kDa ladder markers, which corresponds with the expected MW of DG02740 which is 79.7 kDa. From this gel it was determined that the 3 fractions eluted using 100 mM imidazole would be pooled and dialysed overnight at 4^{o}C into **Buffer A**. **Figure 3.3b** shows the size exclusion gel purification of DG02470, purified on a HiLoad® 16/600 Superdex® 200 pg (Cytivia) column on an AKTA Pure system used during crystallography trials much like **Figure 3.3a,** the band corresponds to the correct size of DG02470.

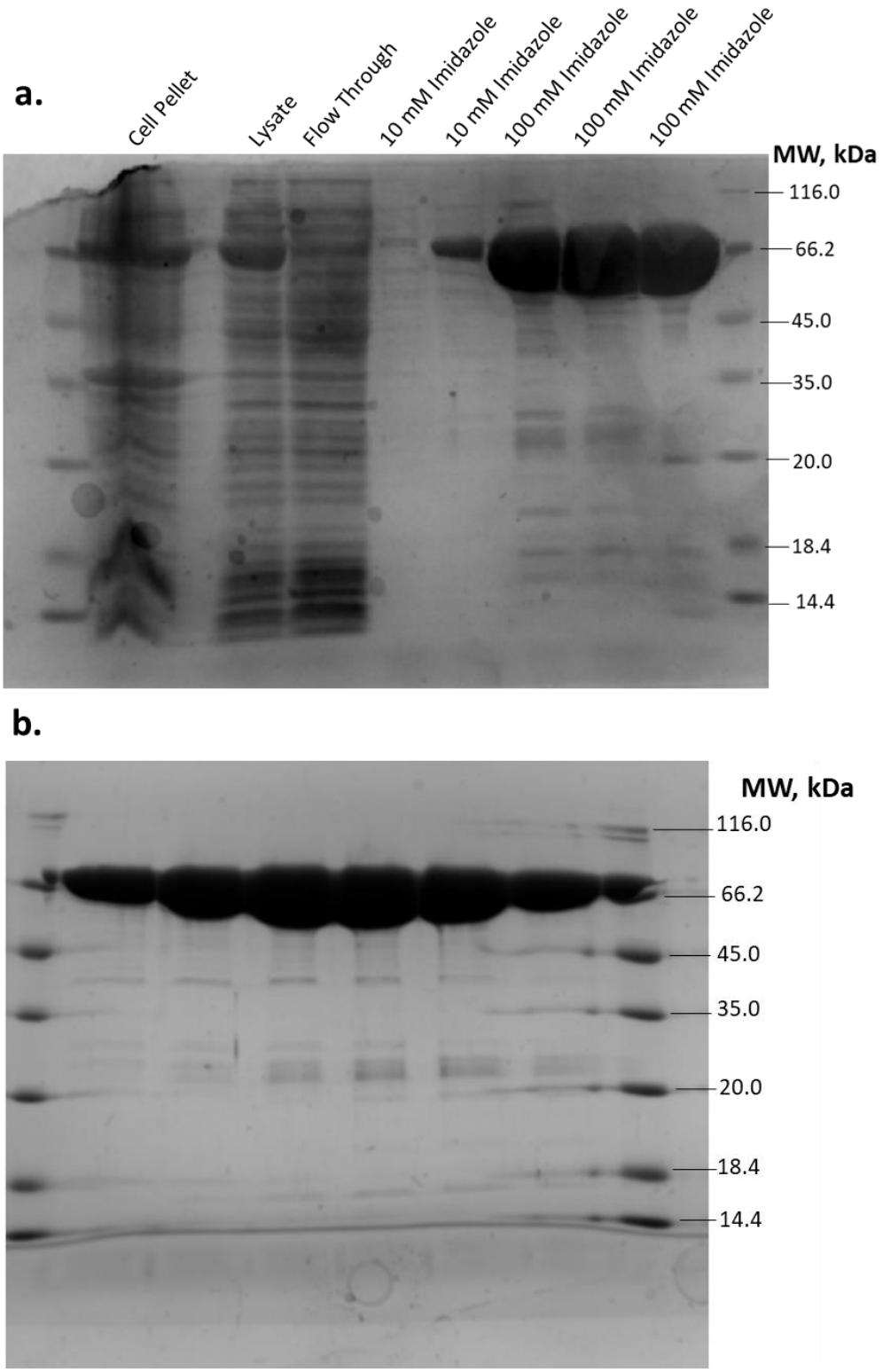


Figure 3.3 SDS of the of IMAC and gel column purification of DG02470. 12.5% SDS gel visualised for **A.** IMAC purification of DG02470 and **b.** fractions selected after gel filtration purification of DG02470

3.2.3 Enzymatic activity

3.2.3.1 Initial activity assay

To determine if DG02470 has any enzymatic activity, it was incubated with *M. smegmatis* arabinogalactan overnight, and the results analysed by TLC **(Figure 3.4)** and compared against an arabinose standard.

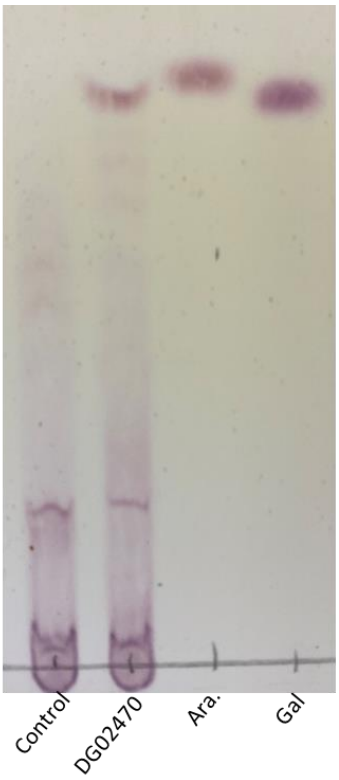


Figure 3.4 TLC analysis of DG02470 vs AG. TLC analysis of 1 µM DG02470 incubated overnight in **Buffer A** with 2 mg ml^{-1} AG stained with orcinol.

The results of the analysis show the production of arabinose monosaccharide but no oligosaccharide bands, suggesting that DG02470 is likely an exo-acting arabinofuranosidase.

3.2.3.2 HPAEC-PAD analysis.

To confirm that arabinose is the only product formed by DG02470, we used high-performance anion exchange chromatography with pulsed amperometric detection (HPAEC-PAD) which allows the quantification of arabinose produced as well as sensitive detection of any other oligosaccharides produced. All assays to be analysed by HPAEC-PAD were performed in 20 mM Phosphate citrate buffer pH 7.4**.**

The initial reaction analysed was 2 mg ml^{-1} was incubated for 120 minutes with 20 nM DG02470. The assay was analysed on HPEAC-PAD using a 100 mM NaOH with a 0-60% 0.5 M acetate gradient on a CARBOPAC PA-300 anion exchange column to fully analyse both monosaccharide and oligosaccharide production, with larger oligosaccharides having a longer retention time requiring a higher acetate concentration to be eluted from the column than smaller ones. The results were analysed and visualised on GraphPad Prism shown in **Figure 3.5**. **Figure 3.5a** confirms the results observed on TLC, that arabinose is released from AG by DG02470.

The oligosaccharide profile of AG did not obviously change after enzyme incubation, again confirming that DG02470 is likely an exo-D-arabinofuranosidase (**Figure 3.5b**).

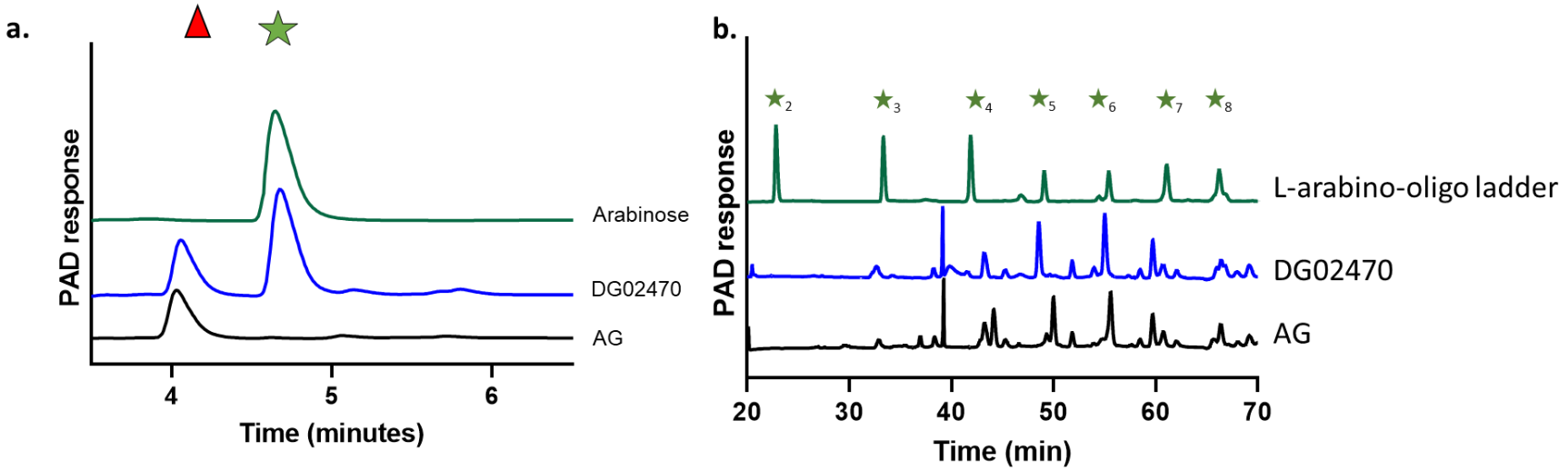


Figure 3.5 HPEAC-PAD chromographs of DG02470 vs AG. 20 nM Dg02470 incubated for 120 minutes with 2 mg ml^{-1} AG, run on HPAEC-PAD on a CARBOPAC PA-300 anion exchange column. **A.** monosaccharides analysis of products with internal standard fucose (red triangle) and standard L-arabinose (green star). Comparing the presence of arabinose after 120-minute incubation without DG02470 (Black) and with DG02470 (blue) **b.** oligosaccharide analysis of products compared to L-arabino-oligo ladder indicated by green stars indicating oligomerisation state.

3.2.3.3 DG02470 Kinetics.

We had determined that DG02470 is an exo-D-arabinofuranosidase, therefore we could use HPAEC-PAD to determine the kinetic rate of DG02470.

To test this a time point assay was performed with 20 nM enzyme at 1, 2, 5, 10, 15 and 30, minutes and analysed by HPAEC-PAD, using 50 µM fucose as an internal standard, and an arabinose standard curve to quantify product. The results are shown in **Figure 3.6a.**

At the substrate (AG) concentrations tested: 0.5, 0.75, 1.0, 1.5 and 2 mg ml^{-1} using a constant 20 nM concentration of DG02470 the rate is of arabinose release still increased linearly (**Figure 3.6b)**. Due to substrate availability, we were unable to increase the substrate concentration higher therefore we were not able to obtain individual values for K_{cat} and K_m. Therefore, we used K_{cat}/ K_m (**Figure 3.6c)**. Using the slope of the line in **Figure 3.6c** divided by the concentration of enzyme in µM (0.02 µM) the k_{cat}/k_m was calculated as 345 ±32 min^{-1} mg^{-1} ml.

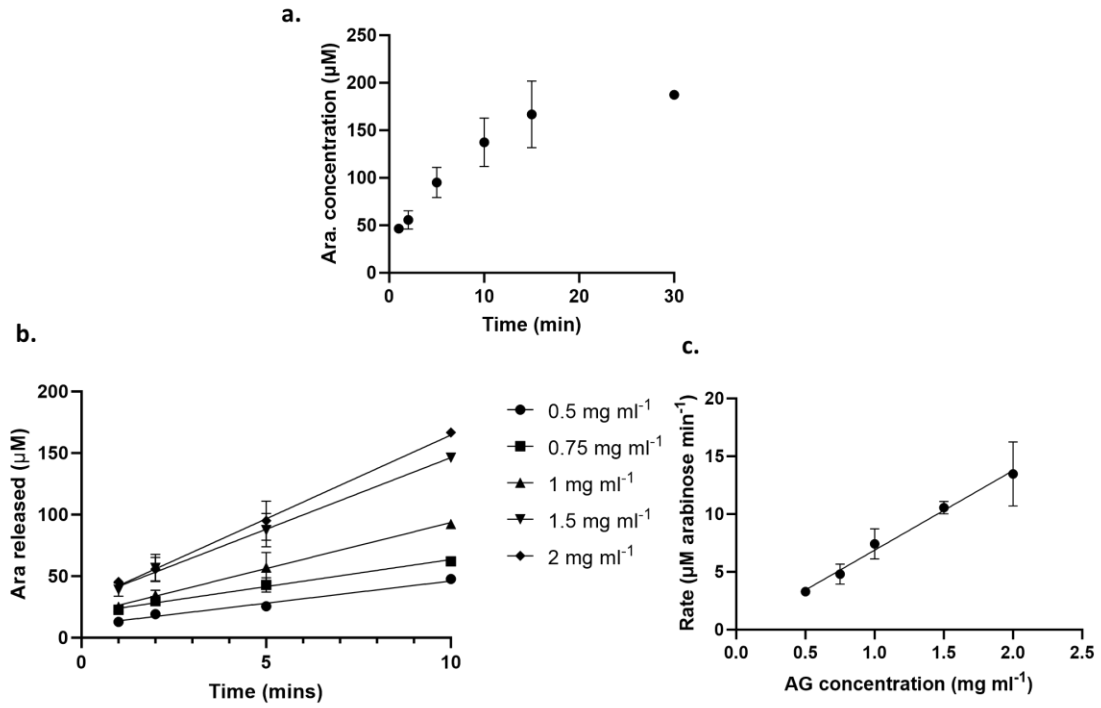


Figure 3.6 HPAEC-PAD analysis of 02470 vs AG. A. 20 nm DG02470 time point vs 2 mg ml^{-1} AG. **b.** Time point assay of arabinose release (µM) with 20 nM Dg02470 Time point assay vs 0.5, 0.75, 1, 1.5, 2 mg ml^{-1} **c.** rate (µM min^{-1}) of arabinose release by 20 nm 02470.

3.2.3.4 Sequential assays

To aid in the determination of bond specificity of DG02470 a sequential assay was set up with previously characterised α-D-arabinofuranosidase Dg_{GH172c} (Al-Jourani et al., 2023) followed by incubation with DG02470, this was also performed in the reverse order of enzyme sequence. Each assay was then analysed on HPAEC-PAD (**Figure 3.7**).

If the release of arabinose increases between incubation steps this indicates a variation in bond specificity between Dg_{GH172c} and Dg02470. The HPAEC analysis of arabinose released is shown in **Figure 3.7** in both cases the amount of arabinose released is increased. This indicates an alternate bond specificity to Dg_{GH172c} which is active on α-D-arabinofuranoside linkages.

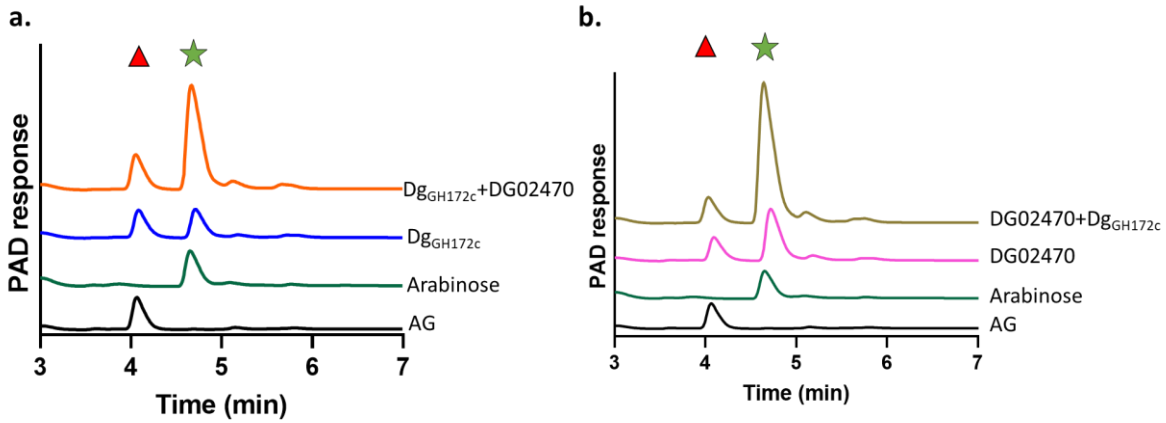


Figure 3.7 HPEAC-PAD chromatograph of sequential assay of AG vs Dg_{GH172c} and DG02470. A. 2 mg ml^{-1} AG was incubated with 100 nM GH_{172c} overnight (blue), the reaction was then stopped by heating, and 1 µM DG02470 was then added and incubated over night (orange). **B.** 2 mg ml^{-1} AG incubated with 100 nM DG02470 overnight (pink) boiled and then incubated with 100 nM Dg_{GH172c} (gold).

3.2.3.5 Activity specificity.

We had characterised the kinetic activity of DG02470 and identified that DG02470 was an enzyme with activity upon exo-D-Ara*f* with a different bond specificity to that of Dg_{GH172}.

We still did not have the bond specificity, to aid in this DG02470 was incubated with β-D-Ara*f*-pNP and α-D-Ara*f*-pNP synthesised by Prof. Spencer Williams (University of Melbourne). When hydrolysed these substrates turn yellow as p-nitrophenol is produced. No activity was detected on either substrate. AG is primarily composed of α-linked arabinan but many of the chains are capped with a β 1,2 arabinose.

In order to test whether these capping residues are the target substrate for DG02470 we made use of a strain of *Corynebacterium glutamicum* in which the arabinosyltransferase B, which catalysis the transfer of Ara*f* onto the arabinan domain forming the β(1,2)Ara*f* linkage (Seidel et al., 2007), has been deleted. This is possible because while the cell wall of *C. glutmanicum* shares the same triple layered structure of *M. smegmatis* including the same arabinogalactan domain, unlike *M. smegmatis* the cell wall of *C. glutamincan* is more tolerant to mutations than the *M. smegmatis* cell wall. With Raad et al. (2010) showing that with a deletion to the *aftb* gene in *C. glutamincan* cells, cells maintained viability, even if perturbed, without β-caps and mycolic acids. While this deletion in *M. smegmatis* leads to a complete loss in cell viability. The final analysis of bond specificity was done using AG purified from *C. glutamicum* with this *aftb* mutation produced and purified by the Moynihan Lab (Birmingham University). Additionally, WT AG purified from *C. glutamicum* (CG AG) was used to compare, in case of any minor differences between *M. smegmatis* and *C. glutamicum* AG.

1 µM Dg02470 was incubated over night with 2 mg ml^{-1} AG from *M. smegmatis* (AG), AG extracted from *C. glutamicum* (CG AG) and *C. glutamicum* Δ*aftb* AG (Δ*aftb* AG) the results are shown in **Figure 3.8.**

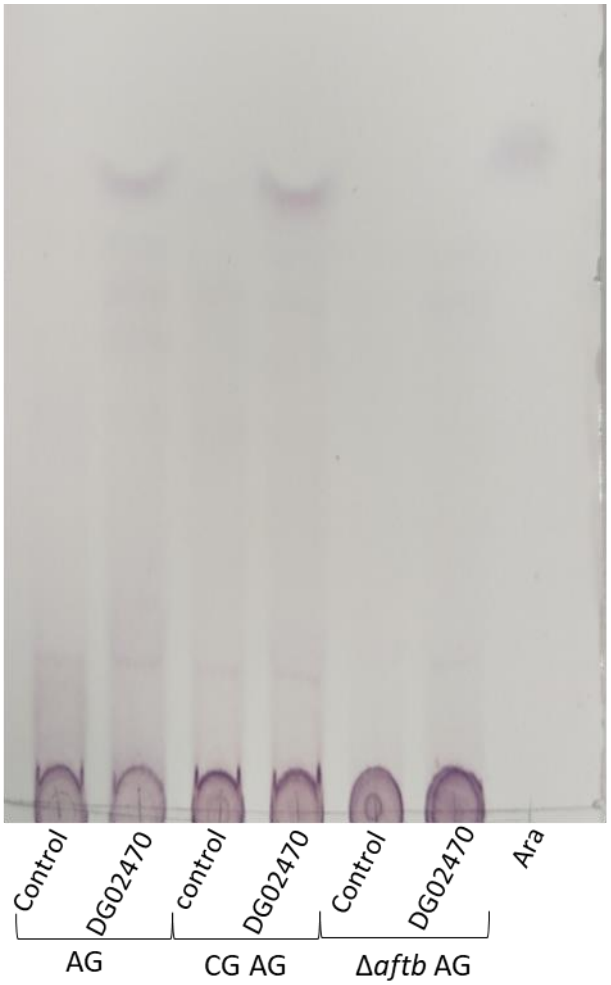


Figure 3.8 TLC analysis of DG02470 vs AG and Δaftb AG. TLC of overnight incubation of 1 µM DG02470 with 2 mg ml^{-1} AG, CG AG and Δaftb AG. Stained with orcinol.

The TLC analysis of the assays show that DG02470 lacks any observable arabinose production when incubated with Δ*aftb*. This suggests that the DG02470 is specific to the β(1,2)-D-ara*f* bonds which cap the arabinan domain. This is reinforced by the production of arabinose when DG02470 is incubated with CG AG, producing a comparable result to that of AG. This aids in the conclusion that it is the β bond upon which DG02470 is active.

3.2.3.6 LAM and Pilin activity.

It is however not only AG that contains arabinose, as previously described LAM in **Section 1.4.4**, present in mycobacterial cell walls and pilins produced by *P. aeruginosa* both contain an arabinan domain.

To determine if DG02470 is active on LAM the Moynihan lab (Birmingham university) kindly provided us with LAM purified from *M. smegmatis,* which also contains β(1,2) cap like that of AG. By incubating 2 mg ml^{-1} LAM and 1 µM DG02470 overnight and analysing on TLC (**Figure 3.9a**) we were able to identify that DG02470 was active on LAM, producing an arabinose band. 2 mg ml^{-1} pilin oligosaccharides which had been purified from *P. aeruginosa* which are composed of α-1,5-D-Ara*f* glycans (Kus et al., 2008) were incubated overnight with 1 µM DG02470. Additionally, $DgGH_{172c}$ and $Dg_{GH4185b}$ were used as a positive sugar control. The reactions were analysed on TLC **Figure 3.9b** when pilins were incubated with DG02470 no band is observable. This was as expected, due to the absence of β-bonds which we had previously hypothesised, based on **Section 3.2.2.3** results, as specificity for β-1,2-Ara*f* bonds. $DgGH_{172c}$ $Dg_{GH4185b}$ and showed arabinose production which is as previously described in greater detail by Al-Jourani et al., (2023).

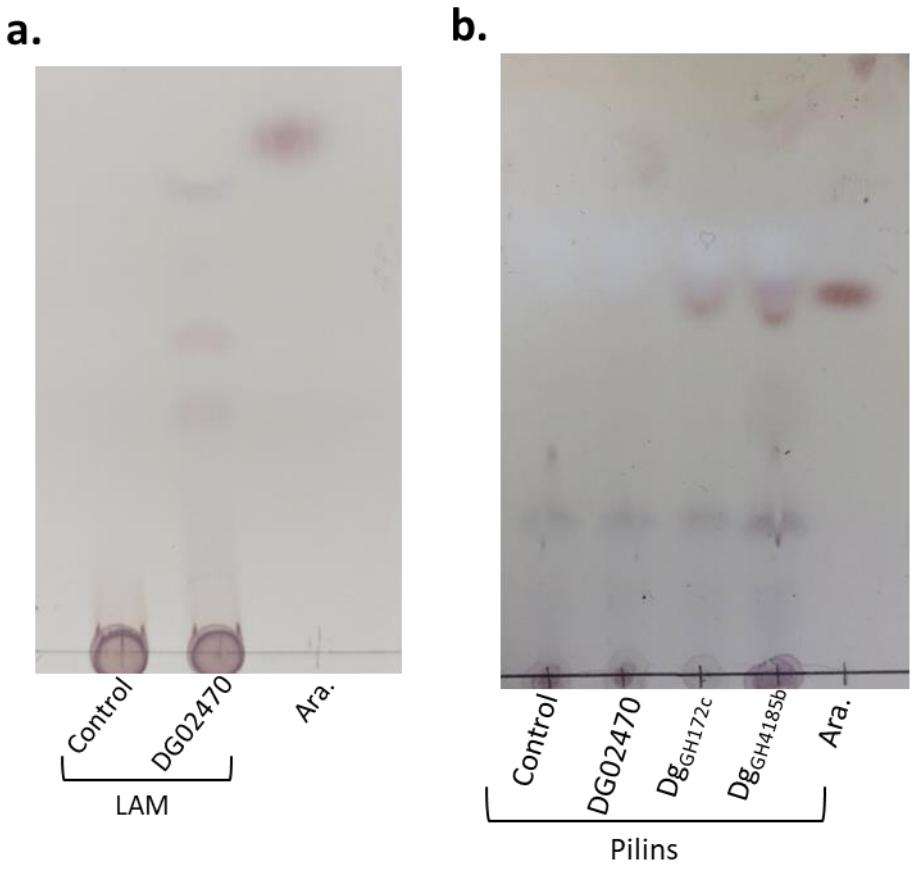


Figure 3.9 TLC of DG02470 vs LAM and pilins. A. 1 µM DG02470 incubated overnight with LAM. **B.** Pilins incubated over night with 1 µM Dg02470 $DG_{GH1722c}$, $Dg_{GH4185b}$. Stained with orcinol.

3.2.4 Structure

Crystal trays were set up to determine the structure of DG02470 purified through a 16/600 Superdex® 200 pg (Cytivia) gel filtration column. Once a single sharp peak was detected, fractions which corresponded with the peak were run on 12.5% SDS to assess purity (**Figure 3.2b**). All visualisations of structure were performed in ChimeraX (V.1.7.1) (Meng et al., 2023).

3.2.4.1 Crystallography trays.

10 nl Index (Hampton Research) and PACT, Structure, Index, JCSG+ (Molecular Dimensions) trays were set up for the crystallisation of DG02470 at 18 mg ml^{-1}. Crystals that developed in 0.1M MMT buffer, pH 6.0 and 2.5% w/v PEG 1500 were selected as a candidate. These were fished by Dr Arnaud Baslé and sent to Diamond Light source for synchrotron analysis. The structure was then solved via molecular replacement by Dr Arnaud Baslé.

3.2.4.2 Structure

The structure was determined by molecular replacement using an AlphaFold model developed by Dr Arnaud Baslé shown in **Figure 3.10**. DG02470 was monomeric in the crystal structure, with a tertiary structure of a C-terminal- (α/α)6-barrel fold with and a N-terminal β-sandwich.

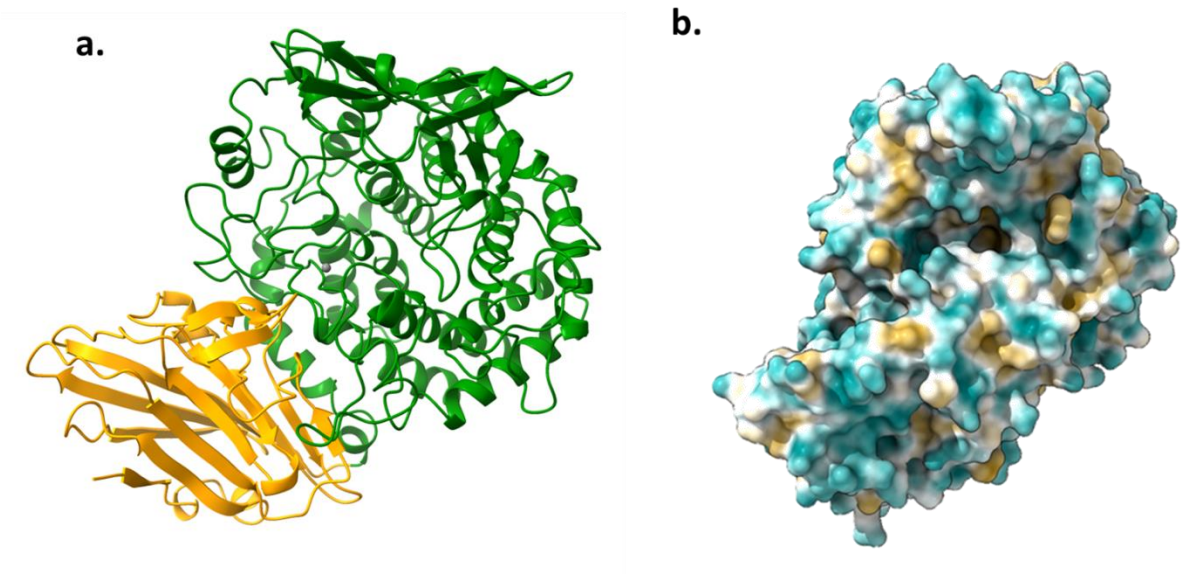


Figure 3.10 Structure and hydrophobicity of DG02470. A. Crystal structure of DG02470 with β-sandwich (yellow) and (α/α)6-barrel (green). **B.** Hydrophobicity map of the crustal structure of DG02470 with blue representing more increasing hydrophilic regions and yellow representing more hydrophobic regions.

A metal ion was observed in the density close to the centre of the (α/α)6 barrel. The PDB file was input in CheckMyMetal (Gucwa et al., 2023) which allows for the determination of a presence of metal within the structure and the probable metal. The metal present was determined to be zinc due to tetrahedral geometry of the retaining residues. These are shown in **Figure 3.11** these residues were identified as H147, C475, D557, H632.

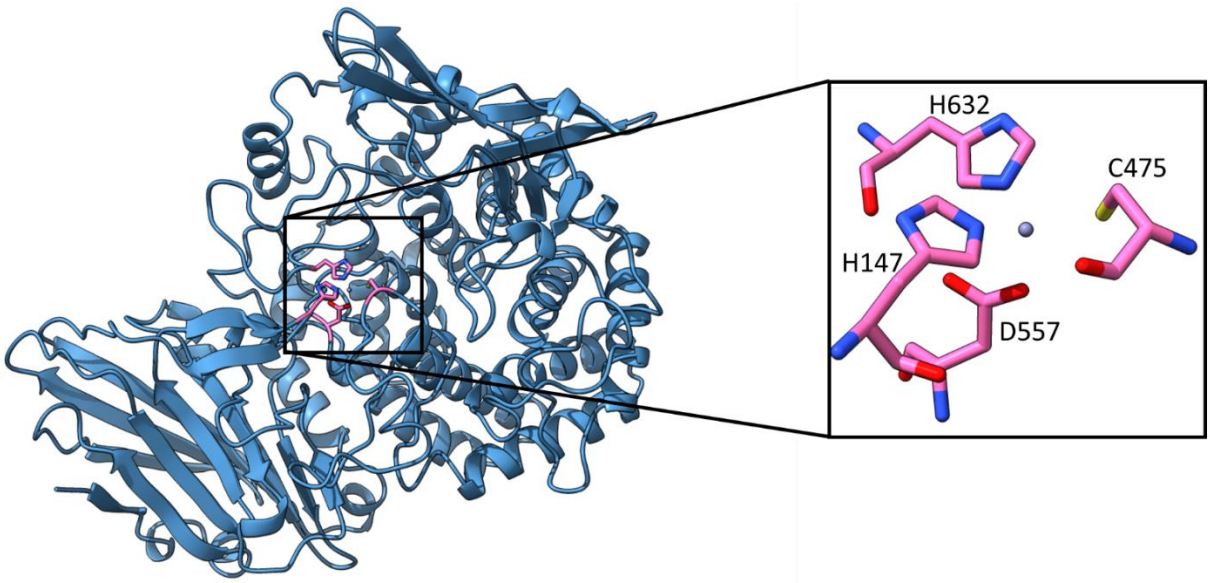


Figure 3.11 Metal retaining residues in DG02470. Structure of DG02470 with the metal retaining residue H147, C475, D557, H632 coloured in pink and the zinc metal in grey. Showing the tetrahedral geometry formed.

3.2.4.3 Structural analysis

Structural alignment of the experimentally determined structure of Dg02470 was done via Foldseek (van Kempen et al., 2023), which aligns tertiary structure across databases, rather than soley amino acid alignment. The results showed a high similarity in structure to the N-terminal (α/α)6-barrel fold, this domain appears to be conserved across several GH families (Lee et al., 2007). The N-terminal β-sandwich domain conversely shows a reduced conservation across identified proteins.

3.2.4.4 Metal sequestration assay

After determination using CheckMyMetal (Gucwa et al., 2023) of the Zn present in DG02470, the next step was to determine if DG02470 is metal dependant. A metal sequestration assay was performed in the presence of 50 mM EDTA which sequesters the native metal and allows the addition of other metals in their Cl form to determine if the metal is essential. The results are shown in **Figure 3.12** indicate that although a zinc is present in the structure as detailed in **Section 3.2.3.2** it is not essential as the post exchange treatment, theoretically containing no metal, is still active upon AG when incubated overnight producing an arabinose band. Additionally, the addition of no other metals tests Ca, Co, Cu, Mg, Mn, Ni and Zn causes a loss in activity of DG02470. Suggesting that although containing a zinc molecule the protein is not metal dependant.

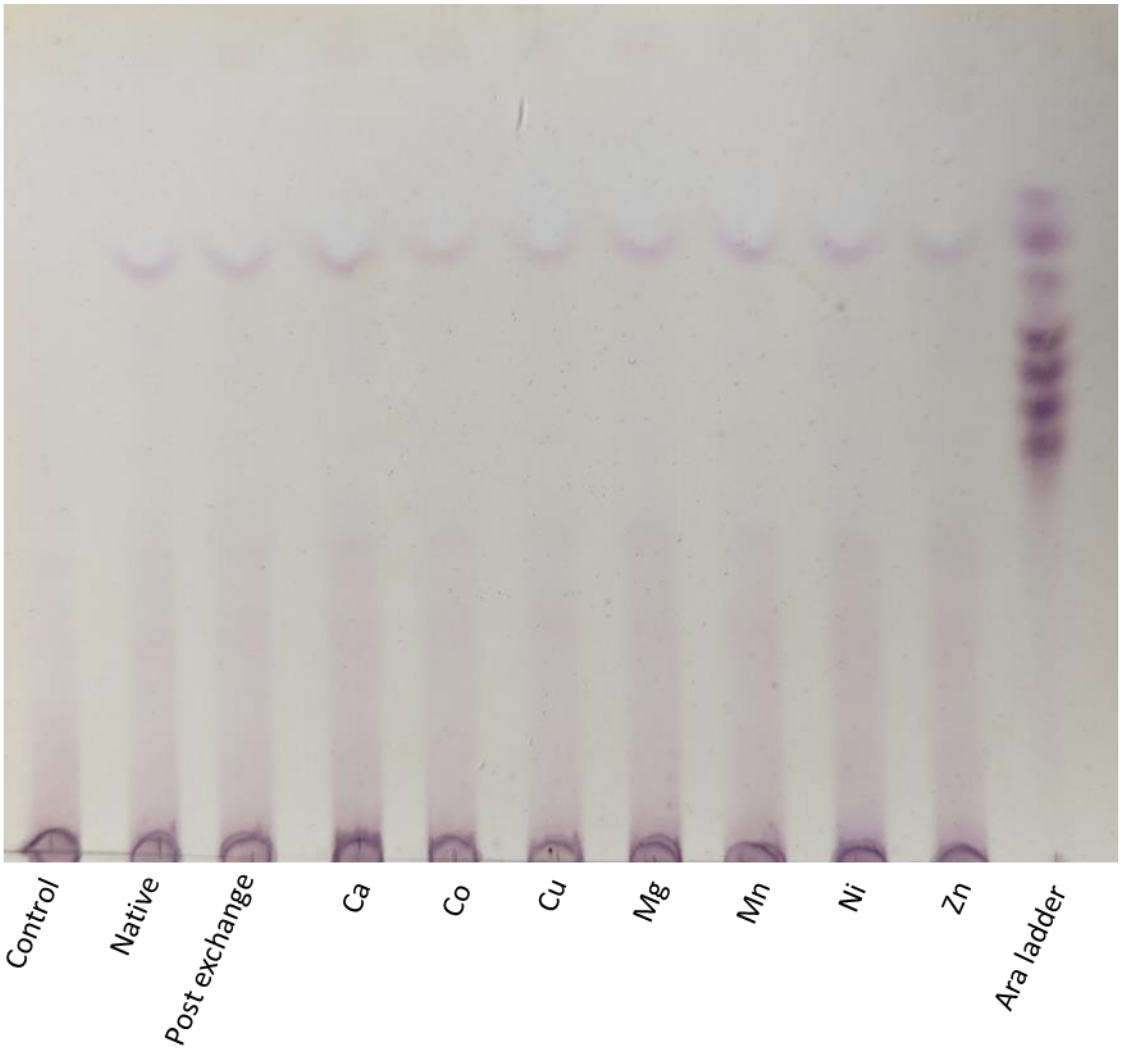


Figure 3.12 TLC of metal sequestration assay of DG02470. DG02470 treated with 50mM EDTA to sequester metal. The reintubated with different metals. Using the native enzyme as a control, post exchange indicating post buffer exchange from EDTA.

3.2.4.1 Alpha Fold comparison

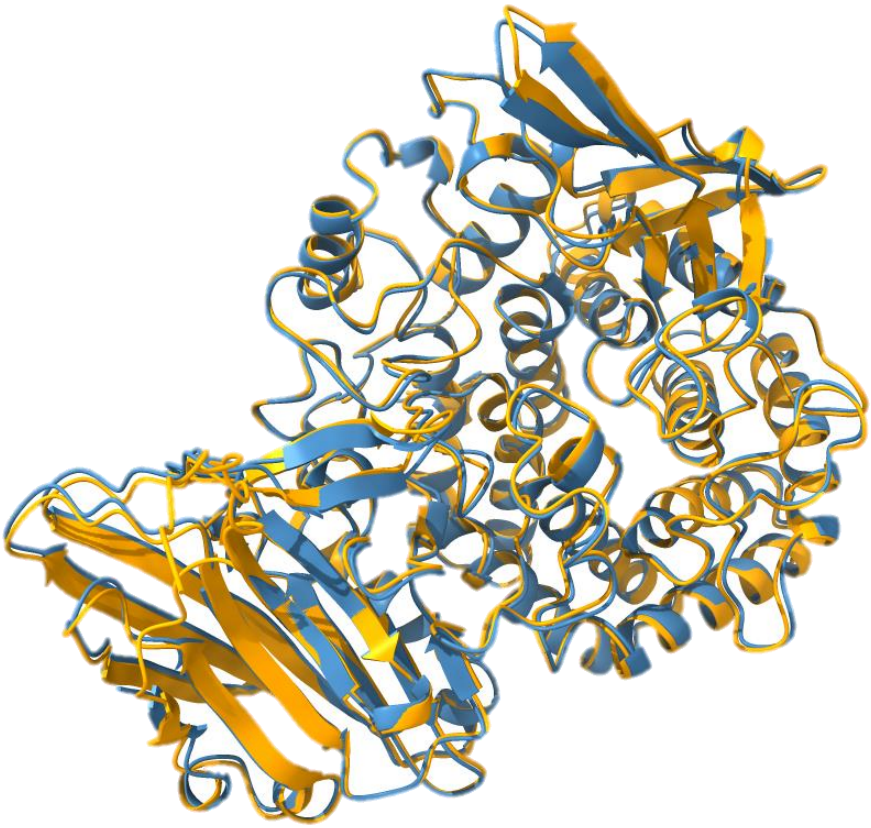


Figure 3.13 Crystal derived structure of DG02470 aligned with alphafold. The crystal derived structure of DG024710 (blue) aligned the ALPHA fold predicted (orange).

When the AlphaFold2 prediction was aligned with the structure obtained from crystallography in ChimeraX (**Figure 3.13**), the RMSD between the two structures was 0.573 Å. This therefore shows a close similarity between the AlphaFold2 predicted structure and the structure determined via x-ray crystallography and molecular replacement.

3.2.1 Site Directed mutagenesis.

To determine which residues were active in DG02470, SDM was used to alter the target residue to alanine. Three methods were used to identify targets: the metal binding residues were selected, Consurf (Ashkenazy et al., 2016) was used to identify highly conserved Glutamic and aspartate residues due to the role they play in the active sites of GH enzymes, aspartic acids often work as a general acid allowing protonation of the

hydroxyl group with glutamic acid acting as a general base. Finally structural analysis was used to identify residues of interest.

All sequences were checked vie Tube seq (Eurofins) to ensure all mutations were successful and correct. All mutants expressed to similar level as the WT.

3.2.1.1 Metal Binding residues.

During the CheckMyMetal analysis, 4 tetrahedral metal retaining residues were identified these were used to produce singly point mutations using site directed mutagenesis, altering the target residue to an alanine, to further analysis the role of the metal in the enzyme. These residues were H147, C475, D557, H632, shown in **Figure 3.12.**

3.2.1.2 Active site determination

The sequence of the cloned DG02470 was analysed through Consurf (Ashkenazy et al., 2016), Consurf analyses the evolutionary pattern of amino acids of a macromolecule revealing areas of conservation that are important for structure or function across homologues (**Figure 3.14**). Residues with a conservation score of 9,9, indicate no change in residue across identified homologues. Conserved aspartic or glutamic acids were selected as targets for site directed mutagenesis to identify if any of them were potentially active residues. Along with the use of Consurf, structural analysis was performed of the crystal structure, to identify any other putative active residues.

Figure 3.14 Consurf of amino acid conservation results for DG02470. Consurf (Ashkenazy et al., 2016) analysis of the conservation of each amino acid across 125 homologues of DG02470 down to 40% similarity. Consurf which analyses homologues of the query identifying regions of high conservation.

Using these two techniques we used side directed mutagenesis to alter the target amino acids to alanine using site directed mutagenesis. The identified residues are shown in **Figure 3.15** these were: W270, W352, E467, D468, E470, E530, E559, D562, E628 and D633.

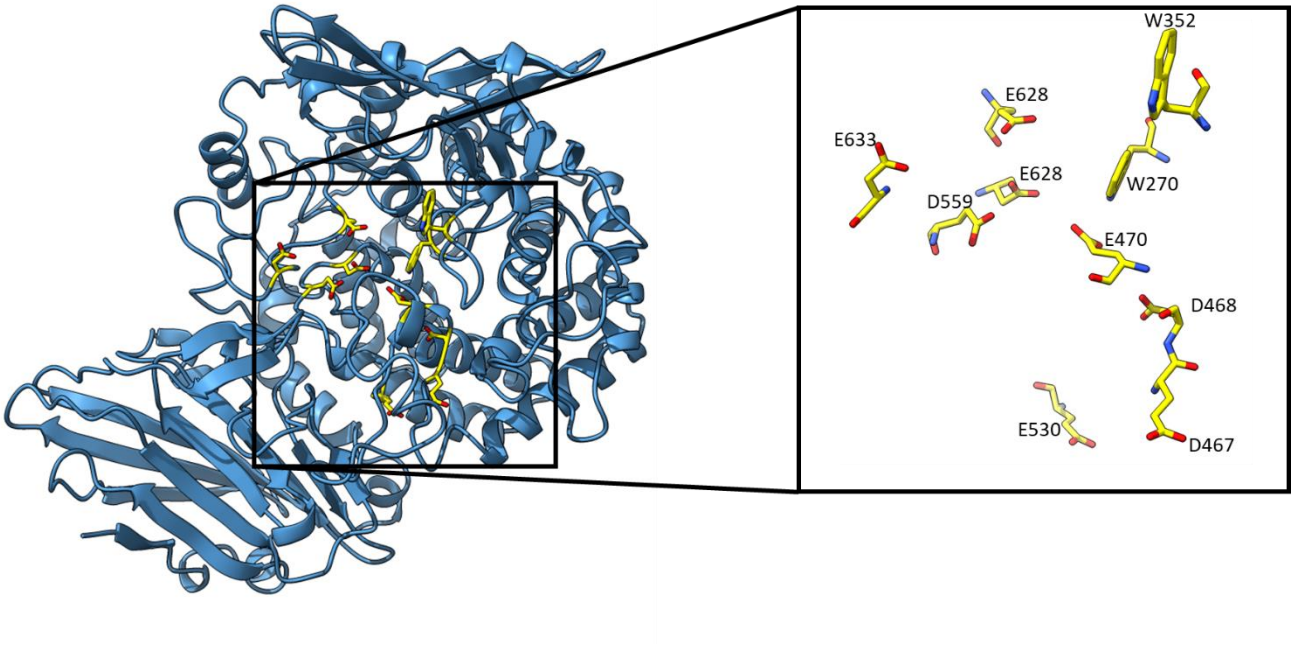


Figure 3.15 Putative active residues in DG02470. Putative active residues (yellow) determined via ConSurf and structural analysis of DG02470.

3.2.2 Activity of DG02470 mutants.

To determine the if mutants knocked out activity if both putative active site residue mutants and metal binding mutants were incubated overnight at 1 µM with 2 mg ml^{-1} the assays were then analysed on TLC shown in **Figure 3.16.**

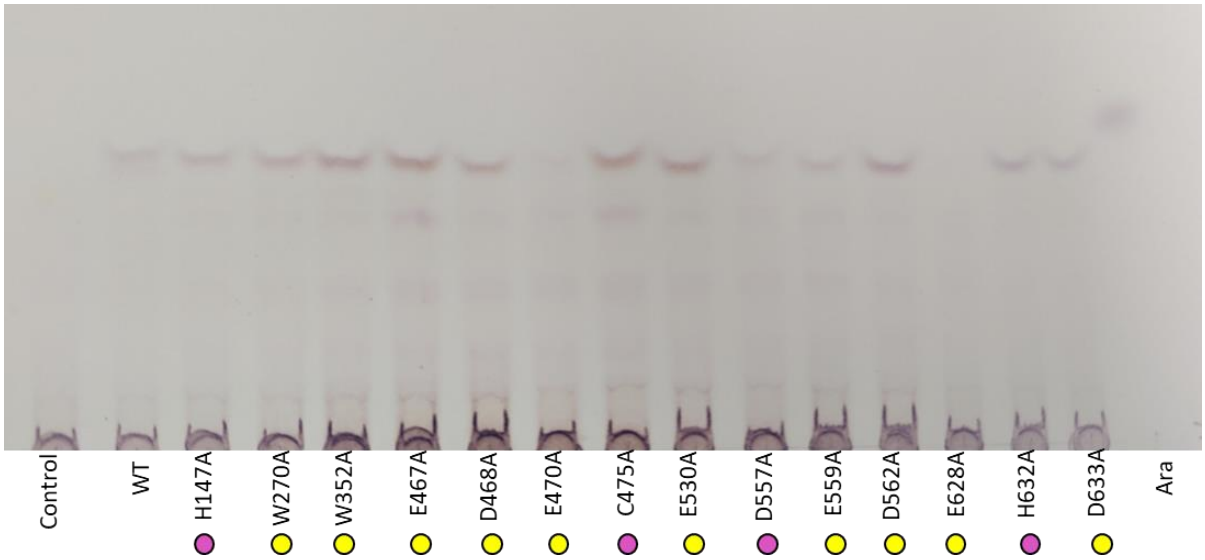


Figure 3.16 TLC of DG0270 mutant's vs AG. TLC of 1 µM DG02470 mutants vs 2 mg ml^{-1} stained with orcinol. Pink filled circles represent metal binding residues. Yellow filled circles represent putative active site mutants.

Two mutations appeared to eliminate activity entirely E470A and E628A, both putative active site mutants. No metal binding residue mutants fully inhibited the activity of the DG02470, although this is expected as the EDTA sequestration assay showed that the zinc was nonessential. HPAEC-PAD analysis was not performed to fully analyse the impact on activity, of each mutant and it may be the case that activity is reduced but not entirely knocked out.

The residues which knocked out activity are shown in **Figure 3.17b.** Using this information, we were able to additionally identify that DG02470 has pocket topology, which matches the previously shown hydrophobicity map **Figure 3.11b**, as active sites are often located in areas of high hydrophobicity.

The distance between E470 and E628 residues has been measured at 9.1 Å. This is characteristic of an inverting mechanism in which the residues are typically between 6-11 Å apart.

Identifying that the active residues E470 and E628 of DG02470, we are able to analyse using the hydrophobicity tool in ChimeraX that they are in an area of high hydrophobicity; this also clarified that DG02470 has a pocket topology (**Figure 3.17b**).

A BLAST (Altschul et al., 1990) search was ran of DG02470 with a percentage identity cut off of 65 percent for 100 homologues, only 99 homologues were identifies with this cutoff, The multiple sequence alignment was performed using MUSCLE (Edgar, 2004). The regions identified as active residues (D470 and E628) were visualised in WebLogo (V.3.7.12) (Crooks et al., 2004; **Figure 3.18**). The visualisation is based on the prevalence of each amino acid aligned homologues with an increased conservation of the amino acid being shown as a larger letter. The analysis shows a high level of conservation of residues surrounding the active residues as well. In the case of E470 two conserved tryptophan are highly conserved in close proximity.

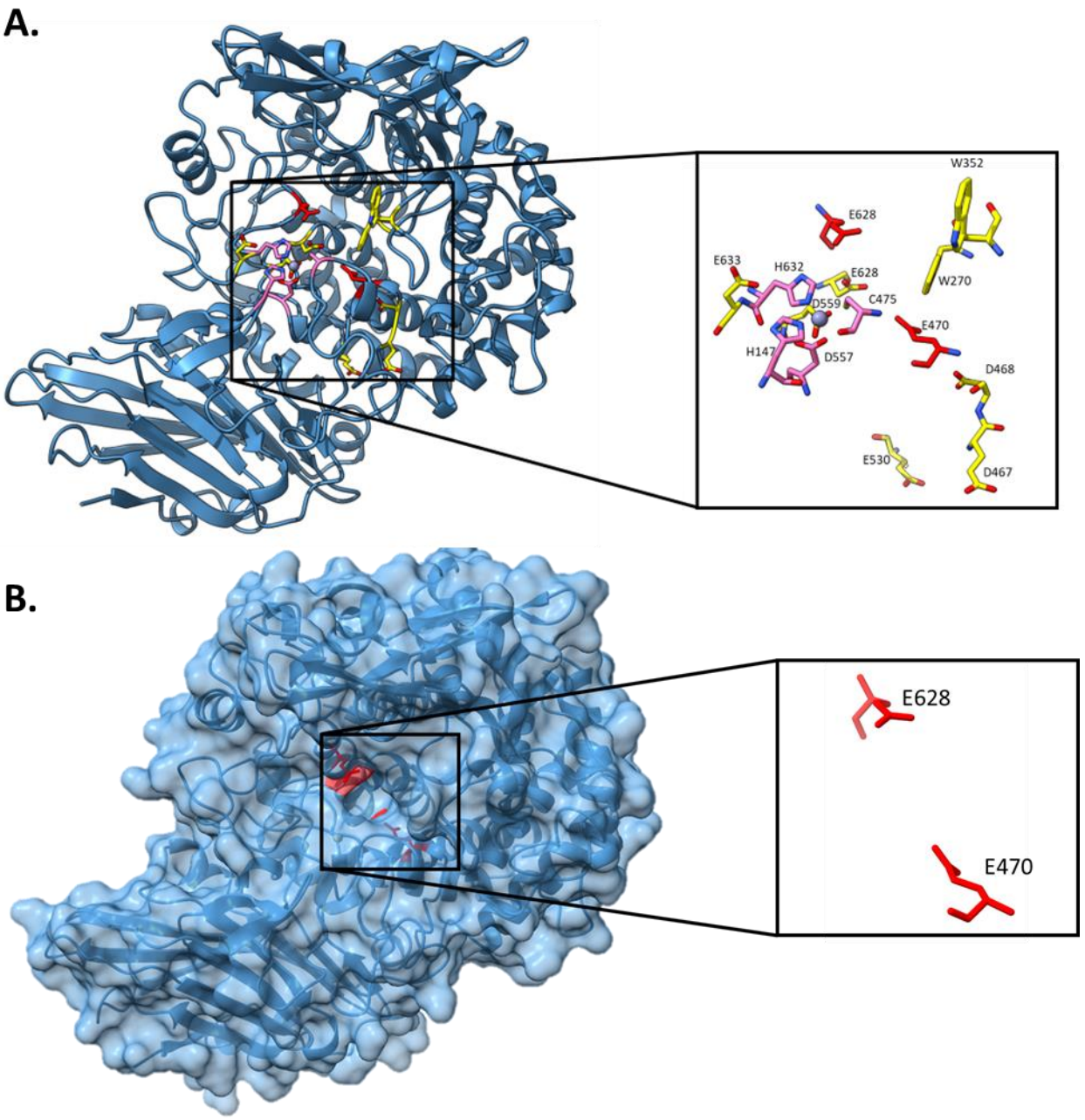


Figure 3.17 SDM mutants of DG0270 and identified active residues. A. SDM residue targets with metal binding residues (pink), putative active residues (yellow) and identified active sites (red). **B.** identified active site residues, showing pocket topology of active site.

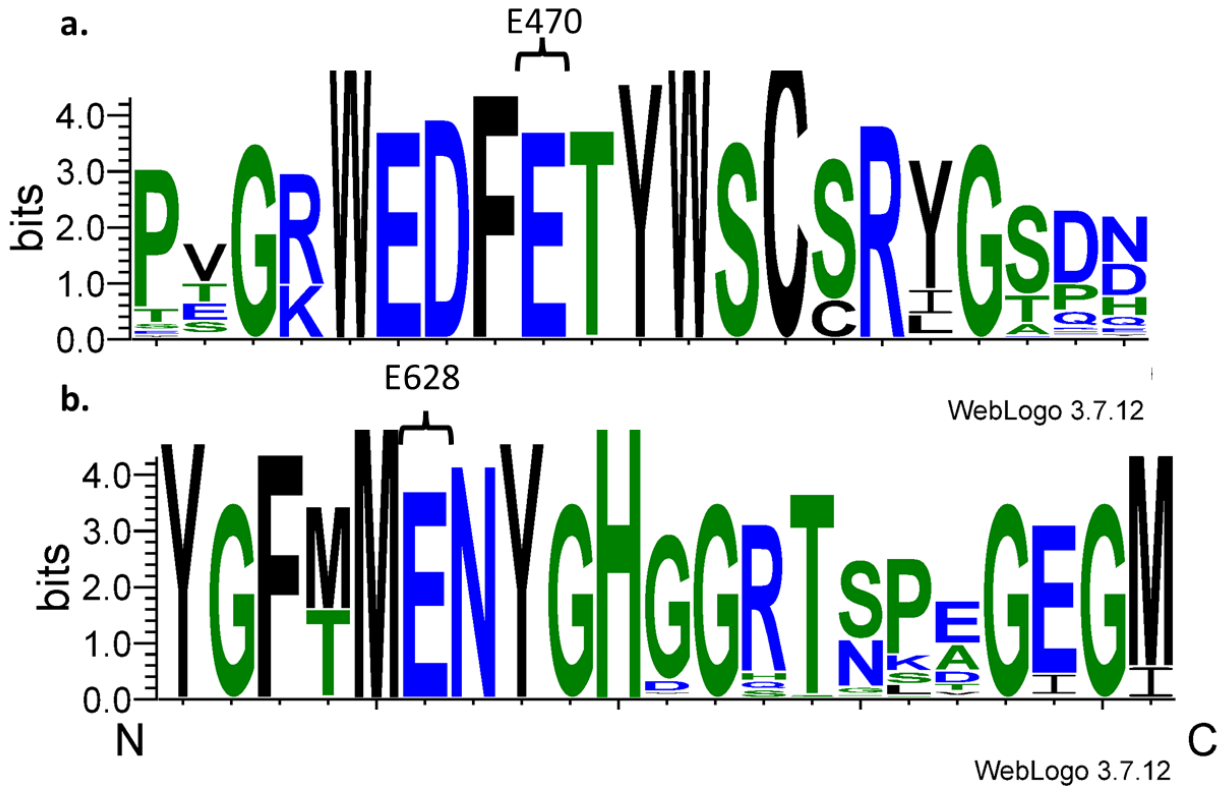


Figure 3.18 Conservation of active site residues across homologues. Conservation of regions of containing active site mutants using MUSCLE (EMBL-EBI) and visualised in WEBLOGO across 99 homologues, which were identified via a BLAST search with a similarity cut off 65% identity.

3.2.3 Ligand soaks

As mentions in **Section 3.2.4** two residues were identified as inhibiting activity E470 and E652. E470A was selected to be used to obtain a sugar bound structure. Optimisation of the original conditions we obtained crystals in was performed.

Lack of activity shown by DG02470 against pNP-β-Ara potentially signified subsite specificity requirement for a longer oligosaccharide. **Experiments detailed in Chapter 4** haveshown the combination of 5 enzymes and DG02470 completely breaks down arabinogalactan into arabinose and galactose (**Figure 3.19**). With the removal of DG02470, oligosaccharides that could not be cleaved by the remaining enzymes were visible on the TLC. These oligosaccharide were used as potential ligands for DG02470 in

soaks. A reaction was set up over night with 1 µM of each of 4 enzymes, $DgGH_{172c}$, $DgGH_{4185b}$, $BfGH_{182}$, $BfGH_{43\_31}$ and with AG.

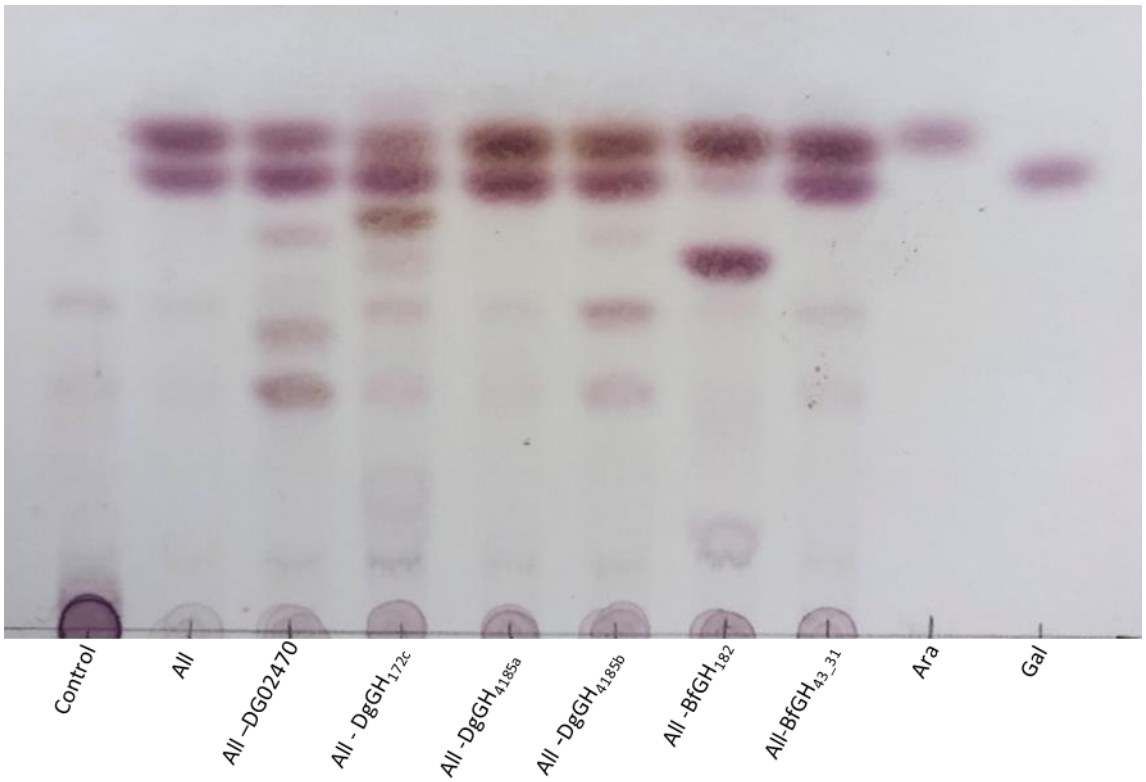
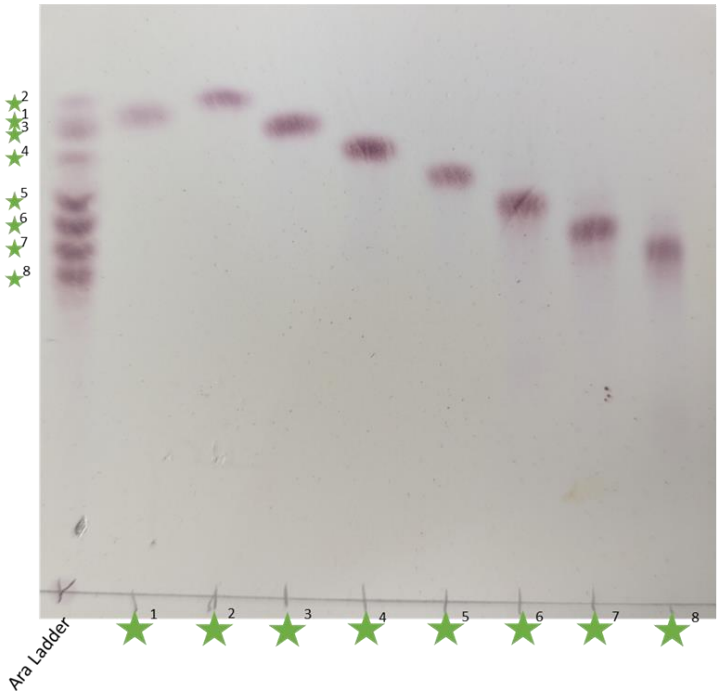


Figure 3.19 TLC analysis of Novel Enzymes vs arabinogalactan. TLC analysis of products generated by incubating 2 mg ml^{-1} AG with 1 µM of each novel enzyme and incubation with the removal of one of the enzymes. Analysing the variation in oligosaccharides with the removal of a novel enzyme.

The resulting reaction was separated via PGC column as described in **Section 2.2.5**. This allows for the separation of arabinose and galactose monosaccharides from oligosaccharides. This is achieved as more polar molecules have a strong retention in the column, with a solvent, in this case butanol being needed to elute them. This also allows separation of oligosaccharides by length. The fractions were then analysed by TLC to identify fractions containing potential substrate. To identify the length of the ligand an arabinose ladder was run on TLC. A comparison of arabinose in different oligomerisation states is show in **Figure 3.20a.**

a.



b.

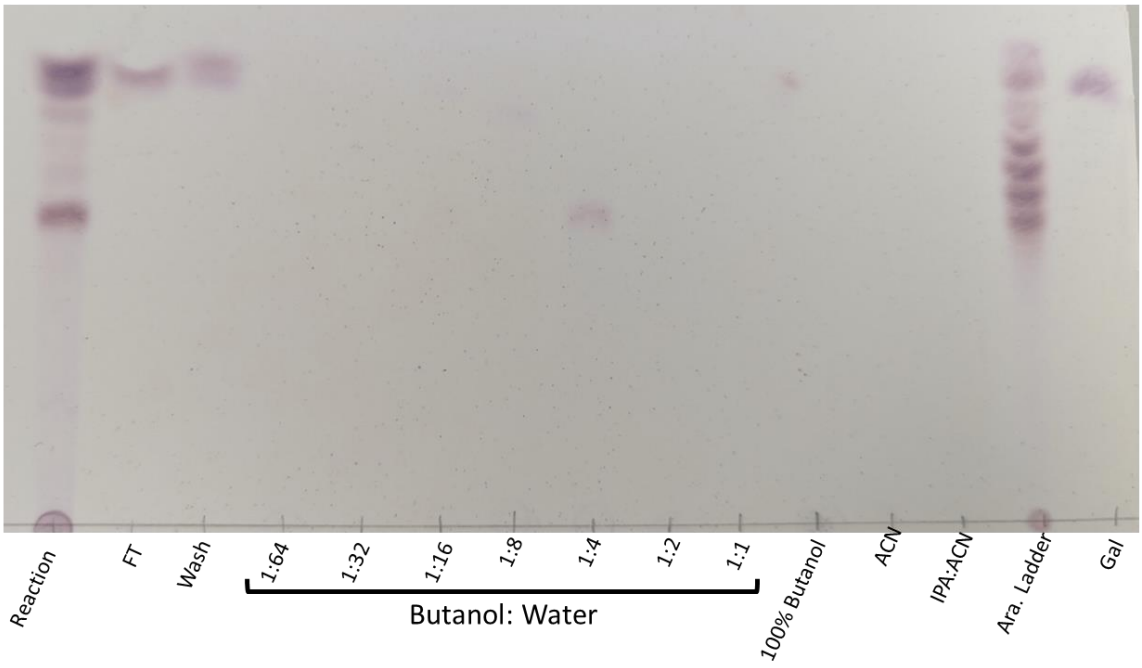


Figure 3.20 TLC of arabinose ladder and PGC fractions. A. L-arabinose ladder of arabinose monosaccharide to L-arabinose octosaccharide. **B.** TLC of PGC fractions obtained from purification of incubation of four proteins: Dg_{GH172c}, $Dg_{GH4185b}$, Bf_{GH43_31} and Bf_{GH182} with AG, producing monosaccharide products and undigested oligosaccharides. Elution fraction increasing Butanol: water ratios.

1:8 and 1:4 fraction had visible oligosaccharides, these two fractions were then lyophilised as described in section x and resuspended in 50 µl of water.

Once E470A crystals had been left for 2 weeks potential candidates were chosen for soaks having the inactive mutant in the presence of the ligand would hopefully give a ligand bound structure. Crystals did diffract, however there was no ligand present.

3.2.4 Sequence similarity network (SSN)

As DG02470 is not a member of an existing GH family, our data suggests it is the founding member of a new enzyme family. To identify other members, sequence similarity network (SSN), which allows the visualisation of the relationships among protein sequence, was performed by Arashdeep Kaur (University of Melbourne) using EFI Enzyme Similarity Tool (EFIEST) (Zallot et al., 2019). This allowed them to create clusters of the proposed family from BLAST similar results, using a hidden Markov model the results were then rerun to increase fidelity. Various cut offs were used to ensure the robustness of the results. Using this model, they identify DG02470 as a new GH family containing a potential 263 other sequences. An alignment score (AS) allows for the grouping of genome neighbourhoods based on alignment similarity with higher scores requiring a higher alignment to be grouped. An alignment score (AS) score of >135 was used to produce **Figure 3.21** with related proteins within the AS being shown by connections. If two proteins are not connected it means the sequences are less similar than alignment score cut off. The node containing GH02470 is predominantly formed of Bacteroidetes phylum. Another node is connected which contains mainly proteins from members of the Bacillota phyla.

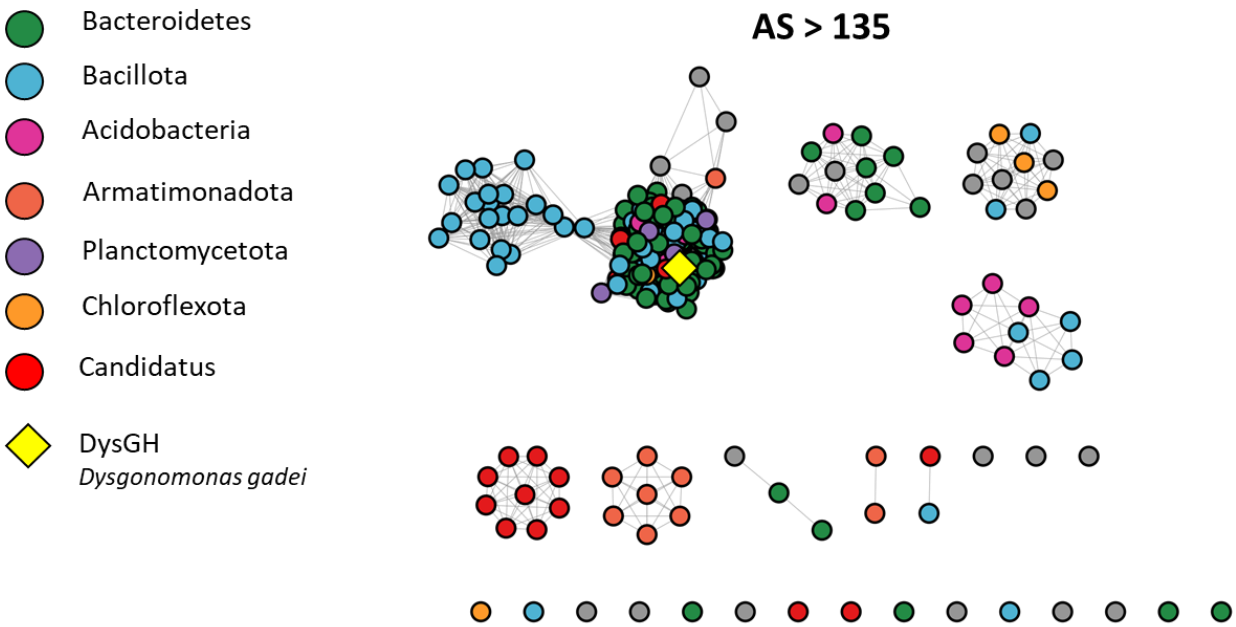


Figure 3.21 SSN of DG02470. SSN of DG02470 using AS>135 grouped into nodes with related proteins within the AS being connected. DG02470 is shown as a yellow diamond. Different coloured circles denote different phyla.

3.2.5 Neighbourhood gene homologues

Genome neighbourhood diagrams (GND) are used to describe genes which neighbour your protein of interest. Using a SSN with an AS of >50 a GND with an open reading frame of +10 and -10 genes was created using EFIEST (Zallot et al., 2019). The network produced is shown in **Figure 3.22** with proteins of the same function shown in the same colour.

The diagram shows the occurrence of each type of protein adjacent to DG02470 homologues. In the neighbourhood groups common genes present are GH172 and GH183s, which are known to be associated with D-arabinan degradation (Al-Jourani et al., 2023).

Within the GND there are also GH76 and GH125c which are both α-mannan degrading GH families (Gregg et al., 2011; Solanki et al., 2022). There is potential that this indicates that there are homologues of DG02470 which are responsible of the degradation of LAM, which as described in **Section 1.4.4** has a β-Ara*f* cap. Therefore homologues of DG02470 are potentially specialised for the caps of LAM. Interestingly there are also α-L-fucosidases, such as GH29 (Grootaert et al., 2020), within the GND and sulfatases. However, although sulphated L-arabione has been described (Ciancia et al., 2020; Surayot et al., 2016), no evidence of known sulphated D-arabinose was found. The same was true of searches for D- arabinose fucose containing glycans. Therefore, this leaves a question as why these are present.

The bacteria in which GH29s, identified in the GND, are present is *Thermotoga maritima* (PF16757) (Sulzenbacher et al., 2004), a hyperthermophilic organism first discovered in sediment around a marine geothermal vent (Huber et al., 1986) and *Drosophila melanogaster* spermatozoa (PF01120) (Pasini et al., 2008). Two organisms unrelated to *D. gadei.*

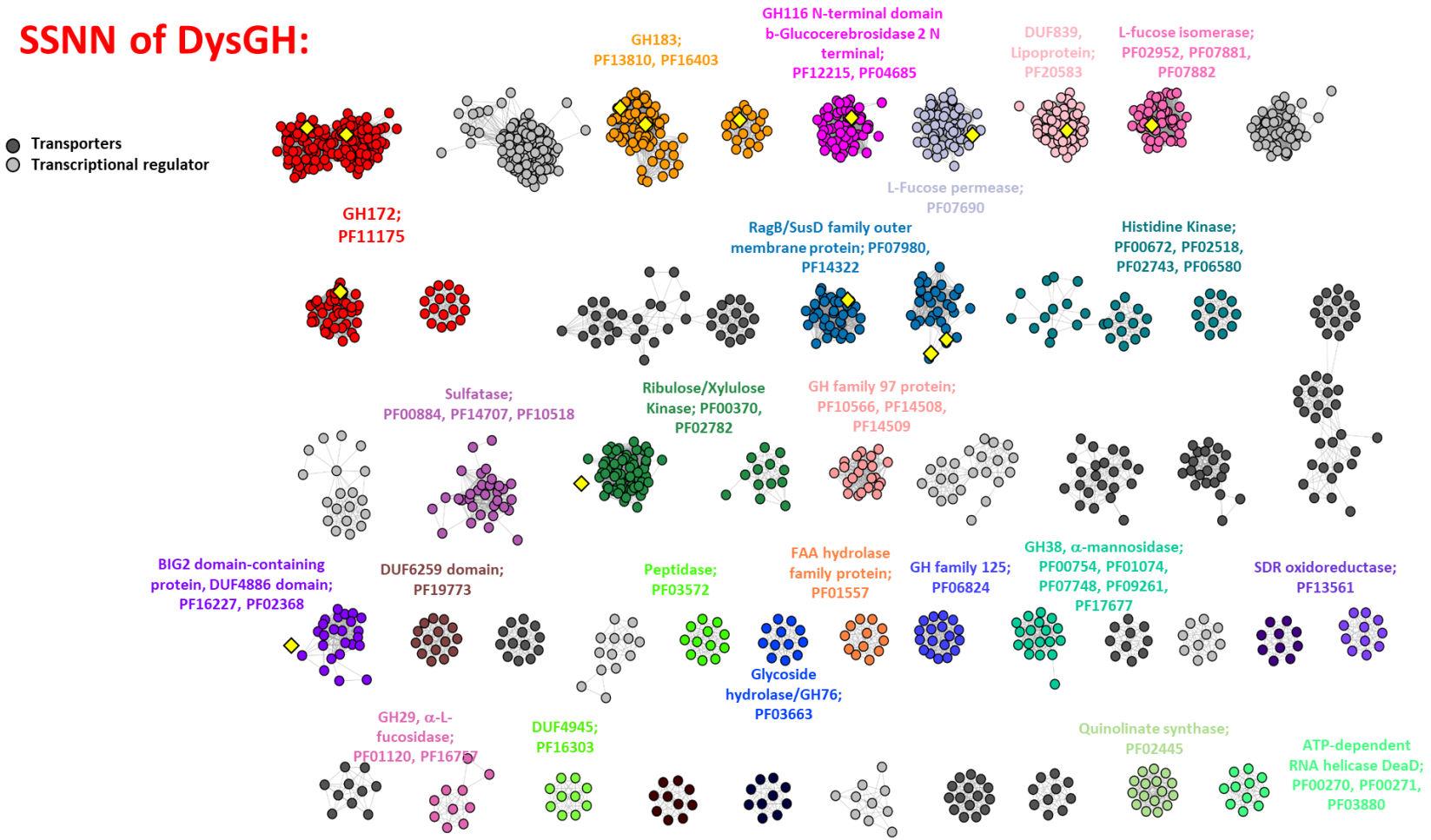


Figure 3.22 Neighbourhood gene analysis. Neighbourhood gene analysis of AS>50 of homologues at DG02470 ±10 open reading frames. DG02470 is shown as a yellow diamond. Families of proteins are shown by different colours.

3.2.6 Similarity to previously identified endo-β-arabinofuranosidase

As previously mentioned, recently Shimokawa et al., (2023) identified ExoMA2$_{GH116}$, an exo-β-arabinofuranosidase, from *M. arabinogalactanolyticum*. To assess the similarity of these the sequences, they were aligned in VectorBuilder and the alignment identity was 21.02% with a similarity of 35.00%.

Structural alignment of ExoMA2$_{GH116}$ and DG02470 is shown in **Figure 3.23** both structures consist of a C-terminal catalytic (α/α)6-barrel fold. Both have an N-terminal domain β-sandwich, however these two domains do not align structurally.

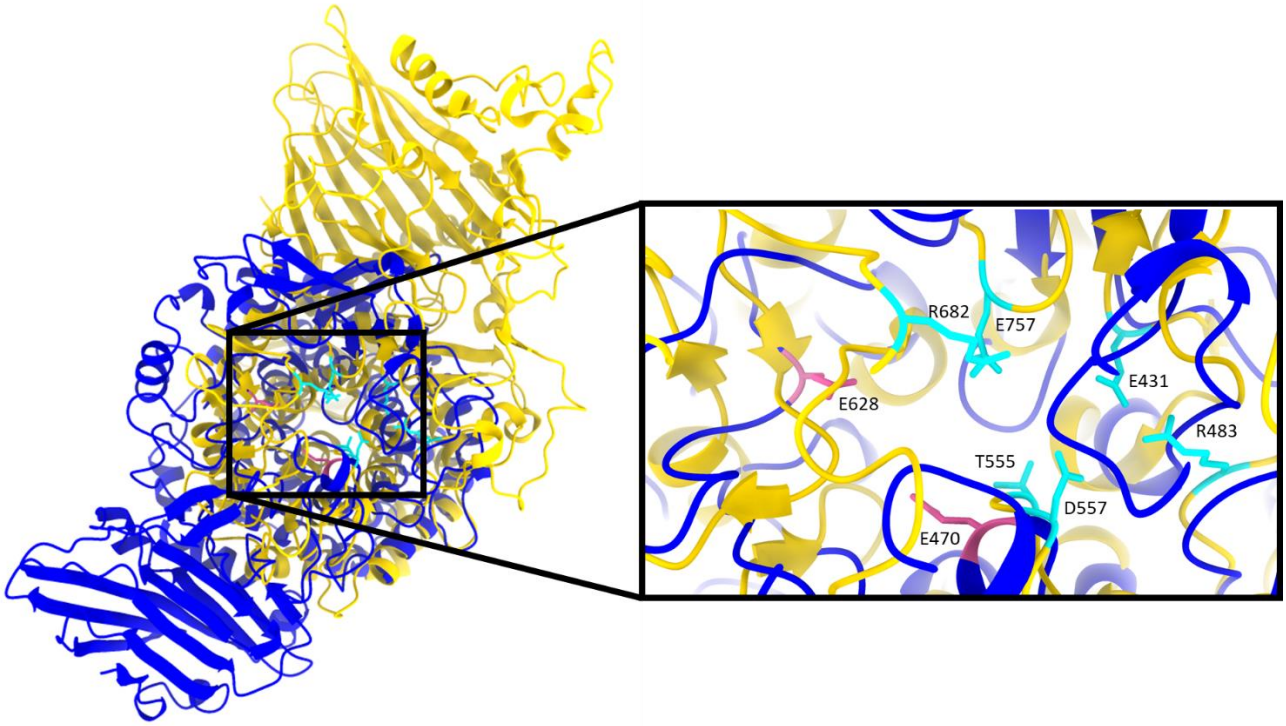


Figure 3.23 Structural alignment of DG02470 and ExoMA2$_{GH116}$. Structural alignment of DG02470 (blue) and ExoMA2$_{GH116}$ (yellow) with the active residues of DG02470 shown in pink and the active residues of ExoMA2$_{GH116}$ show in cyan.

3.3 Discussion

In this chapter we have characterised, through biochemical, structural and sequence analysis, a periplasmic exo-β-D-(1,2) arabinofuranosidase enzyme, which is the founder member of a new GH family from PUL 42 of *D. gadei* in which GH172s and GH182, both with α-D-Ara*f* activity, have been previously identified.

ExoMA2$_{GH116}$ is the only exo-β-D-(1,2) Ara*f* identified (Shimokawa et al., 2023). Although sharing activity however our sequence and structural analysis has shown that DG02470 is not related to GH116, with only a 21.02% identity and a poor structural alignment.

Through sequence analysis using SignalP6.0 (Teufel et al., 2022) it was determined that DG02470 is a periplasmic protein. Initial TLC analysis showed DG02470 is active upon AG and identify it was an exo-D-ara*f*ase. This was corroborated through further analysis using HPAEC-PAD analysis of the products. The activity was also shown via TLC analysis on LAM but not pilins, which allowed us insight into a potential specificity, as the only difference between the two glycans was that LAM has β-D-Ara*f* bonds which are not present in pilins. Using this we hypothesised that DG02470 was active upon β-(1,2)-D-Ara*f* bonds.

Unfortunately, we were not able to confirm the activity of DG02470 using β-ara*f*-PNp, however we were able to show a lack of activity when incubated with Δ*aftb* AG, which was provided by the Moynihan Lab (Birmingham). The discrepancy in results could be due to the subsite specificity of DG02470 which requires additional subsites, which are not present on β-ara*f*-PNp, to be occupied for productive binding. There is also the potential that the pNP molecule interferes with the activity of DG02470, when compared to WT AG. This would have been identified by a substrate bound complex, however as stated we were unable to obtain one.

X-ray crystallography allowed the determination of the structure of DG02470 to 1.8 Å; that of a monomeric protein with a C-terminal- (α/α)6-barrel fold with and a N-terminal β-sandwich. This allowed us insights into potential active sites and residues. It also allowed us to determine that DG02470 contained a zinc molecule in close proximity to the active site, although no experimental assays suggested DG02470 was metal dependant. Structural similarity analysis has shown that the C-terminal (α/α)6-barrel fold region is

conserved across structures analysed, which was expected due to the common nature of the (α/α)6-barrel fold within glycoside hydrolases being characteristic of several GH families (Lee et al., 2007), however the N-terminal β-sandwich roll has little homology across the search.

Although not conclusive we can theorise that the mechanism of hydrolysis is inverting due to structural analysis of the active site residues, which they are 9.1 Å apart which is consistent with inverting mechanism (McCarter & Stephen Withers, 1994), with both E470 and E628 acting as the general acid and general base for the inverting mechanism. It should be noted that there were highly conserved residues present in the protein however, due to time constraints, these were not checked. In close proximity to E470 is a highly conserved tryptophan (W466) due the anomeric amino acid being so close to the active residue it this is possible this residue plays a role in sugar stacking. There is scope for further work in this area.

It is hypothesised that in conjunction with other enzymes characterised within PUL42, such as the GH182s and GH172, the complete breakdown of the arabinan domain of AG is facilitated to provide monosaccharides for use by the cell.

The reasoning for the presence of PUL 42 which facilitates the breakdown of D-Ara*f*, a polysaccharide which itself is not found in *D. gadei* although it is an interesting area of research as to why gut bacteria require a PUL dedicated to the breakdown of a polysaccharide found in Mycobacteriales, which are not reported to be found in the human gut microbiota.

3.4 Future work.

Future work will include: (i) the identification of other residues which are essential for the activity of DG02470 using similarity searches and (ii) full characterisation of the new GH family to which DG02470 is a member. Additional work on homologues could aid in understanding of why this family of enzymes is present in a species such as *D. gadei* which does not necessarily encounter D-arabinan molecule in nature, however it has a PUL with GH172s and a GH183 which target D-arabinan.

There are still further proteins to characterise within PUL42 of *D. gadei.* Additionally, as stated previously only highly conserved glutamic and aspartic residues were mutated, it is likely there are other essential residues that were not identified in this study. This is an area for future work.

Chapter 4. Enzymatic Lysis Protocol development

DNA extraction from mycobacteria has proven particularly difficult to extract which in turn causes sequencing difficulty due to low yield or poor quality DNA, with a great deal of this down to the aforementioned multilayered cell wall (**Section 1.4**). This has led to a plethora of issues, which are not as dominant among other bacteria, including a difficulty in diagnosis as well as disease tracking (Satta et al., 2018).

Traditional DNA extraction techniques for mycobacteria involve harsh chemicals such as phenol, long experiment times or abrasive mechanical means such as bead beating or sonication which while effective produce low quality sheared DNA often inadequate for downstream analysis which has become increasingly important with the rise of WGS (Amaro et al., 2008; Bouso & Planet, 2019; Epperson & Strong, 2020; Odumeru et al., 2001; Votintseva et al., 2015). This has resulted in the need for a scalable efficient method that yields high enough quantity and quality of gDNA for downstream processing to be of high importance.

There are several issues with current DNA extraction techniques for mycobacterium. Phenol-chloroform extraction, a common technique for long read DNA extraction, involves hazardous substances that not only complicate the extraction process but also the disposal of such reagents can prove difficult. Thus, its use is largely unsuitable in the field, or in less specialised labs where the fume hoods or the correct disposal routes can be rare. Bead beating also has issues aside from the highly sheared DNA obtained, there is large upfront cost to the cost of the equipment as well as the logistical issues of transportation due to the weight. Enzymatic lysis is a solution to this, as the reagents are often low hazard and relatively inexpensive, with lyophilised enzymes, such as lysozyme being able to be stored for long periods without specialist equipment. However, there is a lack of specialised commercially available enzymes targeted towards the mycobacterial cell wall.

Although a rapid detection kit does exist for Mtb, GeneXpert (MTB/RIF), with a high turnaround speed and the ability to identify the presence of Mtb as well as the resistance to rifampicin and isoniazid. It is limited to Mtb detection and even this common diagnosis test is quite prohibitive in its cost with each test costing $20-$30 which can also vary between countries (Kaso & Hailu, 2021).

4.1 DNA extraction principles.

DNA extraction can be done in many ways, with varying time usage and cost. The most commonly used commercial kits are based on gravity flow column, spin columns or magnetic beads. Each of these comes with benefits and downsides. Gravity flow columns and magnetic beads produce higher length fragments which is ideal for Nanopore sequencing, however this comes at a higher cost per sample and higher user error due to the increased level of skill needed to use. Spin columns work by binding DNA to a matrix via centrifugation. The quality of the DNA obtained from this process tends to be shorter fragments due to the rougher treatment during extraction, however at a lower sample cost (Quick & Loman, 2018).

4.1.1 Cell lysis

The overall extraction of DNA using the majority of commercial kits is done in three stages: breakdown of the cell wall, lysis of the remaining cells and DNA capture.

Cell lysis is the first step in DNA extraction the aim of this is to break down the cell wall in an osmoprotective buffer. Commonly this is done with lysozyme, with a buffer containing SDS. In the case of mycobacteria, a commercially available lysozyme alternative is not available however work on LysA an amidase which is specifically active upon the PG of mycobacteria has shown anti-tuberculosis activity, so might be an option for future work (Gil et al., 2010).

The secondary lysis step introduces a chaotrope such as guanidinium chloride, which causes de stabilisation of van der waal bonds within the cell membrane, as well as being a strong protein denaturant which with the addition Protinease K aids in the denaturation of nucleases to aiding in the maintenance of DNA integrity (Aşır et al., 2016). Ethanol is then used to precipitate the DNA. The method by which the DNA is captured varies, with spin columns the DNA is bound to a silica matrix. While bead extraction relies on binding the DNA to glass bead and washing. Both are eluted with a high salt elution buffer.

Although no specific lipolytic enzymes or endopeptidase are currently commercially available for mycobacterium, as previously described novel enzymes recently identified

by Al-Jourani et al. (2023) can hydrolyse the AG domain into its constituent monosaccharides have been characterised, along with DG02470 characterised in **Chapter 3**. Due to the essential nature of AG (Alderwick et al., 2015), this makes the domain a prime target in the enzymatic lysis of mycobacteria. Additionally, due to AGs linking to PG, the targeting of AG may also allow commercially available lysozyme to act upon the PG more effectively.

4.2 Objectives.

I. Compare enzymatic lysis to existing DNA extraction techniques for mycobacteria.
II. Optimise the lysis conditions for time and yield.
III. Test across different acid –fast bacteria.

4.3 Results

The two bacteria selected for the optimisation of the protocol were *M. smegmatis* Mc2 155, due to its wide use as a non-pathogenic laboratory model for Mtb, and *M. abscessus* subsp. abscessus, herein referred to as *M. abscessus*, due to its previously described importance in a clinical environment.

To normalise gDNA yield for cell concentration, we initially tested CFU, correlating this to OD_{600}. However, due to the growth pattern of mycobacterium, especially that of *M. smegmatis* on solid media making identification of individual colonies difficult and with the addition of slow growing species to the testing panel making time also became a factor, with BCG taking up to 3 weeks to grow, we decided to use wet weight for the normalisation of DNA quantity, as it is a much more reliable method than OD-CFU. Bead-beating was always used as an internal control even if not shown on graphs for clarity of data representation. GenElute™ Bacterial Genomic DNA Kit (Merck), including their recommended lysozyme concentration of 45 mg ml^{-1}, were used for the extractions. Other kits were considered however these offered a balance between yield, quality, time and ease of use. This allowed for a protocol which is more accessible with a low training time and concise protocol. Without the need for, previously discussed, phenol: chloroform extraction, which ideally needs a fume hood for utilisation and correct disposal methods (Trigodet et al., 2022). gDNA quantification for *M. smegmatis* and *M. abscessus* subsp. *abscessus* was done by qPCR as described in **Section 2.1.24.1.** Other species DNA were quantified via Qubit due to the time constraints of qPCR primer design. All results were analysed using One-way ANOVA statistical tests within GraphPad Prism 9.5.0.

4.3.1 M. smegmatis and M. abscessus results.

4.3.2.1 *Testing enzymatic lysis*

In order to test whether enzymatic lysis was possible, cells were incubated with 5 µM of each enzyme, as well as lysozyme and lipase for two hours, in a thermal mixer at 37 °C. This was carried out in the Gram-positive lysis buffer' provided in the GenElute Kit to ensure it was compatible with the rest of the gDNA extraction procedure. Initial tests to assess the effectiveness of using the novel enzymes in the extraction of gDNA from *M. smegmatis* and *M. abscessus* were performed in the conditions described in **Table 4.2.** The addition of the novel enzymes to the enzymatic lysis step showed a significant increase in gDNA compared to enzymatic lysis without the novel enzymes for both *M. smegmatis* with a mean difference of 26.61 ng μl^{-1} and *M. abscessus* with a mean difference of 11.67 ng μl^{-1} when incubated during lysis step for 2 hours (**Figure 4.1**). When compared to bead beating there is no significant difference in the yield when compared to complete cocktail for both *M. smegmatis* and *M. abscessus.* This initial data suggests that the addition of enzymes able to hydrolyse the AG domain of the mycobacterial cell wall increases the yield of gDNA compared to lysozyme and lipase alone, and comparable yields to the current standard of bead-beating.

Table 4.1 Novel enzymes. Novel Enzymes identified as candidates to increase the yield of gDNA during mycobacterial DNA extraction, previously described in **Section 2.1.24.1** and Novel GH described in **Chapter 3**.

Enzyme (ascension)	GH Family	Bacterial host	Enzymatic target
DG02470 (HMPREF9455_02470)	Unknown	*D. gadei*	Exo-β-Ara*f*
DgGH$_{172c}$ (HMPREF9455_02479)	GH172c	*D. gadei*	Exo-α-Ara*f*
80 (HMPREF9455_02480)	GH4185a	*D. gadei*	Endo-α-Ara*f*
DgGH$_{4185b}$ (HMPREF9455_02481)	GH4185b	*D. gadei*	Endo-α-Ara*f*

$BfGH_{182}$ (BACFIN_04787)	GH182	*B. finegoldii*	Exo-β-Gal*f*
$BfGH_{43\_31}$ (BACFIN_08810)	GH43_31	*B. finegoldii*	Endo-β-Gal*f*

Table 4.2 Concentrations of reagents and incubation conditions a. Reaction component concentrations used in lysis reaction unless stated otherwise. **b.** Incubation temperatures, time and speed of each incubation step used on thermal mixer.

a.

Reagent	Final Concentration
Lysozyme	45 mg ml^{-1}
Novel Enzyme	5 µM
Lipase	1 mg ml^{-1}

b.

Incubation	**Temperature (°C)**	**Time (minutes)**	**RPM**
Lysis incubation	37	120	2000
Proteinase K treatment	55	10	800

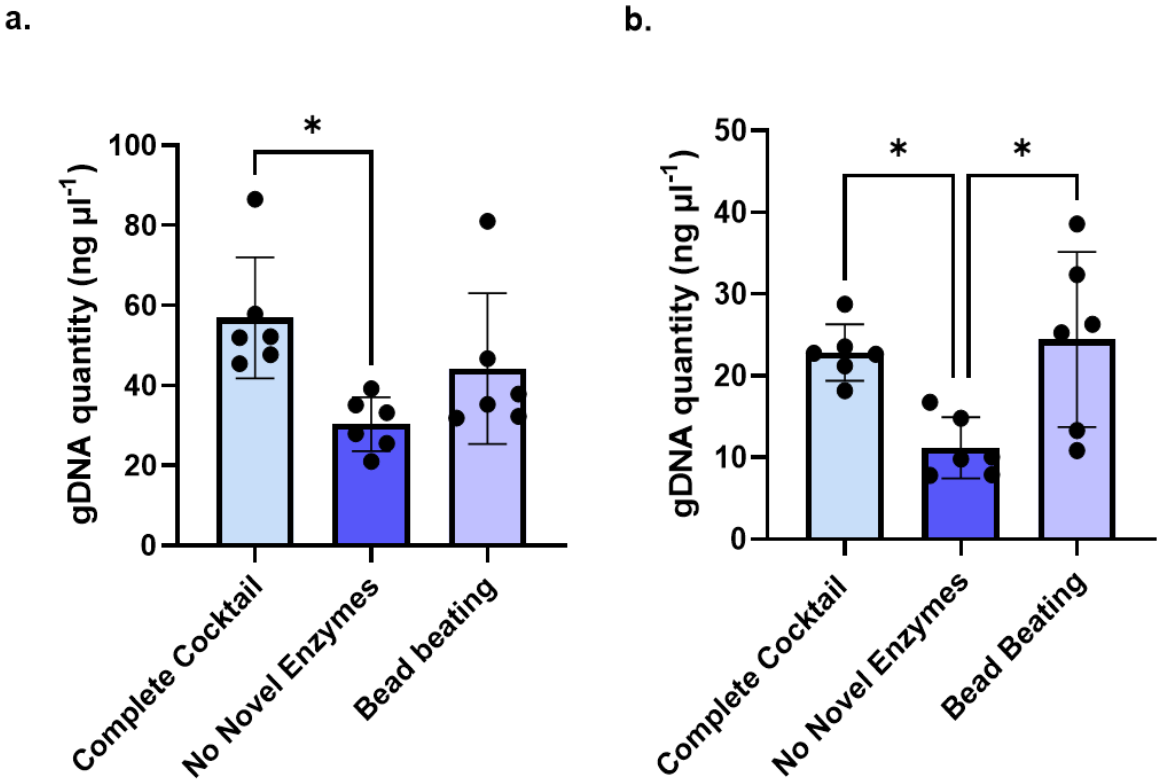


Figure 4.1 Normalised quantification of total gDNA. (**a**) *M. smegmatis* and (**b**) *M. abscessus* comparing yield of complete cocktail, no novel enzymes and bead beating extraction methods using GenElute™ Bacterial Genomic DNA Kits. Statistical significant denoted by * (P>0.05).

4.3.2.2 *Assessing DNA quality*

To gain an understanding of the quality of the genomic DNA, samples were diluted to approximately 5 ng μl^{-1}, 5 µl was mixed with 5 µl running buffer. The 10 µl samples were then ran was run on a 0.4% agarose gel at 40V for 2 hours. This allows determination of gDNA quality in samples, with the tighter and higher the band indicating the higher the quality of the DNA while a more 'smeared' band appearance denoting shearing as the gDNA fragments are of varying length and non-uniform. There was no observable difference between the dominant gDNA bands for both of the enzymatic lysis methods, with dominant bands being present between 20-48.5 kb. This is to be expected due to the lysis methods being the same, apart from the novel enzymes. Although some degradation of the gDNA is present in both of the enzymatic lysis methods this is not unusual, due to the use of spin columns for gDNA capture which is generally not specialised toward HMW DNA, so some degradation of the gDNA is expected, with centrifuging damaging the DNA, leading to maximal gDNA qualities less than 60 kb (MERCK, 2023). Both of the enzymatic methods produce clearly higher quality DNA than bead beating, as the beat-beaten samples have no clear band, indicative of sheared DNA with fragments predominantly lower than 15 kb. Although not quantitative, the gel does give a brief qualitative overview of the DNA integrity. This also shows evidence that even if the yield of gDNA from bead beating can be higher than our enzymatic protocol the quality of the gDNA is notably impacted.

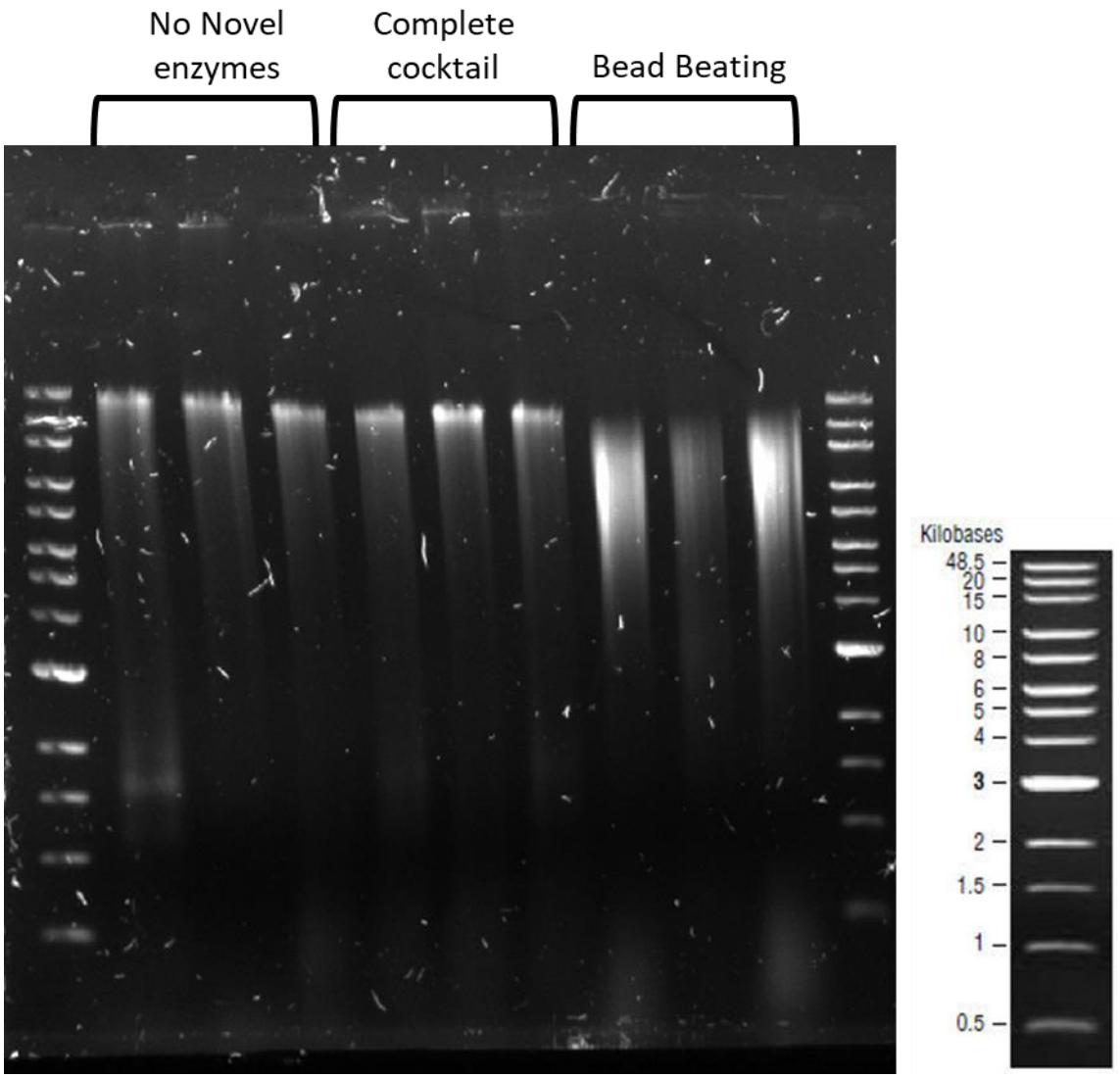


Figure 4.2 Qualitative agarose gel of *M. smegmatis* gDNA. 0.4% agarose gel run at 40 V for 2 hours stained with ethidium bromide. Extracted *M. smegmatis* gDNA from three different conditions in triplicate; no novel enzymes, complete cocktail and bead beating. Ladder is Quick-Load® 1 kb Extend DNA Ladder (NEB), see right for DNA sizes.

4.3.2.3 *Effects of Novel enzymes, Lysozyme and Lipase on mycobacterial cell wall glycans*

To understand the role of each of the novel enzymes in the breakdown of AG into its constituent monosaccharides, we incubated 2 mg ml^{-1} AG overnight with 1 µM of each enzyme (listed in **Table 4.1**) in a multi-enzyme cocktail and analysed by TLC the production of monosaccharide and oligosaccharides (**Figure 4.3**). To understand the contribution of individual enzymes, reactions were also conducted when a single enzyme had been removed from the cocktail. When AG is incubated with all six novel enzymes, two clear bands are visible corresponding to the arabinose and galactose standards, with no visible oligosaccharide bands. With the removal of DG02470, three spots appear that are not present when DG02470 is included. As discussed in **Chapter 2,** this is likely due to Dg02470 being an exo-β-Ara*f*-ase*,* required to hydrolyse the capping β-1,2-Ara*f* motifs in AG. As the other arabinose targeting enzymes all act on α-linkages, these will not be hydrolysed in the absence of DG02470, leaving undigested terminal fragments. When $DgGH_{172c}$ is removed, several oligosaccharide bands appear. As previously described in detail by Al-Jourani et al. (2023), the activity of $DgGH_{172c}$ is exo- α1,5-Ara*f*-ase. With no other exo-α-arabinofuranosidases, a range of oligosaccharides are left from the action of the endo arabinanases. $DgGH_{4185a}$ and $DgGH_{4185b}$ both have the same activity, endo-α-D-arabinanase, which would lead to the prediction that the removal of either enzyme would not have an overall effect, the removal of $DgGH_{4185a}$ having little impact on the breakdown of AG when compared to the complete cocktail. Interestingly, however the removal of $DgGH_{4185b}$ has a notable effect on the oligosaccharides. The removal of $BfGH_{182}$ clearly reduces the amount of galactose with an undigested oligosaccharide band clearly present. Finally, the removal of $BfGH_{43\_31}$ does not appear to produce any oligosaccharides with the results looking similar to those of All, indicating that $BfGH_{182}$ is able to cleave the long chain β(1-5), β(1-6) galactan without the need for an endo acting enzyme. Although the removal did not have a visible effect up on the TLC, we still included it for initial testing as results on TLC only show the break down from purified AG and the removal of the sole endo-glactatofuranosidase would only leave $BfGH_{182}$ and exo-galactofuraosidase which

would not show the production of any oligosaccharides and only produce the monosaccharides, therefore it may be that there are un cleaved residues that are left.

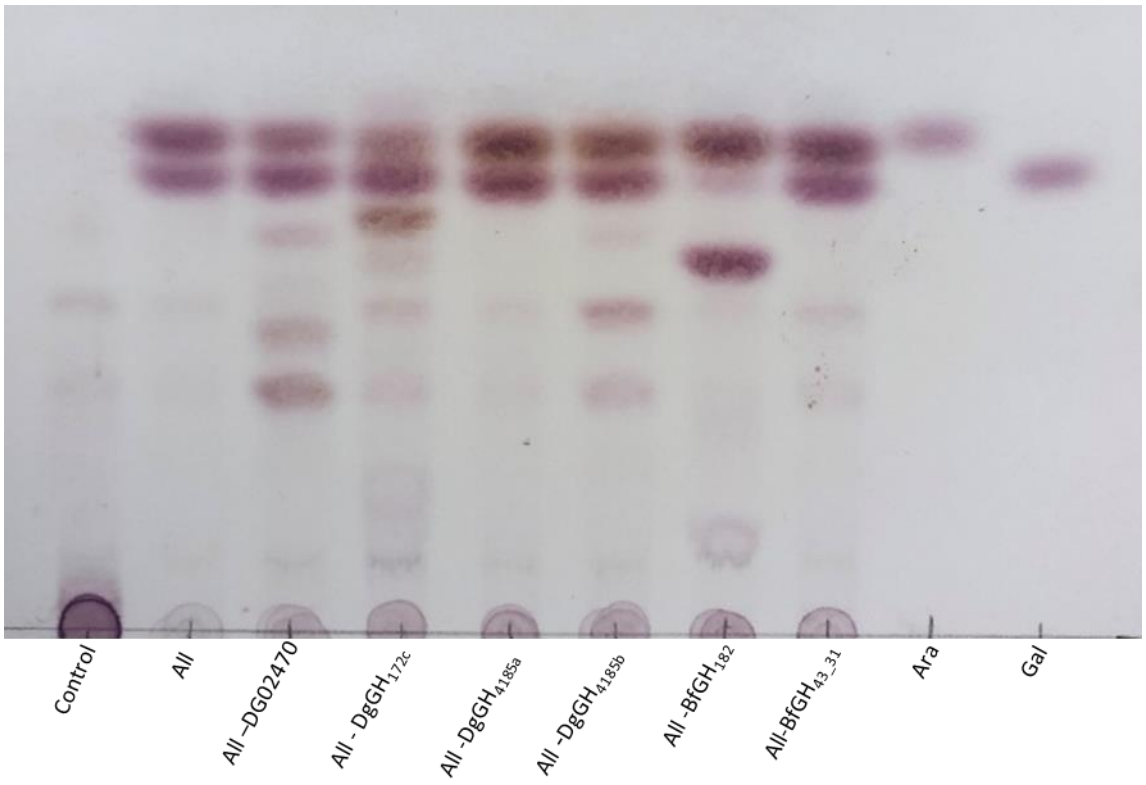


Figure 4.3 TLC analysis of Novel Enzymes vs arabinogalactan. TLC analysis of products generated by incubating 2 mg ml^{-1} AG with 1 µM of each novel enzyme and incubation with the removal of one of the enzymes. Analysing the variation in oligosaccharides with the removal of a novel enzyme

Although we had determined the effects of the novel enzymes on purified AG, we still did not have data for the breakdown of the cell wall glycans by lysozyme, lipase and the novel enzymes during the lysis step on whole mycobacterial cells. To analyse this the lysis step was performed as described in **Table 4.2** on *M. smegmatis* with combinations of the reaction components shown in **Figure 4.4.** After the incubation step the reaction was centrifuged at 11 000 x *g* for 5 minutes and the supernatant analysed by TLC (**Figure 4.4**). As shown in **Figure 4.4**, the addition of different reaction components alters the composition of glycan products in the supernatant. The addition of lysozyme produces a band corresponding to GlcNAc, which is not clearly present in the absence of lysozyme, indicating that the lysozyme is able to hydrolyse the PG in the cell wall of *M. smegmatis* even with the modified PG structure present in mycobacteria. When incubated with the novel enzymes, AG is broken down into arabinose and galactose, which are not present in the absence of the enzymes. Bands are present when *M. smegmatis* was incubated with lipase that are not present in the lysis buffer only control, however a standard was not available for the product of the action of lipase upon the mycolic acids. When added with lysozyme and the novel enzymes a broader band is present for arabinose, as the lipase hydrolyses the mycolic acids allowing for more arabinan to be accessible to the novel enzymes potentially due to the β-caps being more accessible for DG02470 to cleave. When Lipase, lysozyme and the novel enzymes are used in the incubation the GlcNAc, arabinose and galactose bands indicating a greater breakdown of the cell wall glycans.

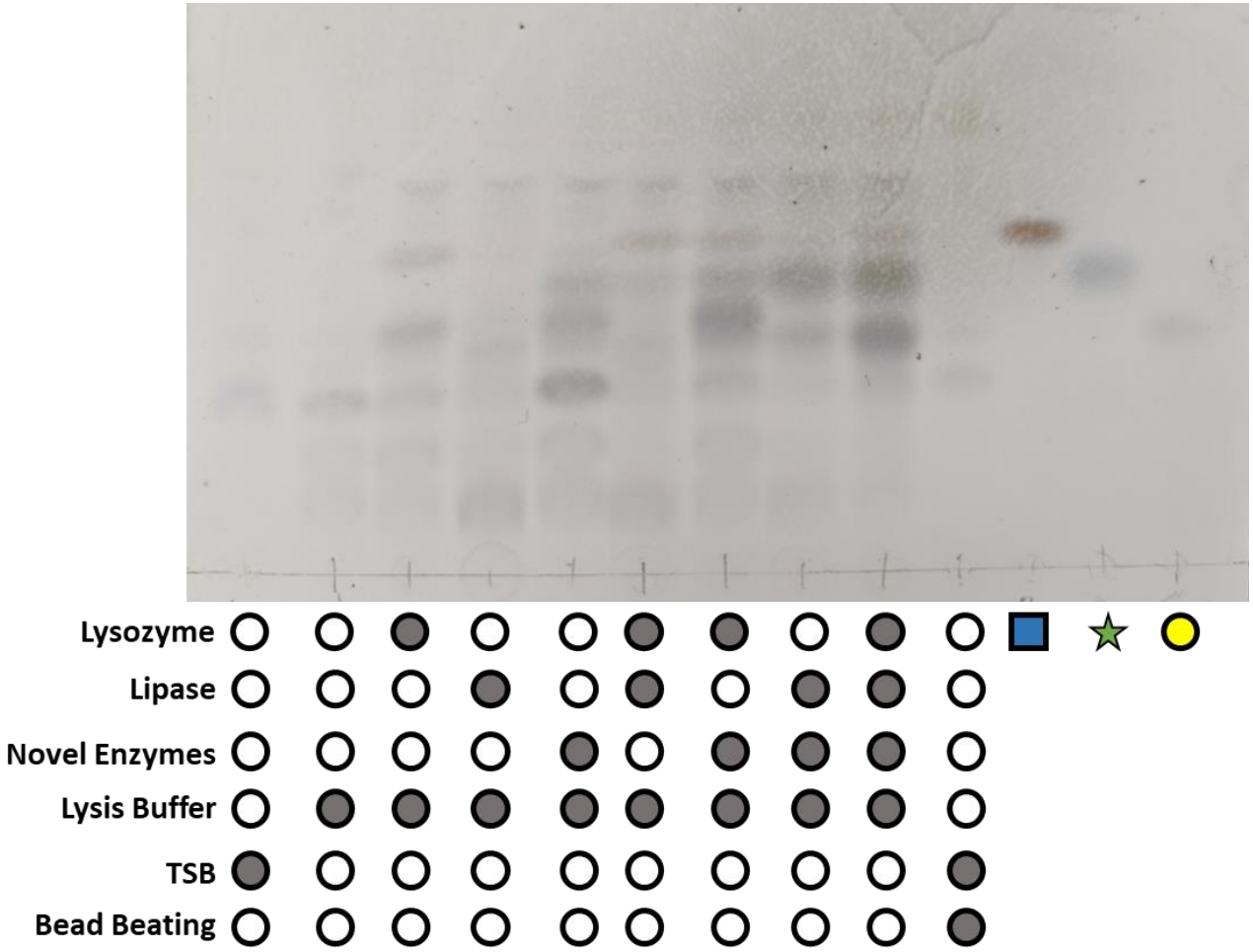


Figure 4.1 TLC analysis of the supernatant after 120 minutes incubation of the lysis step of the DNA extraction of *M. smegmatis* in varying conditions. Grey filled circles indicating presence of a given reaction component. Run with the standards GlcNAc (blue Square), arabinose (green star) and galactose (yellow circle). Run in 5:4:2:1 butanol: methanol: ammonium hydroxide: water. Stained with DPA.

4.3.2.4 *Removal of DgGH$_{4185a}$*

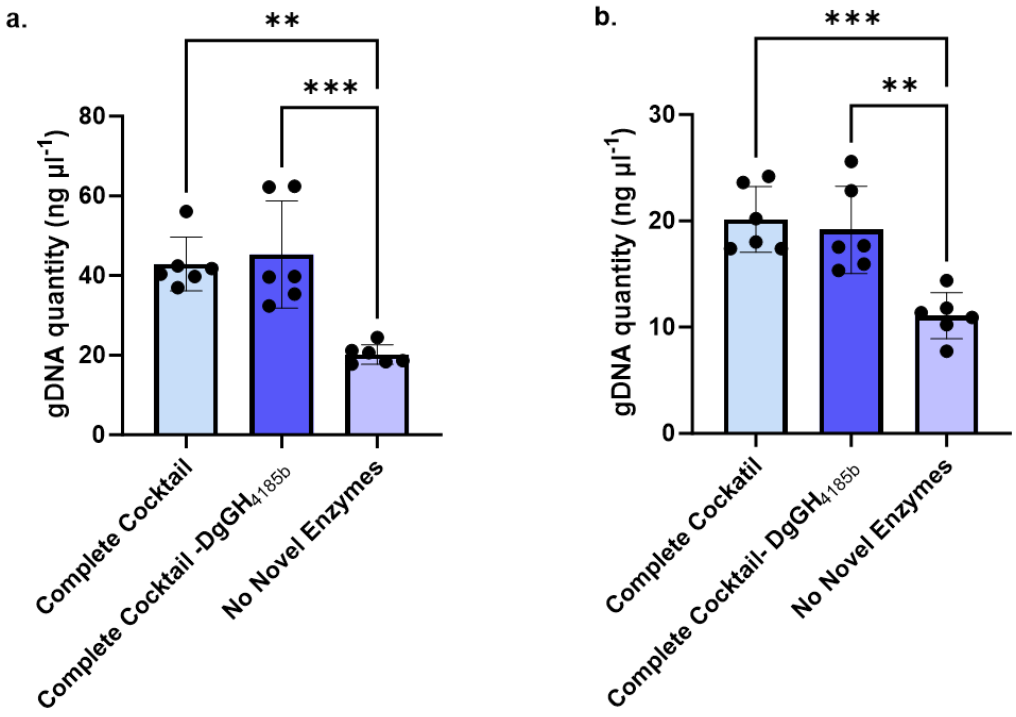


Figure 4.2 Normalised data of mycobacterial gDNA extraction with the removal of DgGH$_{4185a}$. (**a.**) *M. smegmatis* (**b.**) *M. abscessus* comparing the normalised results between the complete cocktail, with the removal of DgGH$_{4185a}$ and no novel enzymes. Significant difference denoted using * (P<0.05), ** (P<0.01), *** (P<0.0005).

Based on the analysis of the TLC in **Figure 4.3** and lack of effect DgGH$_{4185a}$ had upon the breakdown of AG, we tested to analyse if there was any effect upon yield of gDNA when DgGH$_{4185a}$ was removed from the cocktail. Shown in **Figure 4.5** are the results of the analysis, there was no reduction in yield for either *M. smegmatis* or *M. abscessus.* This concurs with the TLC results, likely due to the fact that DgGH$_{4185b}$ and DgGH$_{4185a}$ are both

endo-α--D-arabinase enzymes and therefore fulfil the same role in the cocktail. Due to these results $DgGH_{4185a}$ was removed from the cocktail going forward, increasing the simplicity of the purifications and protocol.

4.3.1.1 Impact on yield of individual novel enzymes.

To analyse the effect upon the overall impact on yield the removal of an individual enzyme has, we removed one enzyme from the cocktail at a time to see if it impacted the yield. The results shown in **Figure 4.6** show a variation across *M. smegmatis* and *M. abscessus*, as to the effect each enzyme has upon yield. In the case of *M. smegmatis*, removal of any single enzyme does not significantly reduce the yield when compared to the complete cocktail. With the removal of $DgGH_{172c}$ and $DgGH_{4185b}$ individually, the yield is still significantly higher than without any novel enzymes. The data suggests that $DgGH_{172c}$ and $DgGH_{4185b}$ play a lesser role in the cocktail when applied to *M. smegmatis.* However, when tested again *M. abscessus* removing any one of the enzymes significantly reduces the yield of gDNA compared to the complete cocktail. Due to the variation in yield between species, with the removal of any one enzyme, going forward no single enzyme was removed, as the objective was to provide a protocol that can cover a wide variety of species.

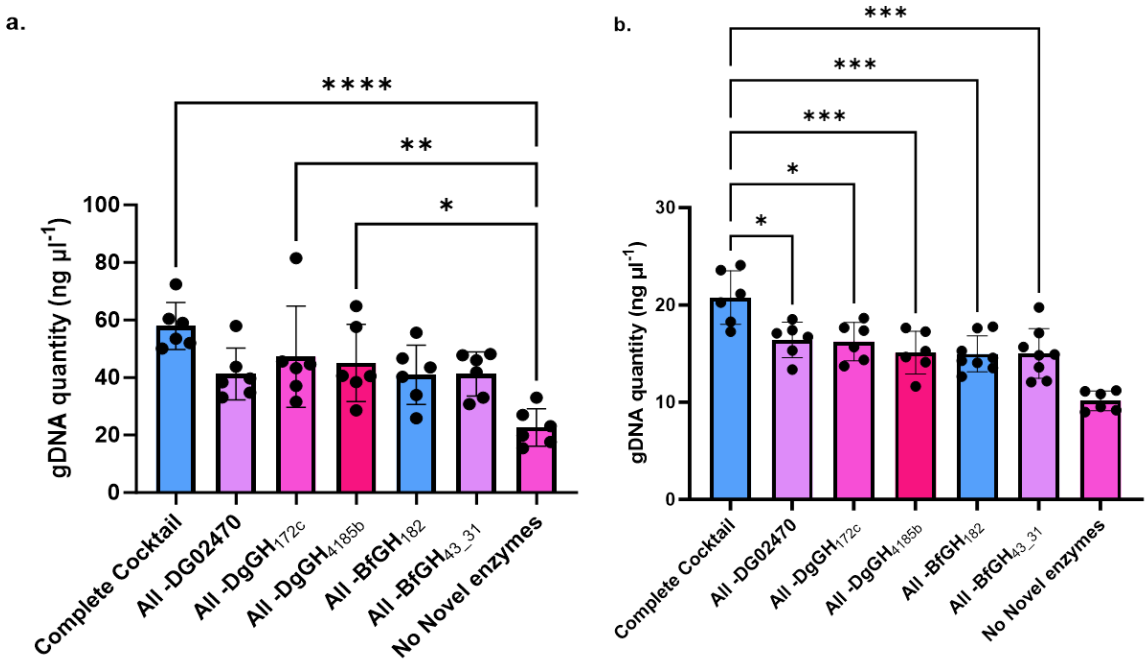


Figure 4.3 Normalised data of mycobacterial DNA extraction with the removal of single enzymes. (**a**.) *M. smegmatis* and (**b.**) *M. abscessus* all extractions with novel enzymes are

significantly higher than no novel enzymes for M. abscessus however for clarity of data representation these statistics are not shown. Significant difference denoted using * (P<0.05), ** (P<0.01), *** (P<0.0005), **** (P<0.0001).

4.3.2.5 *Time point assay*

Although two-hour incubation showed significant lysis, shorter protocol times are always desirable. To test the effect of changing incubation time, same conditions described in **Table 4.2** were used, altering the enzymatic lysis incubation times to 15, 60 and 120 minutes. No novel enzyme incubation was maintained at 120 minutes, the results are shown in **Figure 4.7.** For both *M. smegmatis* and *M. abscessus* incubation time has a significant effect on gDNA yield. In the case of both species there was no significant difference in the yield of gDNA between a 15 minute incubation with the novel enzymes and 120 minute incubation without the addition of the novel enzymes, indicating if needed a short incubation can be used, however this does have the possibility to impact downstream uses due to the lower yield, although for purposes such as PCR, 15 minutes will suffice. Altering the incubation time for *M. smegmatis* from 60 to 120 minutes significantly increased the yield. However, this was not the case for *M. abscessus* where there was no significant increase in the gDNA yield when increasing incubation time from 60 minutes to 120 minutes although the mean yield does increase. Going forward two hour incubations were used for this project due to the increased yield and to ensure the incubation times between control and novel enzyme incubation were consistent.

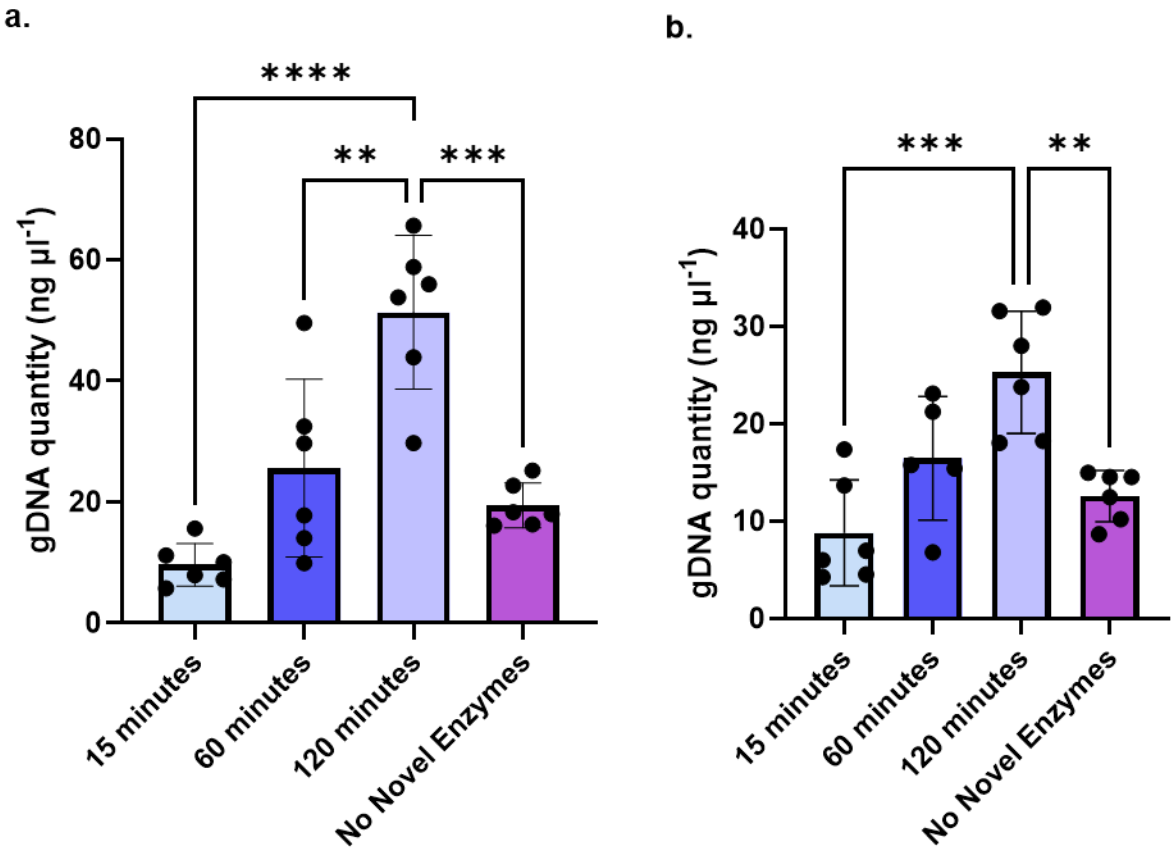


Figure 4.4 Normalised qPCR analysis of DNA yield with varying lysis step incubation times for (**a.**) *M. smegmatis* and (**b**.) *M. abscessus*. Significant difference denoted using) * (P<0.05), ** (P<0.01), *** (P<0.0005), **** (P<0.0001).

4.3.2.6 *Analysis of efficacy of reaction components.*

To determine the effectiveness of the individual reaction components in addition to the arabinogalactan degrading enzymes, we tested the efficacy of lysozyme, lipase and the novel enzymes individually when incubated for 120 minutes in lysis buffer (**Figure 4.8)**, the concentrations and conditions of each were as described in **Table 4.2.** For both *M. smegmatis* and *M. abscessus* the largest increase in yield of DNA when an additional reaction component is added to the lysis buffer is lysozyme, significantly increasing the yield respectively when compared to lysis with only lysis buffer. This was expected due to the importance of PG in the maintenance of the cell wall integrity and while lysozyme has a reduced efficacy against mycobacteria (Kanetsuna, 1980) using a specific PG-endopeptidase such as RipA could be important to increase yield significantly at lower concentration (Healy et al., 2020). In *M. abscessus,* the use of lysozyme as the sole enzyme increases yield of DNA significantly when compared to the addition of lipase or novel enzymes alone. The addition of lipase or novel enzymes alone did not increase the yield of gDNA significantly when compared to lysis buffer for both species tested. All tested individual enzymes were still significantly lower in yield for both species when compared to the complete cocktail, suggesting degradation of all three layers of mAGP is necessary for total lysis.

To potentially simplify the protocol, we tested the impact the removal of lysozyme, lipase and novel enzymes have upon the gDNA yield when compared to each other and the complete cocktail. The results (**Figure 4.9**) show that the removal of any one reaction component significantly affects the yield of gDNA when compared to complete cocktail for both *M. smegmatis* and *M. abscessus.* The removal of a single component has similar effects to the yield of gDNA extracted for *M. smegmatis.* The removal of lysozyme significantly reduced the yield gDNA from *M. abscessus* compared to the extraction using lysozyme and novel enzyme components. With the overall decrease in yield with the removal of individual components it was deemed that each component played an

important role in extraction of gDNA, however when combined with the data in **Figure 4.8** it shows that though each component is important to the overall lysis, individually they have little effect on the yield of gDNA when compared to an absence of any enzyme.

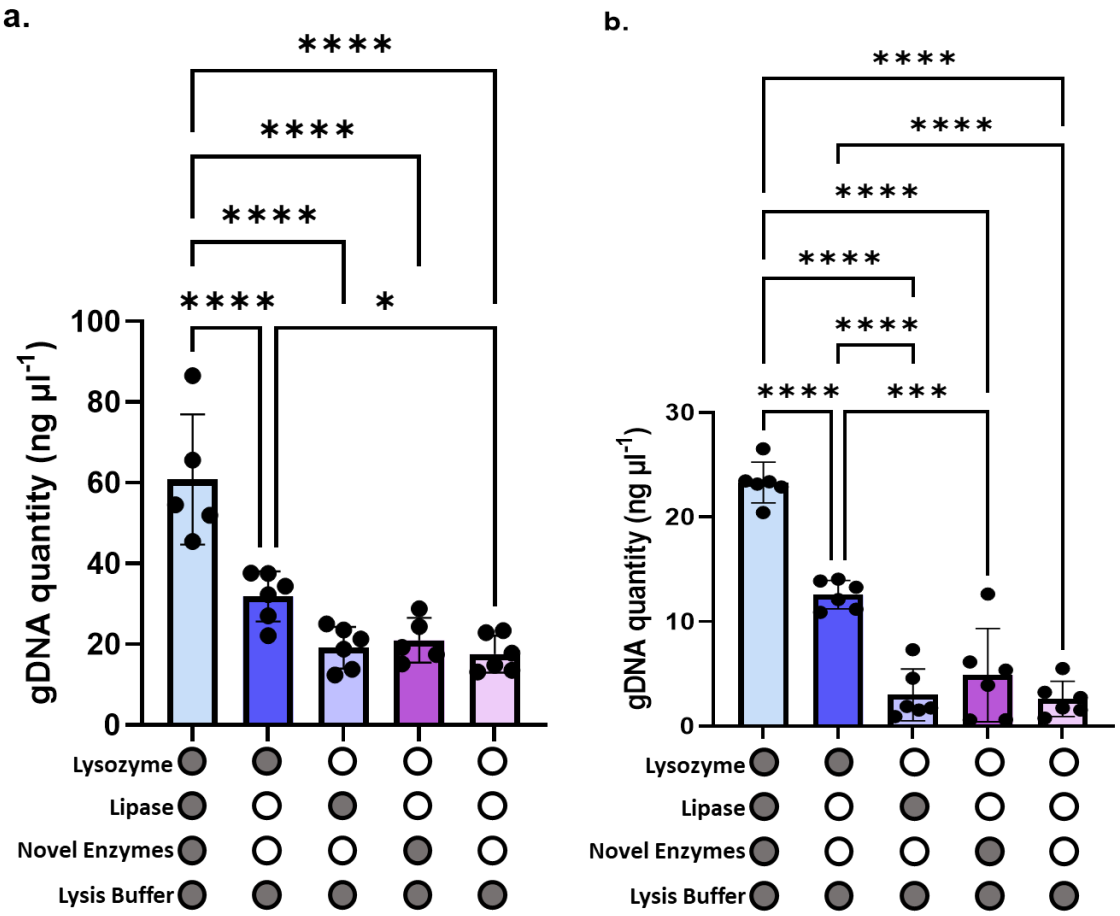


Figure 4.5 Normalised gDNA extraction using one reaction component. (a.) *M. smegmatis* and **(b.)** M. *abscessus* DNA extractions using one reaction component during the lysis step with lysis buffer. Filled in circles denote the presence of reaction component. Significant difference denoted using * (P<0.05), ** (P<0.01), *** (P<0.0005), **** (P<0.0001).

4.3.2.7 *Lysozyme concentration*

Finally, we tested how changing the lysozyme concentration would affect the yield of gDNA. We therefore compared the yield when reducing the concentration of lysozyme from 45 mg ml^{-1} to 20 mg ml^{-1} in both the presence and absence of novel enzymes. The lipase concentration was maintained at 10 mg ml^{-1}. The results shown in **Figure 4.10** show that the yield of gDNA extracted from *M. abscessus* is lysozyme dependant with yield increasing significantly when the concentration is increased from 20 mg ml^{-1} to 45 mg ml^{-1} both with and without novel enzymes. Interestingly the yield from 20 mg ml^{-1} with novel enzymes is comparable to 45 mg ml^{-1} Lysozyme without novel enzymes. The increase yield obtained from *M. abscessus* at 45 mg ml^{-1} compared to 20 mg ml^{-1} is potentially due to the previously mentioned variations in the cell wall when compared to *M. abscessus.*

Conversely *M. smegmatis* gDNA yields did not increase when lysozyme concentration was increased from 20 mg ml^{-1} to 45 mg ml^{-1} in the absence of novel enzymes. When at a concentration of 20 mg ml^{-1} lysozyme with no novel enzyme yield is compared to that of 20 mg ml^{-1} lysozyme with novel enzymes the yield significantly increases. There is however no significant difference in yield between 20 mg ml^{-1} with novel enzymes and 45 mg ml^{-1} without novel enzymes.

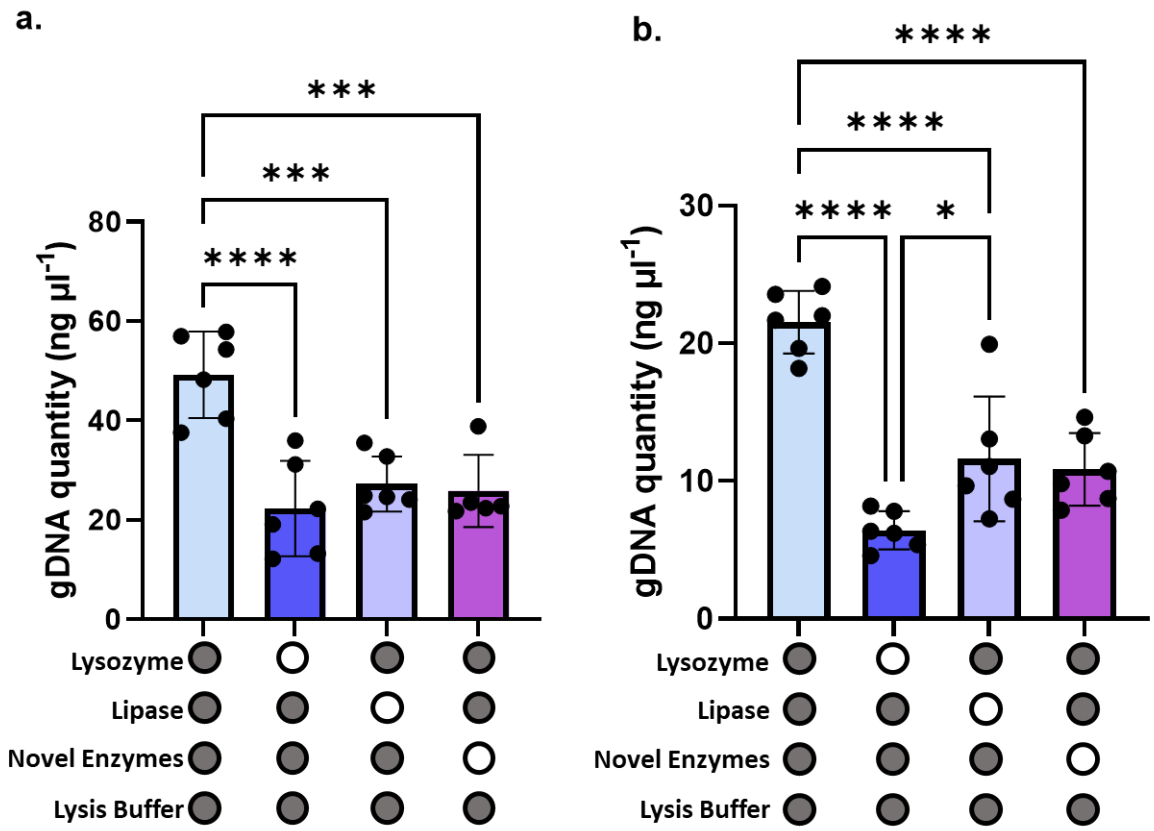


Figure 4.9 Normalised DNA extraction of using two reaction components. **(a)** *M. smegmatis* and **(b)** *M. abscessus* when removing lysozyme, lipase and the novel enzymes individually under the conditions described in **Table 4.2. Si**gnificant difference denoted using * (P<0.05), ** (P<0.005), *** (P<0.0001).

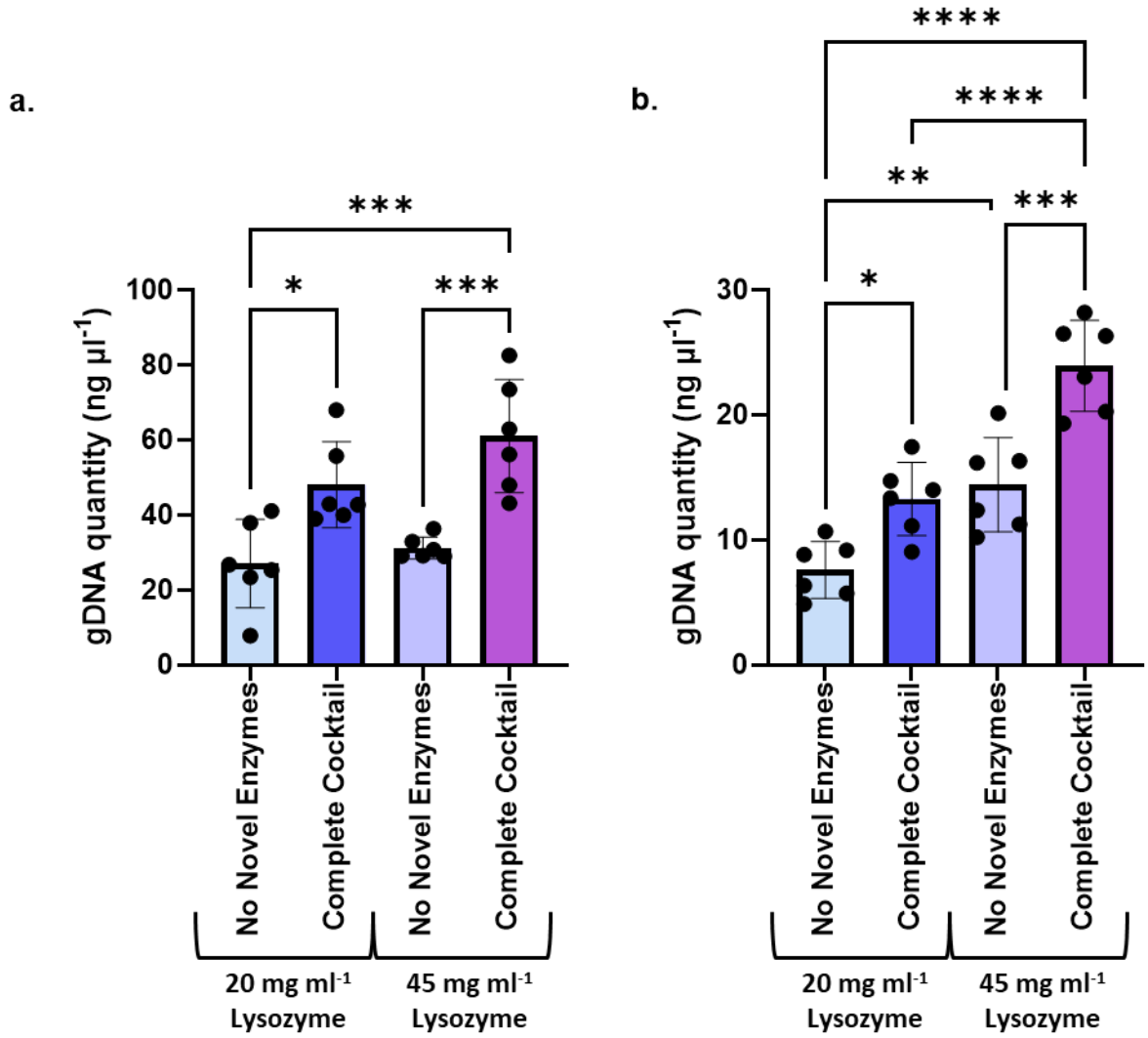


Figure 4.10 Normalised results for the extraction of gDNA with changing lysozyme concentration. From (**a.**) *M. smegmatis* and (**b.**) *M. abscessus*. Comparing two lysozyme concentration 20 mg ml^{-1} and 45 mg ml^{-1} with and without the novel enzymes. Statistical significant denoted by * (P<0.05), ** (P<0.01), *** (P<0.0005), **** (P<0.0001).

4.3.2.8 *Overall extraction results*

Figure 4.11 shows comparative analysis of all data points from the three standard conditions used in every analysis for both *M. smegmatis* and *M. abscessus.* The addition of novel enzymes to the lysis step significantly increases the yield of gDNA when compared to the absence of novel enzymes and also bead beating. Although bead beating has the potential to have a higher yield as shown in **Figure 4.2,** this quality of this DNA is lower, reducing the use for downstream analysis especially in long-read sequencing which is hindered by the sheared DNA obtained through bead beating.

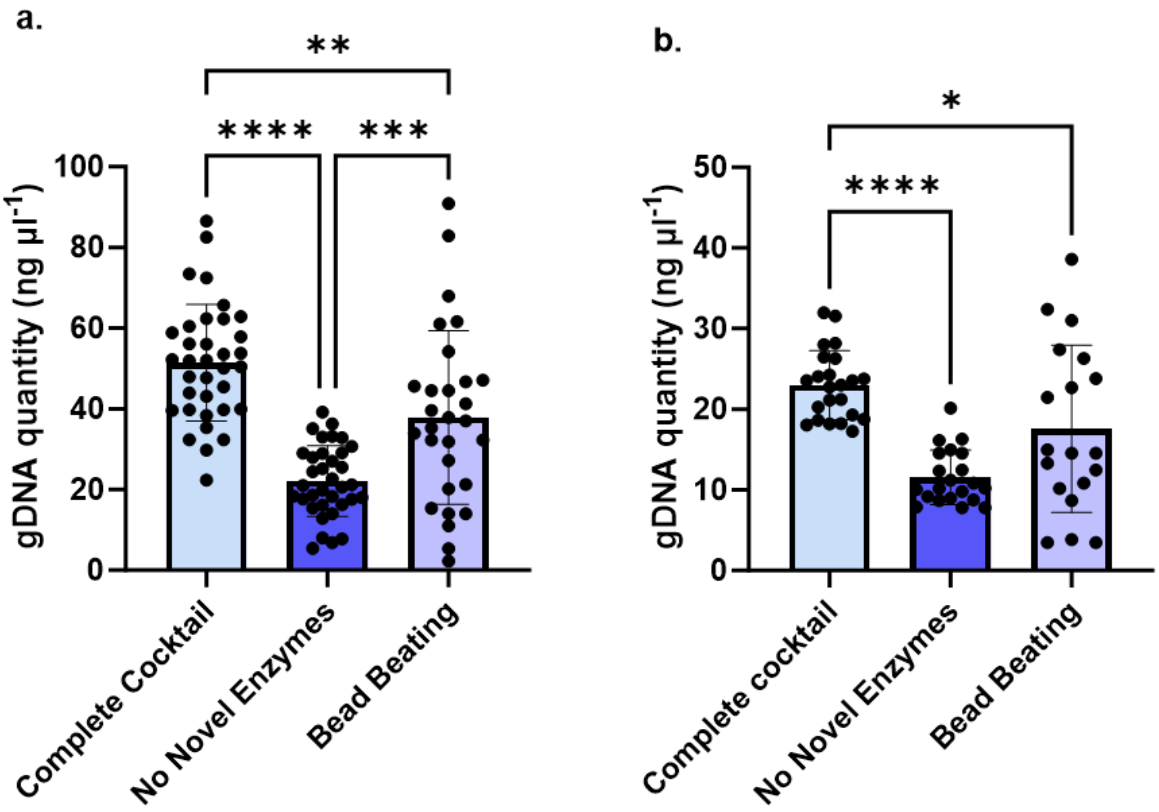


Figure 4.11 Overall normalised DNA extraction for complete cocktail, No Novel enzyme and Bead beating. (**a.**) *M. smegmatis* and (**b.**) *M. abscessus* data for all point across all normalised assays. Statistical difference denoted by * (P<0.05), ** (P<0.01), *** (P<0.0005), **** (P<0.0001).

4.3.2.9 *Freeze drying enzymes.*

Lyophilisation or more colloquially known as freeze-drying is the removal of water by freezing. This is done in 3 stages, the first of which is making the solid matrix by freezing the suspension via rapid methods such as liquid nitrogen. This Is followed by sublimation of the water molecules, the third and final stage is secondary drying where water molecules are removed until desired dryness is achieved (Baolin Liu & Xinli Zhou, 2014).

Lyophilisation is used across a multitude of industries to stabilise compounds allowing for long term storage. The removal of the water reduces ways which active compounds can degrade. The absence of water reduces the risk of oxidation and hydrolysis that can happen to the compounds. This has enabled the widespread availability of pharmaceutically active molecules such as vaccines and medication, as well as more day-to-day compounds such as the food industry and media in labs (Assegehegn et al., 2019; Ward & Matejtschuk, 2021).

Although lyophilisation has enabled a wide variety of storage solutions it is none the less a complex process in which many factors must be accounted for during the lyophilisation process, such as the rapid changed in pH, solute concentration and potentially salt concentrations. This is due to the reduction in volume due to evaporation of water. Additionally other changes may occur such as crystallisation that may irreversibly damage the protein. To avoid this, stabilisers are often used, this often comes in the form of sugars, amino acids and polymers (Bhatnagar & Tchessalov, 2020). Although this still requires a lot of optimisations when selecting the optimal stabiliser to ensure factors such as the glass transition temperature (T_g) do not affect the stored proteins, as a T_g higher than the stored temperature as an increase beyond the Tg causes the protein to misfold and cause chemical degradation (Carpenter et al., 2002; Roy et al., 1991). Of the sugars the most commonly used in stabilisation are sucrose and trehalose, due to their non-crystalising nature (Starciuc et al., 2020). Additionally, the glass transition temperature (T_g) of trehalose and sucrose is 106^{o}C and 60^{o}C respectively making them ideal for long term storage (Roe & Labuza, 2005).

To increase the ease of storage and potential marketability of the cocktail, lyophilisation of the enzymes was performed to test whether the enzymes are still active after storage as a powder. As previously described in **Section 1.5**, lyophilisation requires a volatile buffer containing no salt and potentially a stabiliser to aid in resuspension. The sugars selected as stabilisers were sucrose and trehalose, commonly used in protein lyophilisation as described in **Section 1.5**, in a 1:5 protein: sugar ratio. The samples were then dialysed into 100 mM ammonium acetate buffer pH 7 that evaporates when lyophilised. The proteins were lyophilised as described in **section 2.2.4.**

After resuspension in **Buffer A** each of the treatments were incubated over night at 1 µM with 2 mg ml^{-1} AG and spotted onto TLC to ensure the activity of the enzymes was still present and comparable to freshly prepared enzymes after lyophilisation. The TLC shown in **Figure 4.12** shows that the use of no stabiliser and trehalose have no observable impact upon the efficacy of the enzymes when breaking down AG into arabinose and galactose. However, the use of sucrose as a stabiliser appears to affect the efficacy of the cocktail, with additional oligosaccharide bands being present that are not present in any other treatment this may be a result of the sucrose acting as a inhibitor to the enzyme restricting the active site active site.

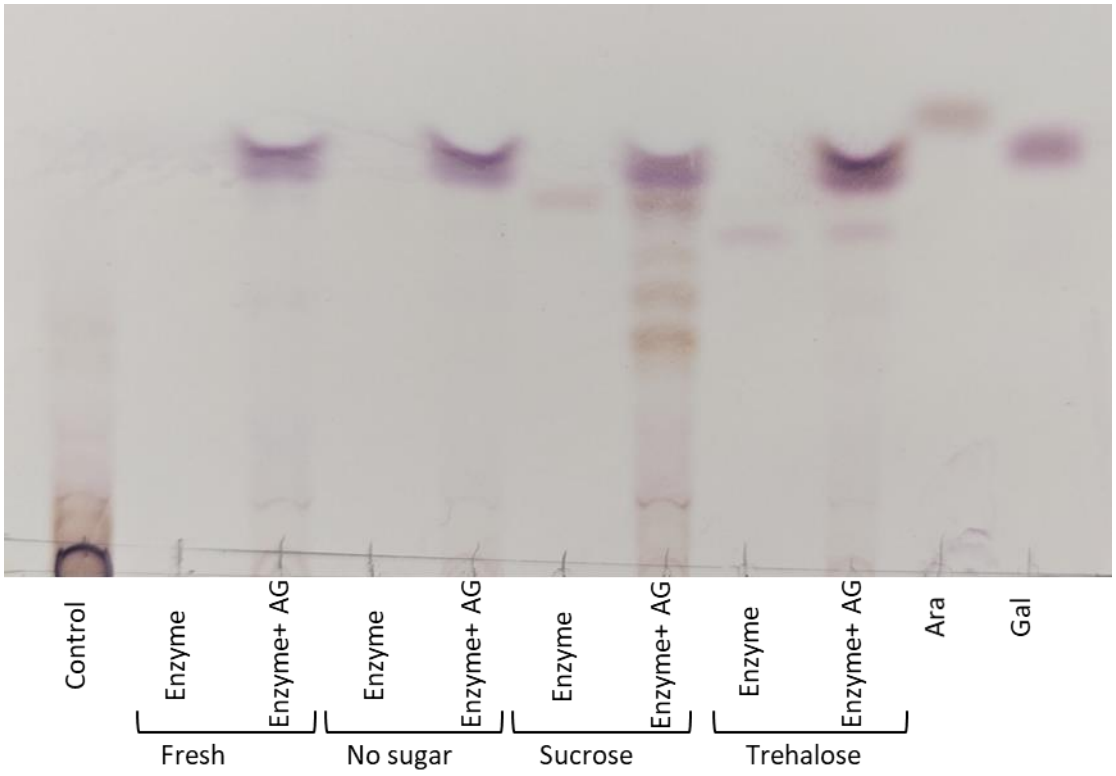


Figure 4.12 TLC analysis of lyophilised enzymes against AG. Lyophilised enzymes resuspended in Buffer A with stabilisers: no sugar, sucrose and trehalose compared to fresh enzyme when incubated at 1 µM overnight with 2 mg ml^{-1} AG. Enzyme controls ran only the resuspended enzymes and buffer A. Stained with orcinol.

4.3.2.9.1 DNA extraction

The suspended enzymes were then tested to see their efficacy for DNA extraction, when compared to freshly prepared enzymes, across the different lyophilisation conditions. The conditions used were as described in **Table 4.2.** The results of the gDNA extractions for *M. smegmatis* and *M. abscessus* are shown in **Figure 4.13.** The freeze drying of the proteins both in the presence of a sugar stabiliser, and in the absence of one significantly decreased the yield of gDNA when compared to fresh enzymes. Yields from both no sugar and trehalose are still significantly higher than no novel enzyme in the case of *M. smegmatis.* However, with *M. abscessus* the yield from lyophilised proteins is all significantly lower than with fresh enzymes. These initial results are promising and suggest

lyophilisation may be a possibility however more testing would be needed to make it more comparable to freshly prepared enzymes.

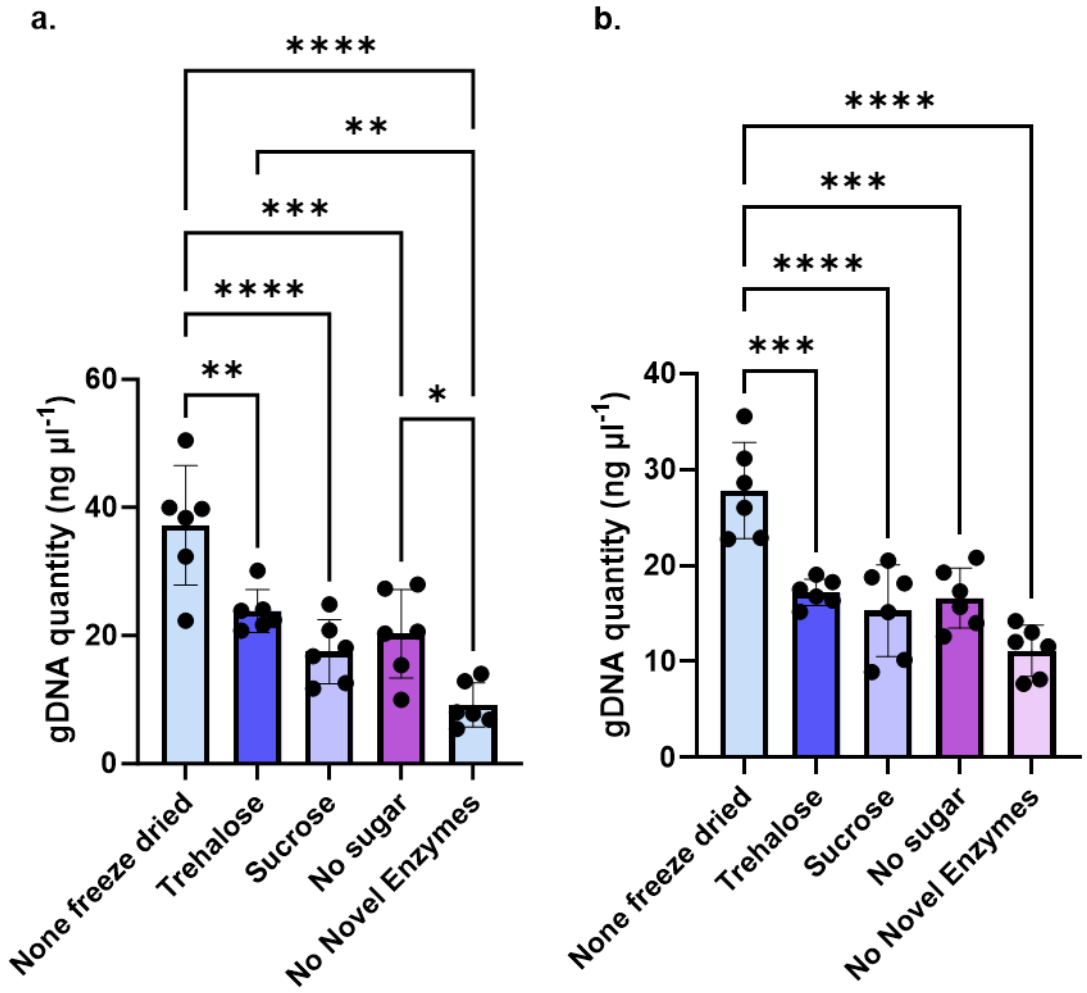


Figure 4.13 Normalised DNA extractions using lyophilised enzymes. (**a.**) *M. smegmatis* and (**b.**) *M. abscessus* comparing the DNA yield from lyophilised enzymes using trehalose, sucrose and no stabiliser (no sugar), with freshly prepared enzymes and no novel enzymes. Statistical significance denoted by * (P<0.05), ** (P<0.01), *** (P<0.0005), ****(P<0.0001).

4.3.3 Additional species

4.3.3.1 *Mycobacterial species.*

To analyse the efficacy across other mycobacterial species *M. avium* sp. *paratuberculosis* and *M. bovis* BCG were used, *M. avium* due to its clinical and BCG due to both its clinical significance and its comparableness to Mtb. The data shows that adding the enzymes significantly increases the yield of gDNA for both *M. avium* and BCG. The addition of novel enzymes also significantly increase the yield when compared to bead beating for *M. avium* subsp *paratuberculosis,* however there is no significant change between the complete cocktail treatment and bead beating with *M. bovis* BCG. It should also be noted that the yields from *M. bovis* BCG are very low and without a large initial pellet, downstream sequencing is unlikely to work as well with other mycobacterial species tested, but it is sufficient to perform multiplex PCR with short amplicon sizes as shown in **Figure 4.15**.

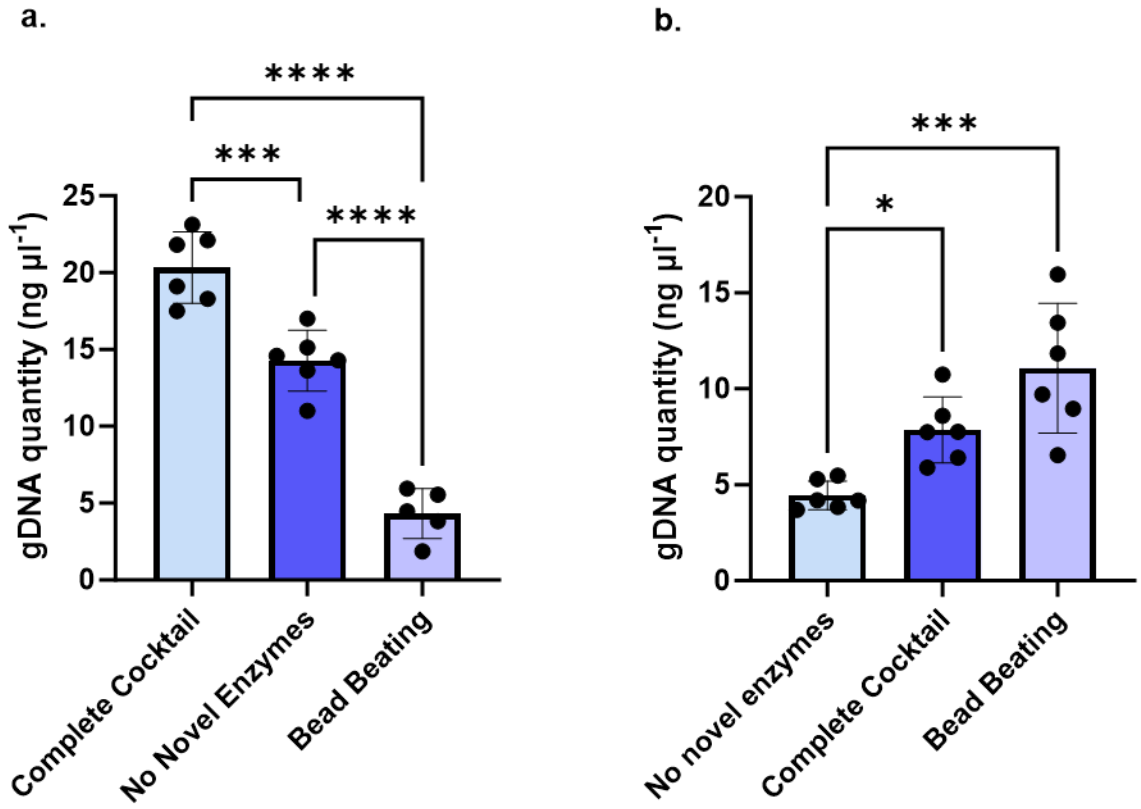


Figure 4.14 Qubit results for the gDNA extraction of additional mycobacterial species. (**a.**) ***M. avium*** subp. *paratuberculosis* and (**b**) *M. bovis* BCG. Comparing the results using 10 mg ml^{-1} lipase, 45 mg ml^{-1} lysozyme without novel enzymes, with novel enzymes and bead beating. Significant difference denoted * (P<0.05), ** (P<0.01), *** (P<0.0005), **** (P<0.0001)

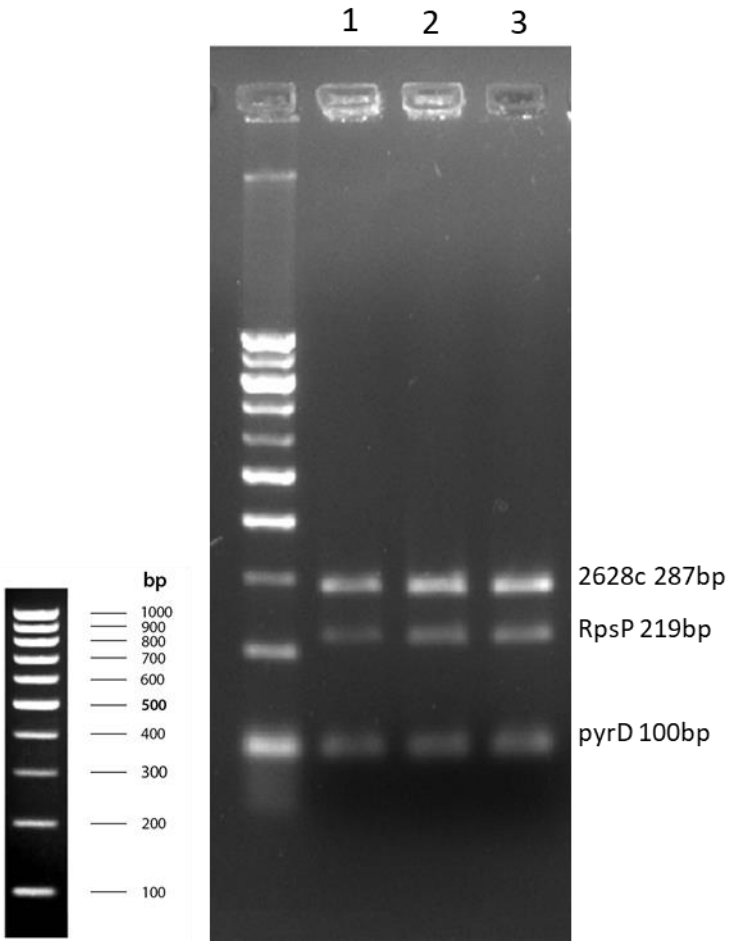


Figure 4.15 Agarose gel of multiplex PCR of BCG gDNA. 2% agarose gel electrophoresis of multiplex PCR of *M. bovis* BCG extracted with different conditions: 1. No Novel Enzymes. 2. Complete cocktail 3. Bead beating.

4.3.3.2 *Nocardia farcinica*

Finally, we tested the cocktail on another member of the Mycobacteriales order: *Nocardia farcinica* containing the same AG domain as mycobacteria (Daffe et al., 1993). The results shown in **Figure 3.16** show the results from this experiment. The addition of the novel enzymes significantly increases the yield when compared to no novel enzymes. Although the addition of novel enzymes has no significant difference in yield when compared to bead beating, although as previously discussed and shown this DNA will be more damaged, although bead beating will suffice for less powerful tools such as small amplicon PCR.

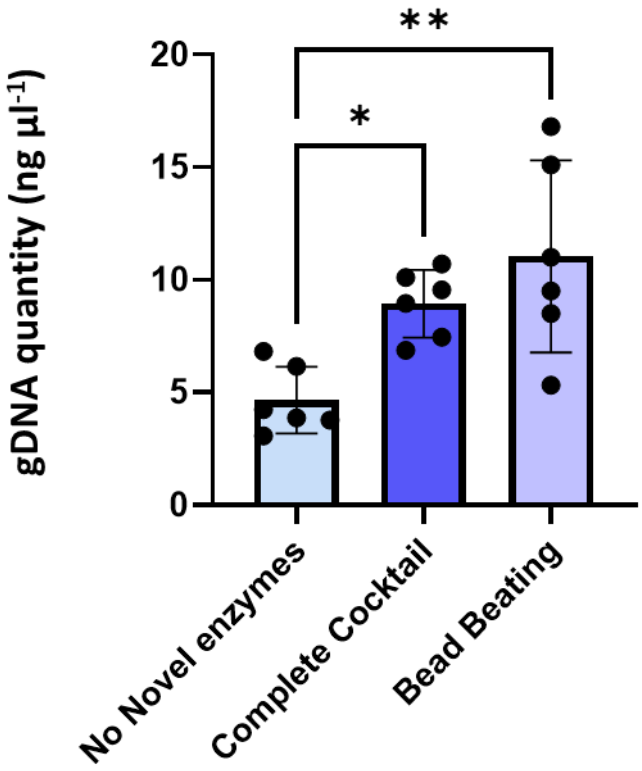


Figure 4.16 Qubit data from gDNA extraction from *N. farcinica*. Qubit results for the gDNA extraction from *N. farcinica* using No Novel enzymes, complete cocktail and bead beating. Significant difference denoted by * (P<0.05), ** (P<0.01).

4.4 Discussion

In this chapter we have provided a protocol adapting the commercially available GenElute™ Bacterial Genomic DNA Kit (MERCK) spin columns using novel enzymes identified by Al-Jourani et al. (2023) and characterised in **Chapter 3** which provides significantly increased yield of gDNA when used on *M. smegmatis* and *M. abscessus,* as well as higher quality DNA than obtained via bead beating. Additionally, preliminary data on the efficacy of the protocol on *M. bovis* BCG, *M. avium* subsp. *paratuberculosis* and *N. farcinica.*

In **Figure 4.4** we showed that the enzymatic cocktail used is shown to break down the PG and AG into its constituent elements of GlcNAc, arabinose and galactose, and mycolic acids potentially being cleaved by lipase, however additional work would be needed to verify this. We hypothesise that the additional of the lipase and novel enzymes aid in degradation of the cell wall, allowing greater access to the PG for lysozyme to act and therefore increased availability of gDNA once the secondary lysis buffer is introduced in the proteinase K digestion step. It may also increase the solubility of the individual component of mAGP, as the mAGP complex together is insoluble (Brennan, 2003) allowing for greater yield through the spin column.

The final combination of novel enzymes (DG0270, $DgGH_{172c}$, $DgGH_{4185b}$, $BfGH_{182}$, $BfGH_{43\_31}$) allow for a complete breakdown of AG with DG02470, $DgGH_{172c}$ and $DgGH_{4185b}$ cleaving the Ara*f* domain and $BfGH_{182}$ and $BfGH_{43\_31}$ cleaving the Gal*f* domain. As described by Al-Jourani et al. (2023), $DgGH_{4185b}$ and $DgGH_{4185a}$ are GH183 enzymes with overlapping endo-arabinanase activity, therefore only one would be required in the cocktail, which we verified in assays and in gDNA extractions. This reduced the number of enzymes in the cocktail from six to five, enabling easier production.

We showed that not only did the yield of gDNA increase when cultures are incubated for 120 minutes with novel enzymes compared to 120 minutes with no novel enzymes, we also showed that a 15 minute incubation with novel enzymes provides the same yield

from both *M. smegmatis* and *M. abscessus* as 120 minute incubation without novel enzymes. Although there is no significant change in in yield between 60 and 120 minutes for *M. abscessus,* the mean yield does increase, and as the protocol is aimed at a broad range of species, 120 minutes was selected for testing. This does however allow for flexibility in the protocol, e.g., with a large cell pellet incubation time could be reduced to still yield sufficient DNA or reduced if DNA is only need for sensitive assays such as PCR which requires a much lower yield.

Interestingly, from the study it has shown that lysozyme concentration impacts the yield of *M. abscessus* gDNA*,* however the reduction from 45 mg ml^{-1} to 20mg ml^{-1} did not impact the yield from *M. smegmatis* significantly, although we cannot be certain of the exact reason. It may be due to the GLPs present in the cell wall of *M abscessus* (**Section 1.2.3.1.1**) affecting enzyme access**.** This is also in line with previous studies which show *M. smegmatis* being more susceptible to lysozyme than other mycobacterial species (Gordon & Barnett, 1977; Kanetsuna, 1980). This would also explain the difference in gDNA yield obtained throughout the study as *M. abscessus* genome is 5 Mb while *M. smegmatis* 6.9 Mb, therefore you would expect the yield of gDNA to be ~1.4x higher from *M smegmatis* compared to *M. abscessus,* when using equally weighted pellets, however our results show a mean increase between *M. smegmatis* and *M. abscessus* was ~2x higher. Suggesting the general increased permeability of the cell wall leads to a higher yield of gDNA from *M. smegmatis.*

It is possible that using a mechanical disruption step along with an enzymatic lysis step would increase the overall yield of gDNA during extractions, however the implications of doing this would impact the quality of the DNA drastically. Although this may aid sensitivity if the sample is of low bacterial load.

Previous literature regarding DNA extraction from NTMs has involved phenol and/or bead beating (Bouso & Planet, 2019; Käser et al., 2010). Phenol is a toxic chemical and the latter requiring an expensive piece of equipment not necessarily accessible to all labs. The results we obtained from our protocol are comparable with many of those protocols such as Bouso & Planet (2019) which included both a step of bead beating. Their optimised extraction gave a 4.17 µg total mean from a washed normalised cell weight of 26.4 mg, while for *M. abscessus* the overall average we obtained was 20 ng $µl^{-1}$ when normalised

at 50 mg, which when converted to the same normalisation equates to 2.1 µg total for *M. abscessus* and 5.5 ng for *M. abscessus.* Epperson & Strong (2020) do not use phenol or bead beating, instead using CTAB in their extraction they obtained between 100 ng and 10 µg although washed weights of bacterial are not provided so direct comparison is difficult.

Sucrose and trehalose were selected as lyophilisation stabilisation sugars due to the ability of reducing sugars to resist crystallisation and are widely used in pharmaceuticals as lypo-protectants (**Section 1.5**) (Mensink et al., 2017; Sundaramurthi & Suryanarayanan, 2010). Through initial TLC analysis we were able to determine that the use of trehalose and no sugar appeared to impact the enzymes within the cocktail the least when resuspended in **Buffer A**. Sucrose appeared to interfere with the activity of the enzymes, it is unclear as to the reasoning for that and we were not able to determine the cause. This finding was also corroborated by the gDNA extraction data for both *M. abscessus* and *M. smegmatis* with trehalose stabiliser having significantly higher yield for both bacteria tested when compared to no novel enzymes for both bacteria. In addition to the aforementioned loss in efficiency it should also be noted that the lyophilised protein containing sucrose was difficult to resuspend and may have been the reason for some of the variations, however no additional experiments were done to improve this.

Although we were able to analyse the breakdown of arabinose and galactose in a qualitative form with the use of TLC, we were not able to quantitatively able to determine the change in efficiency after freeze dying, although this would be relatively easy to do using a HPAEC-PAD.

This protocol is not designed to replace those protocols which are specifically designed for HMW DNA extractions. However, it can work alongside providing a protocol which is both quicker than HMW extractions providing an alternative with a reduced cost and speed that provides quality of DNA that can be used for downstream analysis.

In conclusion we have shown a high throughput, low-cost protocol using an adapted GenElute™ Bacterial Genomic DNA Kit (MERK) protocol for mycobacteria that allows for easy modifications depending on the end user needs and potentially uses across the Mycobacteriales order.

The cocktail is under patent: ENZYMES AND USES THEREOF PCT/GB2023/050457. Under the names of inventors Josep Manion, Patrick Moynhan, Amar Gudka and Elisabeth Lowe.

4.5 Future work

In order to optimise lyophilisation, further work would be needed to determine the optimal method to maintain the function of enzymes after resuspension which would allow for marketability of the protocol as well as the allowing the use of the protocol for those who do not have skill in protein purification. To verify if the sucrose is having an inhibitory effect upon the DG0470 adding the sucrose to fresh enzyme and analysing the breakdown products would give more conclusive evidence.

Additionally determining how the complete cocktail lyses the cells, to identify the effects that each of the components: lysozyme, lipase and novel enzymes have upon cell integrity, viability and morphology *in vivo* can be done with the use of microscopy, would give a clearer understanding of the process in which the enzymatic lysis works using that to determine potential alteration to the protocol. Optimisations for yield or length of the extracted gDNA could also be performed, such as altering the speed of the lysis and proteinase K digestion to gauge how these affect gDNA. The use of more specialist HMW extraction protocols could also be included into an adapted protocol.

Due to the use of generic lysozyme which is non-specific to mycobacteria future work into both a specialised lipase and lysozyme, such as RipA or RipB as a lysozyme alternative and LysB as a lipase alternative for mycobacteria, which have been previously identified (Gil et al., 2010; Healy et al., 2020; Martinelli & Pavelka, 2016). Although these specific esterase's and peptidases may well increase yield and efficiency to a greater extent, they have not been produced to a commercially viable concentration, so optimisation would be needed for this.

Chapter 5. Comparison of Nanopore sequencing of DNA extraction techniques across species.

5.1 Introduction

5.1.1 Whole Genome sequencing

DNA is the principal code for all life, sequencing it and understanding it is a powerful tool in a multitude of fields. Allowing for more accurate disease tacking, drug discovery and a deepening understanding of genomics. Since the first sequencing of λ bacteriophage in 1982 (Sanger et al., 1982) and the first complete cellular genome sequence of *Haemophilus influenzae* in 1995 (Fleischmann et al., 1995), the field of DNA sequencing has evolved at an exponential rate, becoming increasingly accessible and affordable.

The first steps towards Whole Genome Sequencing (WGS) were pioneered by Sanger using radio labelling of partial digests of DNA. Using Dideoxynucleotides (ddNTPs) lacking the 3' hydroxyl group and therefore not allowing the addition of more dNTPs. Radio labelled ddNTPs randomly bind to the elongation making DNA fragments of every possible length, it was using this technique that the first protein coding gene, the coat protein of bacteriophage MS2, was sequenced (Jou et al., 1972). However, even as these techniques were developed and refined by many groups, the process remained costly and time consuming.

Second generation sequencing can broadly be characterised into two major categories; sequencing via hybridization and sequencing by synthesis (SBS) (Slatko et al., 2018), many current WGS techniques are still based on SBS including Illumina or Ion Torrent platforms.

Sequencing by hybridisation: originally developed in the 1980s using arrayed DNA oligonucleotides of known sequence. DNA is repeatedly hybridized to the array then non-hybridized DNA is washed off, this allows for the determination of sequence (Drmanac et al., 2002). This technique is mainly only used for diagnostics now.

SBS requires direct action of DNA polymerase to produce an observable output. Uses a solid support containing micro channels or wells in which the reactions take place. Most

SBS methods do not use dideoxy terminators, instead using luminescent molecules they measure pyrophosphate production which is used to convert pyrophosphate to ATP which is used by a luciferase producing luminescence in real time (Pervez et al., 2022; Rhoads & Au, 2015). Heather & Chain, (2016) reviews this more in depth.

Finally, at present we are in the third generation of sequencing, third generation sequencing involves Single molecule sequencing (SMS). SMS allows sequencing in real time without amplification which reduced bias. One of the key advantages in SMS is the ability to have long reads. It also allows for much easier *de novo* assembly of whole genomes when compared to short read counterparts, however at the cost of higher error rates (Hook & Timp, 2023).

5.1.2 Nanopore Sequencing

Founded in 2005 at the University of Oxford by Dr Gordon Sanghera, Dr Spike Willcocks and Professor Hagan Bayley. The first sequencing data was presented in 2012, with the first commercially available product, the MinION, being released in 2014. The technology allows for long single read sequencing with the longest single read at the time of writing being 4.2 Mb in an internal Nanopore study (Jain, 2023).

The technology relies on nanoscale protein pores that serves as biosensors which are embedded in an electrically resistant membrane (**Figure 5.1**). A constant voltage is applied to produce ionic current through the nanopore. A motor protein is attached to the nanopore. This motor protein has two roles: to act as a helicase and to control the speed of translocation of the ssDNA across the nanopore. The changes that occur in the ionic current are due to different nucleotide sequences passing through the nanopore (Wang et al., 2021).

The key element of the function came with the use of phi29 DNA polymerase and a Nanopore; either α-hemolysin or MspA. The addition of the phi29 allowed for the controlling of the speed of translocation, decreasing the fluctuation in translocation in kinetics, increasing data quality (Cherf et al., 2012). Initial readings from Nanopore however were very error heavy with error rate being estimates at up to 38% (Laver et al.,

2015). Later iterations of the chemistry and pores has reduced error to 6% and 8% for low and high GC reads respectively (Delahaye & Nicolas, 2021). During the course of the technologies lifetime ONT has constantly updated the chemistry of their systems with the chemistry at time of writing being Kit 14 and the nanopore being R10.4.1.

Along with its ability to perform ultra-long reads. Nanopore MinIons and Flongles have the benefit of being a portable with no inbuilt computer they are connected via USB allowing their use in so called 'Labs in suitcases', thus allowing sequencing *in-situ*. These have already been implemented in areas where the Ebola epidemic has spread, as well as for water testing in Ethiopia the latter costing approximately £10,000 for all the necessary equipment for the lab (Halla et al., 2022).

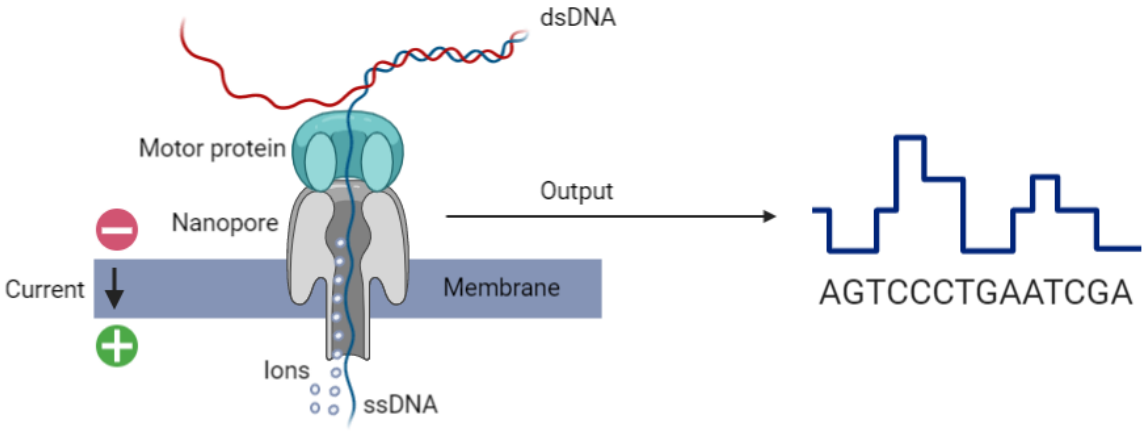


Figure 4.5.1 Representation of the mode of action of Nanopore sequencing. Cartoon respresentation of Nanopore sequencing. Including the motor protein (phi29), transmnmbrane Nanopore. With the outer memberane negative charge and positive internal charge shown.

5.1.2.1 Library preparation techniques

There are two main methods of library preparation for ONT; these are ligation and transposase-based shown in **Figure 5.2**. Rapid Transposase-based sequencing works by randomly cleaving DNA and attaching barcodes via transposome-complexes to the ends of DNA, allowing up to 96 samples to be sequenced in one library (Tyler et al., 2018).

Ligation library preparation, which involves end prep and nick repair, as well as the ability to use size selection, allows for greater control of your samples compared to rapid barcoding, although the time taken to prepare the library is much greater.

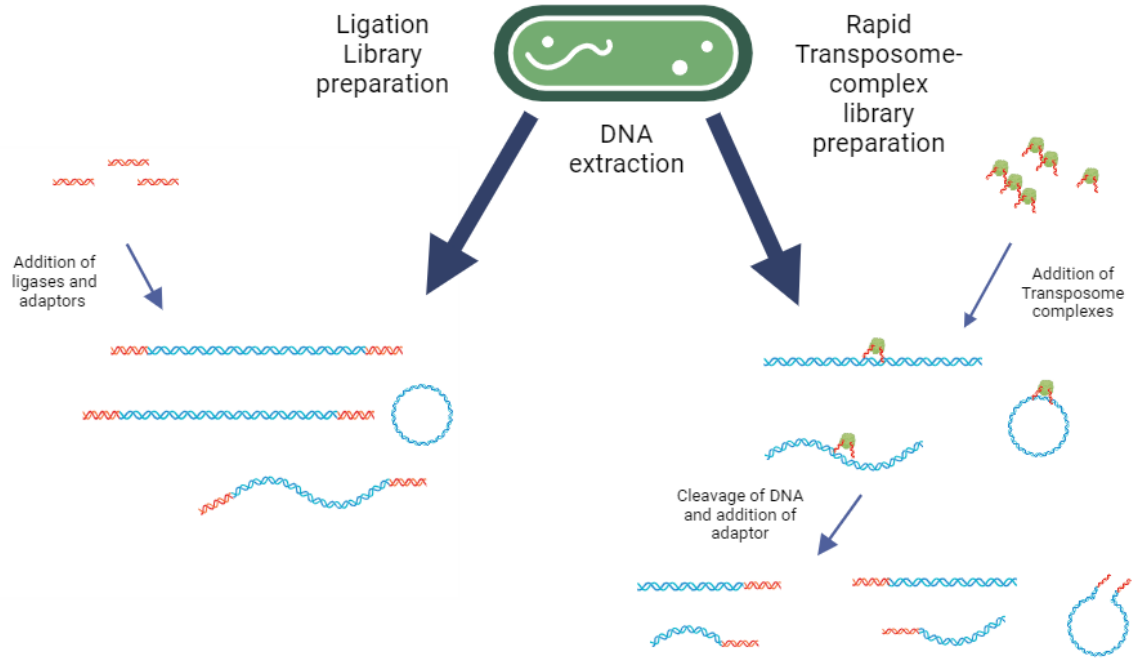


Figure 4.5.2 ONT library preparation techniques. Representation of the two main ONT library preparation techniques - ligation library preparation and Rapid Transposome-complex library preparation, with blue DNA indicating extracted DNA and red DNA indicating adapters or barcodes.

5.1.3 De Novo Assemblies

De novo assemblies are constructed by matching overlapping sequencing regions (Kmers) until ends do not overlap with other Kmers. This builds the contigs. No refence is used for the construction of the genome assembly, allowing for an unbiased construction of the genome. However, this does mean misassembles can occur especially if the quality of the reads is low with the assembly being of a high contig count (Sohn & Nam, 2018).

5.2 Objectives.

I. Evaluate the quality of the DNA extracted through the protocol developed in **Chapter 4.**

II. Develop a pipeline for de novo Assembly of the sequencing data.

5.3 Results

DNA was extracted under 3 conditions; Complete cocktail, No Novel Enzyme, Bead Beating, these conditions are outlined in **Chapter 4.**

5.3.1 DNA sequencing data

Basic analysis of the statistics of each barcode was performed using NanoStat (De Coster et al., 2018) which provides basic sequencing statistics including number of reads, longest read and N50.

All sequencing was performed as described in **Section 2.1.25.** Assembly of the genomes was performed through FLYE (Kolmogorov et al., 2020) with the estimated genomes sizes used for each organism shown in **Table 5.1**. Once contigs had been assembled in FLYE, genome analysis was performed in QUAST and PAThosystems Resource Integration Center (PATRIC) (Gurevich et al., 2013; Olson et al., 2023). Quast allows for the alignment of a constructed genome to a reference allowing the determination of percentage genome alignment. PATRIC is an online tool which allows for the identification of genome as well as the annotation combining a web-based graphic interface with command line.

Table5.1.1 Genome lengths used for de novo Assembly in FLYE.

Organism	Estimated Genome (Mb)	Plasmids
M. smegmatis MC2 155	6.9	0
M. Marinum ATCC 927	6.4	0
N. Farcinia NCTC11134	3.6	4
T. paurometabola DSM 20162	4.4	1

5.3.1.1 M. smegmatis MC2 155

The initial analysis of the Nanopore sequencing of *M. smegmatis* MC2 155 is shown in **Table 5.2**. The initial analysis of the longest read shows very little difference between No Novel enzyme and Complete cocktail with both having a longest read of ~80 kbp. N50s, which are the mean length of the DNA fragments are higher with Novel Cocktail however not to a large degree. However, the variation in genome length when *de novo* assemblies are constructed in FLYE, is shown clearly in **Table 5.2** with all constructs from complete cocktail being >6.8 Mb while No Novel enzymes only one genome construct is >6 Mb.

This disparity is also shown by the variation in % genome alignment to the reference genome *M. smegmatis* MC2 155 via QUAST. Alignment of the Complete cocktail barcode constructs all exceeded 90% alignment while No Novel Enzyme extractions ranged between 24-88% alignment to the reference.

Although not above 90%, its interesting the coverage obtained from No novel enzyme is higher than expected due to the nature and difficulty extracting DNA from mycobacteria, with the highest alignment of barcode constructs being 88.892% to the reference genome, although the quality of the construct is lower than ideal with 109 contigs.

Bead beating unsurprisingly showed a low N50 and longest read. However interestingly we were able to obtain a high genome alignment score at 94.961% from 47 contigs, although the other replicates showed a large discrepancy with the lowest being 41.446% from 136 contigs and this large range matches the variations across the yields by mechanical mechanisms that we see in yields presented in **Chapter 4**.

Table 5.2.2 Nanopore sequencing and genome assembly results for M. smegmatis Mc2155. Nanopore sequencing results for *M. smegmatis* Mc2155 N50 and longest read obtained through NanoStat. Genome length and number of Contigs constructed using FLYE. Aligned genome (%) to reference genome performed in QUAST. Each line represents a different barcode.

Conditions	N50	Longest read	Genome length	Contigs	Aligned genome (%)
Bead Beating	1521	32510	6682918	47	94.961
Bead Beating	2927	17677	2922509	136	41.446
Bead Beating	2918	17881	4888789	194	68.979
No Novel Enzyme	8016	86370	1769565	29	24.927
No Novel Enzyme	2429	26023	5763848	182	79.424
No Novel Enzyme	7991	63585	6253400	109	88.892
Complete Cocktail	9398	59120	6893278	54	98.025
Complete Cocktail	10960	85975	6876478	29	97.942
Complete Cocktail	11928	49012	6978407	19	99.416

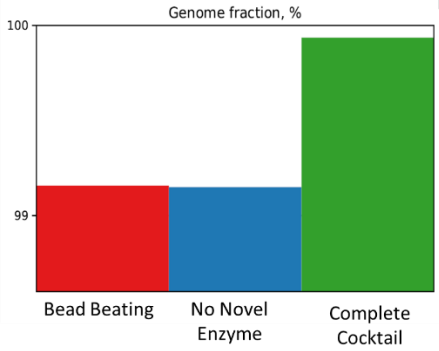
To see if we could obtain increased coverage and enable greater access to genome analysis, the triplicate barcodes for each condition were pooled, FLYE was run with the same constraints as previously, the results are shown in **Figure 5.3. Figure 5.3a** shows the genome fraction (%) that was aligned to the reference genome. When the triplicate barcodes were pooled for bead beating extractions an alignment of 99.157% from 10 contigs was obtained. The reduction in contig count is especially surprising based on the results of the N50s and longest reads, even with pooled barcodes a much more fragmented construct would be expected.

The pooling of the No novel enzyme triplicates produced a construct with an alignment to the reference genome of 99.149% from 11 contigs, which is a substantial increase compared to the individual highest coverage from 88.892%, as well as a reduction in contigs from 109. Complete cocktail alignment also increased from 99.416 to 99.935%,

although a small increase, the contigs count reduced from 19 to 4, therefore less scaffolds were needed aiding in confidence of the assemblies.

Visualisation of the alignments are shown in **Figure 5.3** using MAUVE ran through Patric. Bead beating assembly shows a large amount of misassembles with contigs having not been assembled in the correct order.

a.



b.

Condition	Aligned genome length	Contig
Bead Beating	6 939 323	10
Enzyme	6 961 401	11
Complete Cocktail	7 001 561	4

c.

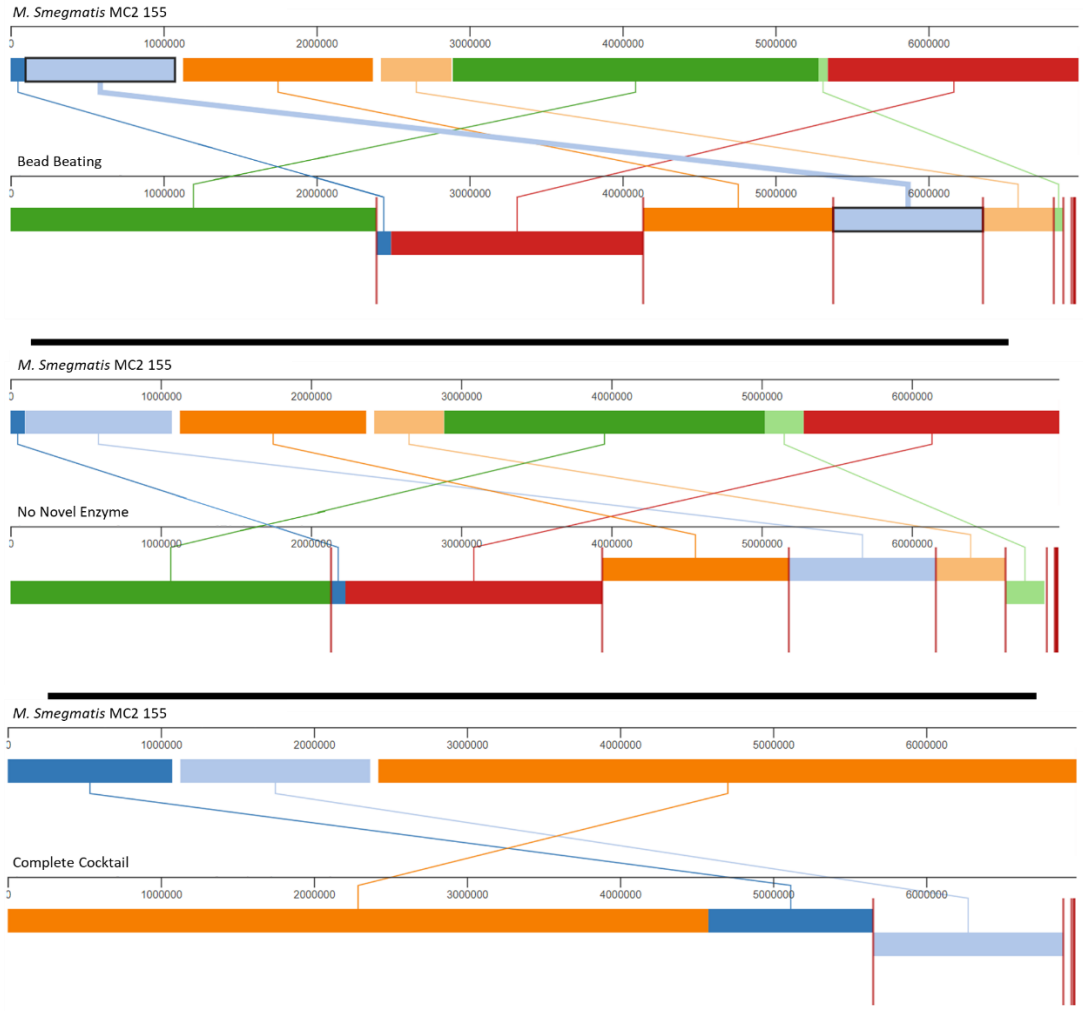


Figure 5.1 Analysis of pooled barcodes for each condition for *M. smegmatis* Mc²155. A. Percentage alignment of the *de novo* assembled genome to the *M. smegmatis* MC2 155 genome, with bead beating in red aligning to 99.157% of the reference genome, No Novel Enzyme in blue aligning to 99.149% of genome and Complete Cocktail in green aligning to 99.935% of reference genome. **B.** data for the aligned genome lengths and number of contigs for each de novo assembly. **C.** MAUVE alignments of the de novo assemblies to the reference genome.

5.3.2.1.1 Phylogenetic determination

Although the DNA extraction was done using a known reference genome, it is important in *de novo* assemblies that you are able to determine an unknown species, for that we implemented the use of Type genome strain server and Mash: fast genome and metagenome distance estimation using MinHash (Ondov et al., 2016).

To construct a phylogenetic tree TYGS (Meier-Kolthoff & Göker, 2019) was used the tree constructed is shown in **Figure 5.4.** TYGS was able to identify all *de novo* assemblies down to species level of *M. smegmatis*, however not strain. All pooled barcodes are however identified as the same stain belonging to the same node. The closest strain identified is *M. smegmatis* NCTC 8159, with only a high-scoring segment pairs (HSPs) score of 96.5. If ignoring the genome length due to the nature of the *de novo* assemblies being draft genomes. This score is too low to be confident that this is the correct strain.

Due to the low similarity results of the phylogenetic analysis, MinHash (Ondov et al., 2016) was used through PATRIC (Wattam et al., 2017). Using the similar genome finder tool in PATRIC the genome constructs were analysed. Through this method the closest genome identified was *M. smegmatis* str. MC2 155 with a distance of 0.0000715897 from the pooled barcodes. Using a combination of TYGS and Mash it can be deduced that the *de novo* assembly is in fact an assembly for *M. smegmatis* str. MC2 155.

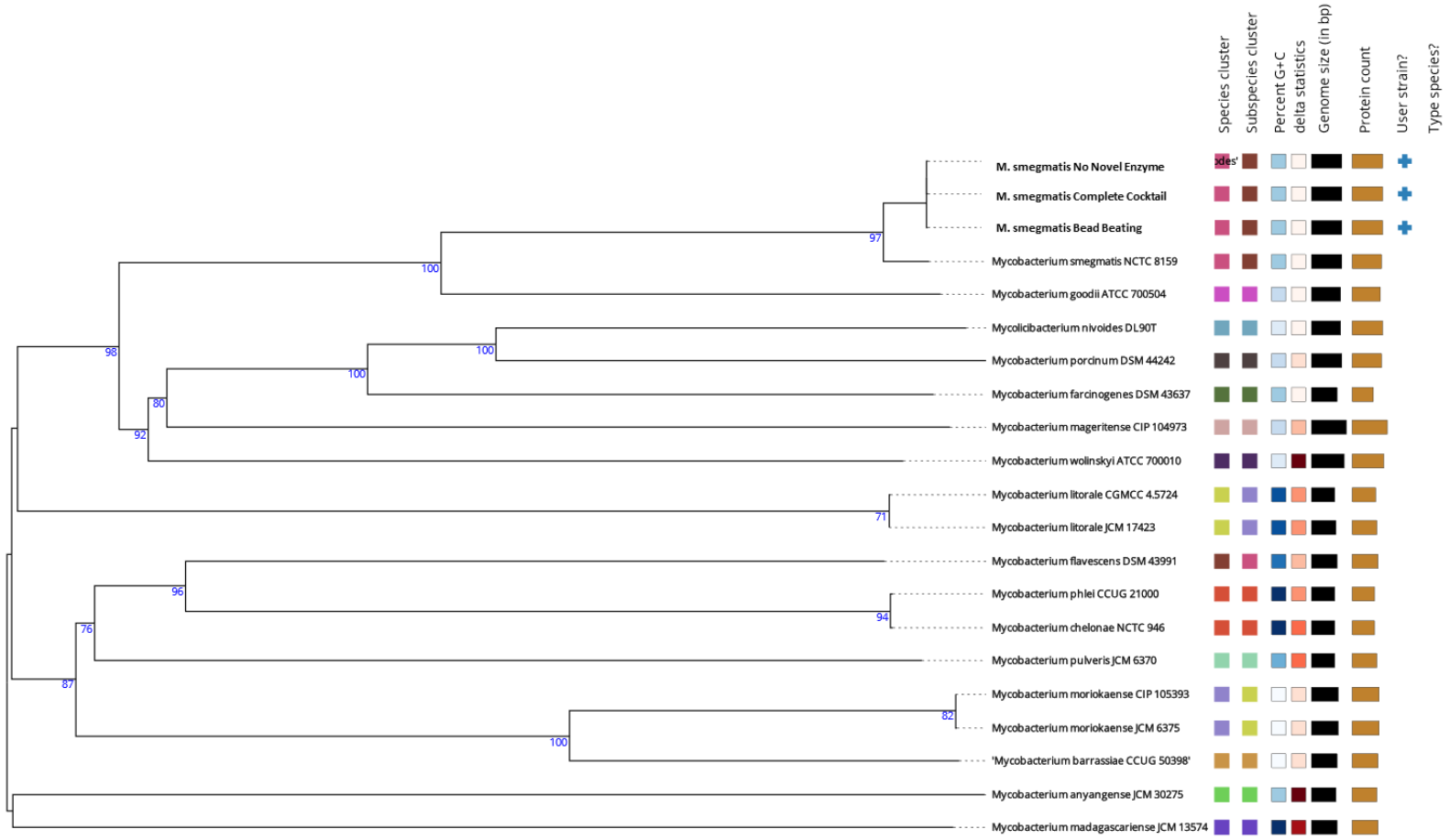


Figure 5.2 Phylogenetic tree of *de novo* Constructs of M. smegmatis. Phylogenetic tree of the *de novo* assemblies of the pooled barcodes for *M. smegmatis* MC2 155 produced in TYGS.

5.3.1.2 N. farcinica NCTC11134

The initial Nanopore reading statistics from NanoStat analysis of the barcodes for each barcode are shown in **Table 5.3.** There is little variation in the N50 or longest read between No novel enzyme and Complete cocktail. Genome lengths of the *de novo* assemblies do vary greatly between conditions, with the highest genome lengths constructed from Complete Cocktail extraction with two replicates being over 6.6 Mb and one being 5.4 Mb. Meanwhile No Novel enzyme extraction constructs had a much more varied length ranging from 271 Kb to 6 Mb with the lowest (271 Kb) only aligning to 1.209% of the reference genome.

Bead Beating provided the characteristically low N50, however genome lengths were much lower than obtained in other samples during this study, with the highest genome alignment from the triplicates only being 15.206%.

The constructed length of all three of the Complete Cocktail triplicates did cause concern with the closely related *N. farcinica* NCTC630 having a genome length of 6.4 Mb. To ensure the correct genome had been sequenced Complete Cocktail barcodes were aligned with QUAST against the *N. farcinica* NCTC630 to analyse construct quality as well as alignment score. With a constructed genome length of 6.5 Mb, constructed genome only had an alignment score of 90% meaning 10% of the constructed genome was not present in the reference genome of *N. farcincia* NCTC630. To verify that the sequence, *N. farcinica* NCTC11134 QUAST was performed as before, including the 4 plasmids present in *N. farcinica* NCTC11134. When they were analysed against these plasmids plus the genomic chromosome, a high percentage of this was covered by the assembly. This confirms that the sequence present was that of *N. facinica* NCTC11134.

Table 5.3.3 Nanopore sequencing and genome assembly results for *N. farcinica NCTC11134.* Nanopore sequencing results for *N. farcinica NCTC11134 N50* and longest read obtained through NanoStat. Genome length and number of Contigs constructed using FLYE. Aligned genome (%) to reference genome performed in QUAST. With each representing a different barcode.

Conditions	N50	Longest read	Genome length	Contigs	Aligned genome (%)
Bead Beating	2825	41339	1673344	74	15.206
Bead Beating	2457	30286	205065	9	1.109
Bead Beating	2635	26488	62892	3	TL
No Novel Enzyme	6622	60784	6042616	58	90.827
No Novel Enzyme	4959	62510	271444	34	1.209
No Novel Enzyme	6288	41370	2986328	61	38.966
Complete Cocktail	4974	49676	6354474	32	97.982
Complete Cocktail	7038	52214	5356969	83	76.969
Complete Cocktail	5825	59392	6352535	42	98.294

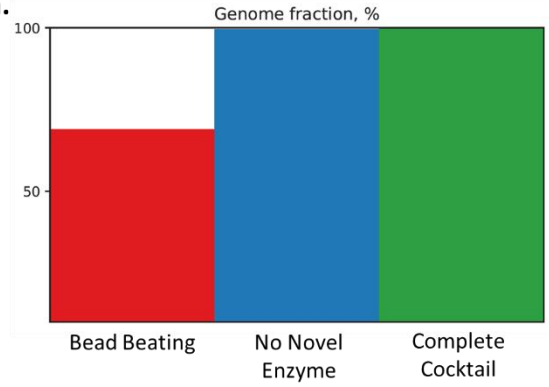
As done previously to understand how much coverage could be obtained from the pooling of all triplicate and re-analysis this was performed.

With the new assemblies QUAST was used to analyse them again the chromosomal genome, the results are shown in **Figure 5.5. Figure 5.5a** shows the % of the genome aligned to the extractions. The complete cocktail extractions cover 100% of the genome from 5 contigs, while No Novel Enzymes are barcodes are only able to be aligned to 99.56% of the reference genome. Although the pooled bead beating genome is only 2.6 Mb aligning to 68.9% of the reference genome, surprisingly low even for bead beating as well as being from 157 contigs, so a low quality.

Overall, the pooled Complete cocktail barcodes for *N. farcinica* provide a higher alignment to the reference genome from fewer contigs when compared to both No Novel Enzyme

and Bead beating. No Novel Enzyme provided the second highest alignment of the conditions tested. With beat beating providing low alignment and a high contig number.

a.



b.

Condition	Aligned genome length	Contigs
Bead Beating	2 574 383	157
No Novel Enzyme	3 638 103	19
Complete Cocktail	3 655 552	5

c.

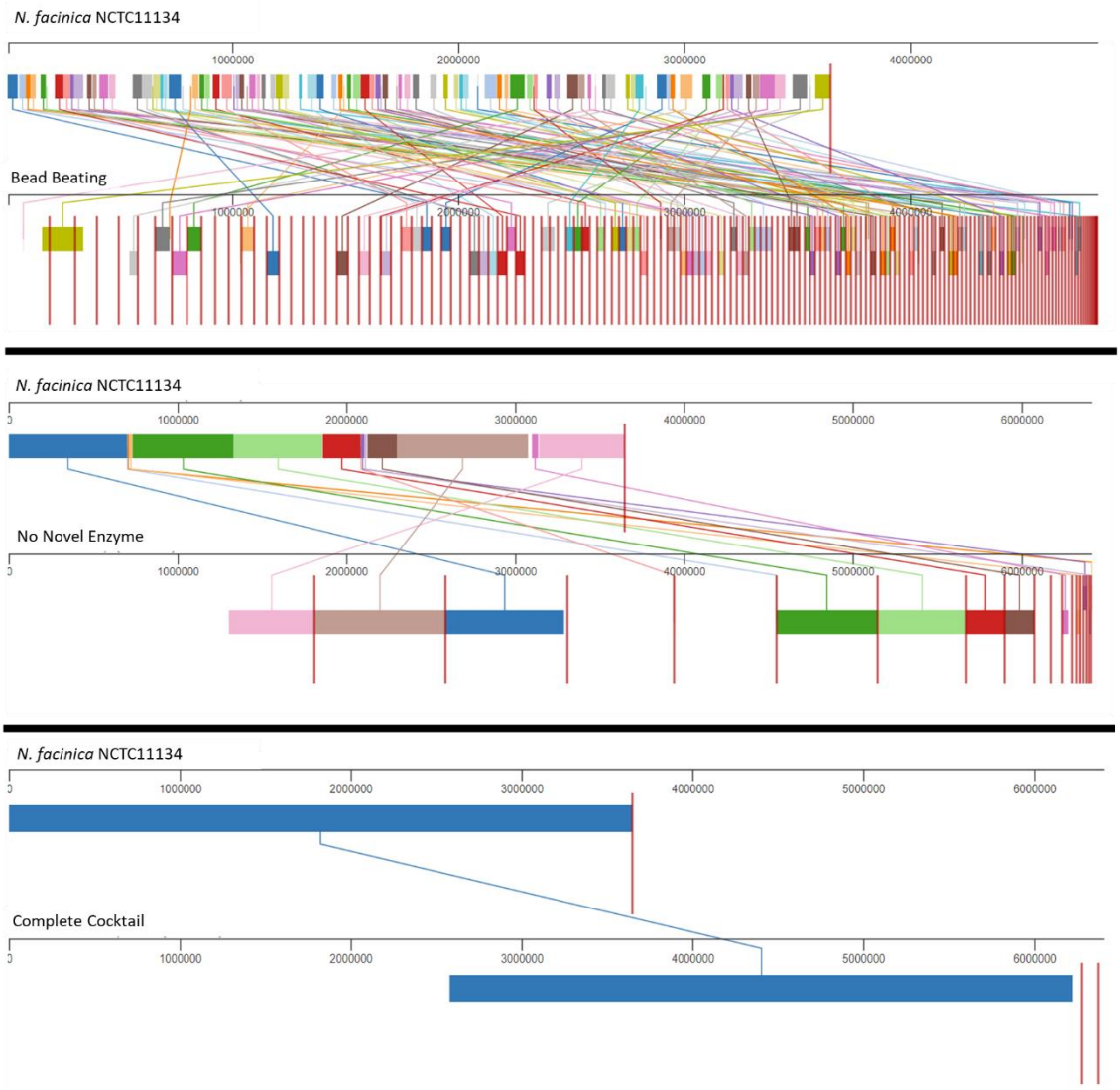


Figure 5.3 analysis of pooled barcodes for each condition for *N. farcinica*. A. Percentage alignment of the *de novo* assembled genome to the *N. farcinica* NCTC11134 genome, with bead beating in red aligning to 68.9% of the reference genome, No Novel Enzyme in blue aligning to 99.56% of genome and Complete Cocktail in green aligning to 100% of reference genome. **B.** data for the aligned genome lengths and number of contigs for each *de novo* assembly. **C.** MAUVE alignments of the *de novo* assemblies to the reference genome, size variations are due to the aforementioned large plasmids.

5.3.1.2.1 Phylogenetic analysis

The final confirmation of the strain identification for the sequence DNA was performed in TYGS as shown in **Figure 5.6**. Both enzymatic extraction and Complete cocktail extraction have been shown to be able to be identified down to species level being identified as *N. farcinica* NCTC11134, while the pooled barcodes extracted using bead beating are only identifiable down to the species, however not the strain. While it was not able to identify the strain from bead-beaten samples, it was still identified down to species level which is still surprising as the genome is only 68.9% aligned to the reference.

This shows that using the *de novo* assemblies you can identify down to the strain which interestingly with both Complete Cocktail extraction and No Novel enzyme extraction.

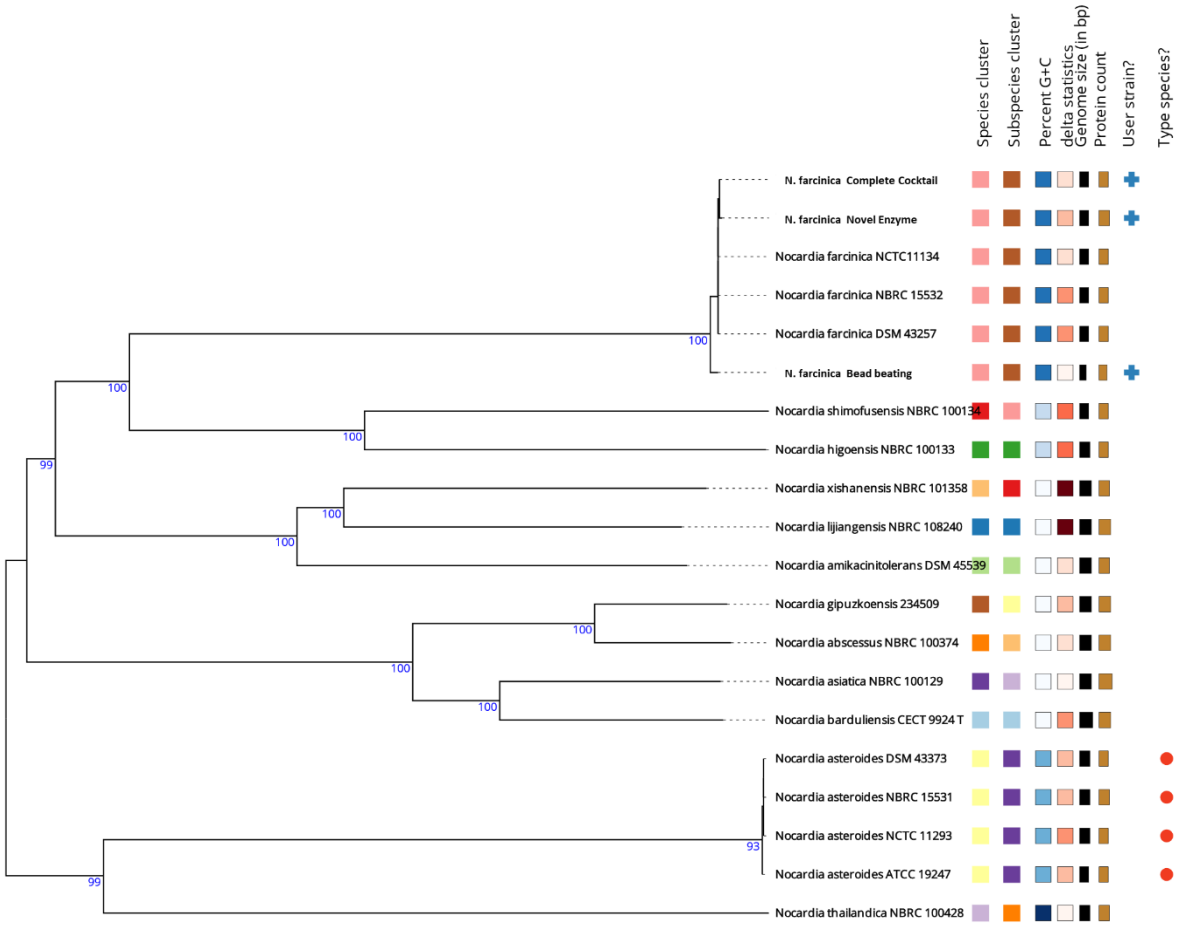


Figure 5.4 Phylogenetic tree of *N. farcinica de novo* Assemblies. Phylogenetic tree analysed and constructed in TYGS of the *de novo* assemblies of Pooled barcodes of the three DNA extraction conditions tested.

5.3.1.3 M. marinum ATCC927

The NanoStat analyse for the sequencing of *M. marinum* are shown below in **Table 5.4.** All conditions were done in triplicate, however two of the Bead Beating replicates and one of the No Novel enzyme barcodes were of too poor-quality reads for genomes to be assembled in FLYE and therefore the data is not shown.

In the case of *M. marinum* although N50s are lower for No Novel enzyme when compared to Complete Cocktail, there was no difference between longest read in the replicates of enzymatic extraction, when analysed in FLYE the assembly lengths for both No Novel enzyme and Complete Cocktail did not vary between the two conditions with them both ranging between 4.8 Mb and 6.1 Mb. All the *de novo* assemblies from the enzymatic lysis methods were also constructed in a similar number of contigs, ranging between 85 and 105, indicating damage to the DNA and poor sequencing quality. When they were aligned to the reference *M. marinum* ATCC 927, the % alignments showed 87.55 to 93.124% coverage for the No Novel enzyme extraction. A surprisingly higher than expected genome alignments.

The complete cocktail extraction aligned to between 72.628 to 89.49%, which interestingly is lower than No Novel enzyme although all three replicates were able to be assembled into genomes, as one of the No Novel enzymes barcodes was not able to construct into an assembly in FLYE.

The results for bead beating are poor with two of the barcodes being unable to be assembled into genomes in FLYE. Additionally, the barcode which was able to be assembled into an assembly had only a genome length of 3 Mb aligning to only 46.122% of the reference genome.

Table 5.3.4 Nanopore sequencing and genome assembly results for *M. marinum* ATCC927. Nanopore sequencing results for *M. marinum* ATCC927 N50 and longest read obtained through NanoStat. Genome length and number of Contigs constructed using FLYE. Aligned genome (%) to reference genome performed in QUAST. With each representing a different barcode.

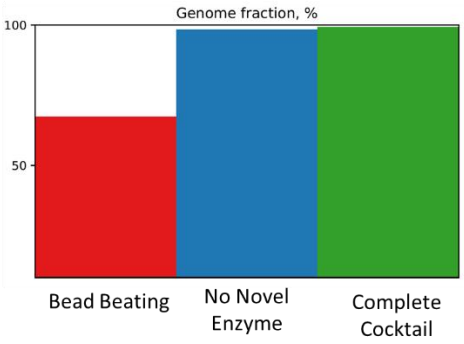
Conditions	N50	Longest read	Genome length	Contigs	Aligned genome (%)
Bead Beating	6745	32510	3048388	80	46.122
No Novel Enzyme	5876	82033	6168311	85	93.124
No Novel Enzyme	2881	40924	5822935	105	87.555
Complete Cocktail	6384	62132	4956583	105	73.352
Complete Cocktail	6613	47726	4838444	100	72.628
Complete Cocktail	6195	65632	5959729	98	89.429

After pooling of the barcodes including those which were not able to the be initially constructed the results are shown in **Figure 5.7.**

The alignment of all the conditions increased when compared to individual barcodes. The alignment of bead beating from 46.122% alignment of the single highest barcode to 67.322% when all three triplicates were pooled and reconstructed from 122 contigs, with a genome length of only 4.4 mb which is still a poor quality of coverage. The highest single barcode for the Complete cocktail was 89.429% which rose to 99.249% alignment with a genome size of 6.4 Mb. Although when pooled No Novel enzyme extractions did increase in coverage, the increase was only from 93.124% alignment to the reference of the highest single barcode to 98.31% alignment when pooled.

Overall, the pooling of barcodes had the greatest increase for the Complete Cocktails, which when pooled had a higher coverage of the genome than No Novel Enzymes in contrast to single barcode constructs. While bead beating maintained a low coverage from a large number of contigs.

a.



b.

Condition	Aligned genome length	Contig
Bead Beating	4 403 626	122
Enzyme	6 353 752	12
Complete Cocktail	6 426 753	18

c.

Figure 5.5 analysis of pooled barcodes for each condition *for M. marinum.* A. Percentage alignment of the *de novo* assembled genome to *the M. marinum* genome, with bead beating in red aligning to 67.322% of the reference genome, No Novel Enzyme in blue aligning to 98.31% of genome and Complete Cocktail in green aligning to 99.249% of reference genome. **B.** data for the aligned genome lengths and number of contigs for each

de novo assembly. **C.** MAUVE alignments of the *de novo* assemblies to the reference genome.

5.3.1.3.1 Phylogenetic analysis

When the pooled barcodes were analysed in TYGS the phylogenetic tree which was produced (**Figure 5.8)**. The closest identified strain to the *de novo* constructs was *M. marinum* CCUG 20998, which also has the passport name *M. marinum* ATCC 927, which corresponds to the known genome.

Interestingly even at 67.332% Bead Beating were able to be identified down to strain level, which is surprising based on that low alignment to the corresponding genome.

Overall, all *de novo* assemblies were able to be identified down to strain level at a high confidence for Enzymatic extractions while Bead Beating, although it was still able to be identified this level the confidence, was reduced.

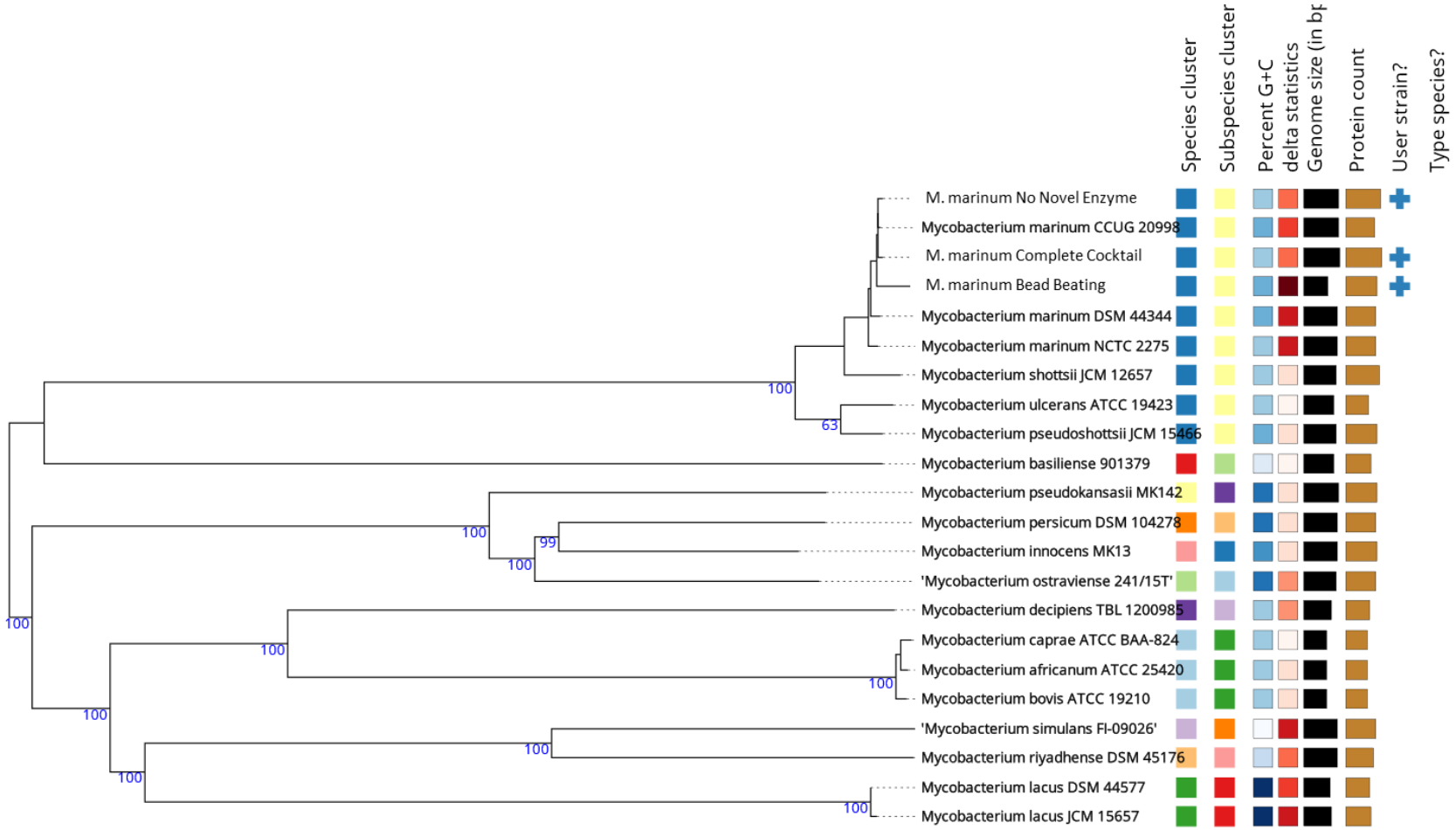


Figure 5.6 Phylogenetic tree of *M. marinum de novo* Assemblies. Phylogenetic tree analysed and constructed in TYGS of the *de novo* assemblies of Pooled barcodes of the three DNA extraction conditions test.

5.3.1.4 T. paurometabola DSM 20162

The results for the sequencing of *T. paurometabola* DSM 20162 are shown in **Table 5.5**, as with other species the initial analysis of the barcodes shows that Bead Beating extraction technique provides low quality reads and constructs, with only one of the barcodes for bead beating being constructed into a genome in FLYE and at 118 contigs and only 64.048% coverage. While *de novo* assemblies were constructed from 3 replicates of No Novel Enzyme extractions, the alignment to the reference *T. paurometabola* DSM 20162 were only 29.016 and 1.814% with the third being too low to give a result, which is expected due to the genome length of the assembly being 7323 bp from 75 contigs. Novel enzyme extractions on the other hand provided high genome length and increased alignment to the reference with one replicated being aligned to 99.88% of the reference from 3 contigs.

Table 5.3.5 Nanopore sequencing and genome assembly results for *T. paurometabola* DSM 20162. Nanopore sequencing results for *T. paurometabola* N50 and longest read obtained through NanoStat. Genome length and number of Contigs constructed using FLYE. Aligned genome (%) to reference genome performed in QUAST. With each representing a different barcode.

Conditions	N50	Longest read	Genome length	Contigs	Aligned genome (%)
Bead Beating	2534	18346	2906276	118	64.048
No Novel Enzyme	4701	45469	7323	75	TL
No Novel Enzyme	3376	69229	1280110	34	29.016
No Novel Enzyme	1321	69519	80080	2	1.814
Complete Cocktail	5589	39739	3442236	75	72.539
Complete Cocktail	5841	43430	2906276	118	64.048
Complete Cocktail	3724	70209	4475772	3	99.88

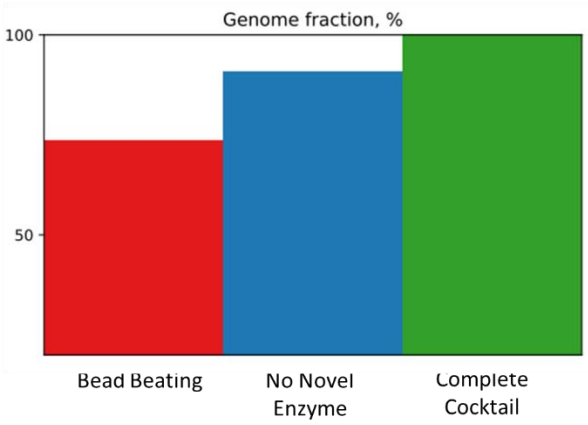
Condition barcodes were pooled into the conditions in which they were extracted and reanalysed using the same perimeters as previously in FLYE.

The results of the reanalysis are shown in **Figure 5.9**. 100% coverage was obtained from pooled Novel Enzyme extractions when the *de novo* construct was aligned to the reference genome of *T. paurometabola* DSM 20162 from 2 contigs.

With the pooling of the No novel enzyme, the coverage went from a peak of 29.016 to 90.897%, a surprisingly large increase when you compare it to the low coverage of the single barcodes. Although this could be due to the low-quality short reads of the barcodes that when pooled there is enough to form contigs, as the 90.897% alignment is constructed from 61 contigs. Conversely to this, a high increase in alignment when bead beating barcodes are pooled resulted in the genome only increased from a max of 2.9 Mb to 3.3 Mb, with a genome alignment percentage increase of only 64.048 to 73.721%.

Overall, The Complete Cocktail extraction provided higher quality *de novo* assemblies than those of both bead beating and No novel enzymes with a much lower contig count.

a.



b.

Conditions	Aligned Genome Length	Contigs
Bead Beating	3 333 711	114
No Novel Enzyme	4 007 092	61
Complete Cocktail	4 393 546	2

c.

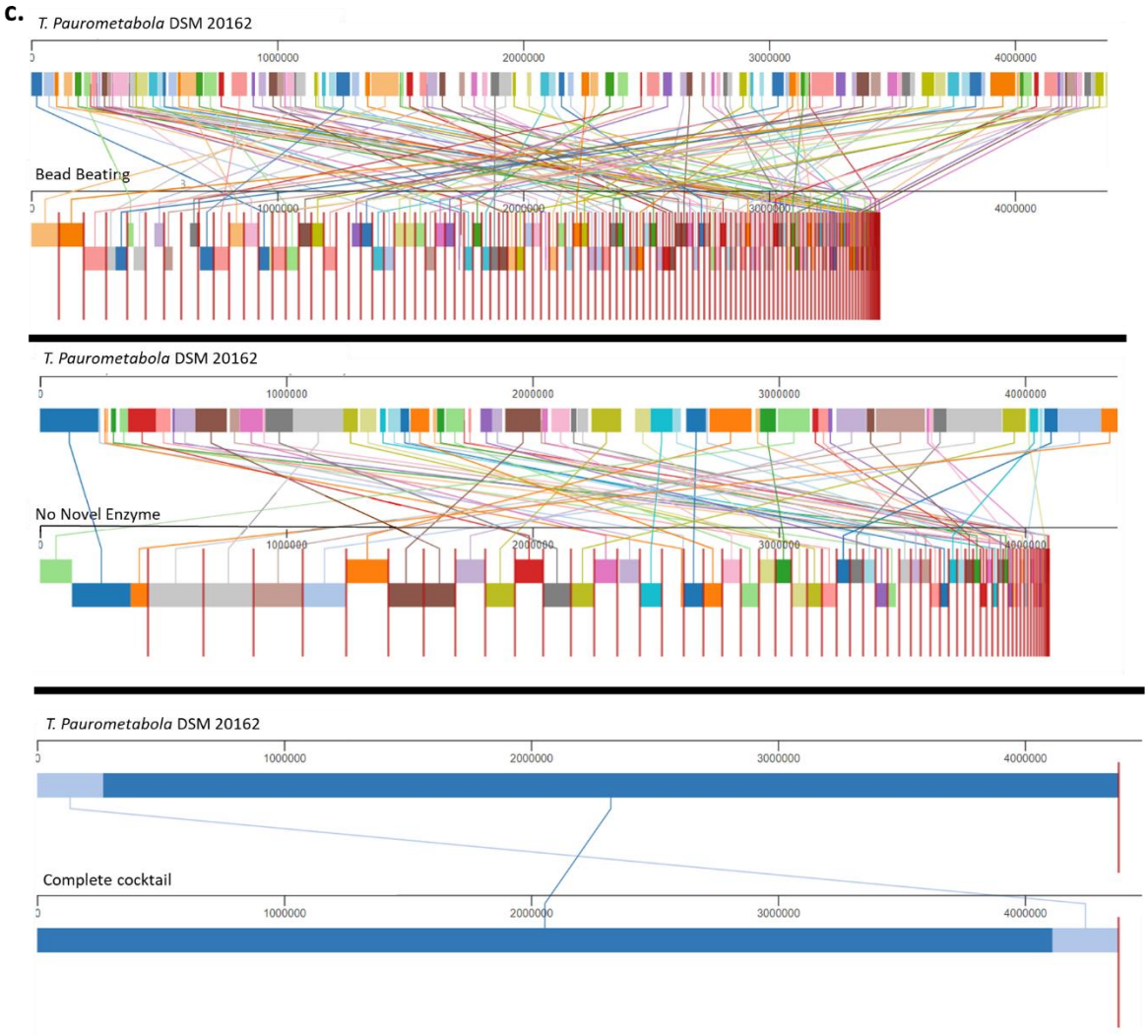


Figure 5.7 Analysis of pooled barcodes for each condition for *T. paurometabola*. A. Percentage alignment of the *de novo* assembled genome to the *T. paurometabola* genome, with bead beating in red aligning to 73.721% of the reference genome, No Novel Enzyme in blue aligning to 90.897% of genome and Complete Cocktail in green aligning to 100% of reference genome. **B**. data for the aligned genome lengths and number of contigs for each *de novo* assembly. **C.** MOUVE alignments of the *de novo* assemblies to the reference genome.

5.3.1.4.1 Phylogenetic determinations

As done previously, the pooled barcodes were input into TYGS for phylogenetic analysis. The phylogram produced is shown in **Figure 5.10.** The Complete Cocktail extraction *de novo* assembly was identified down to strain level *T. paurometabola* DSM 2016. While Bead Beating was able to be identified to species level, being identified as *T. paurometabola.* Meanwhile the bead beating assembly was only able to be identified down to the species having the same node as *T. paurometabola.* The increased specificity of the Complete Cocktail assemblies is useful for many other applications.

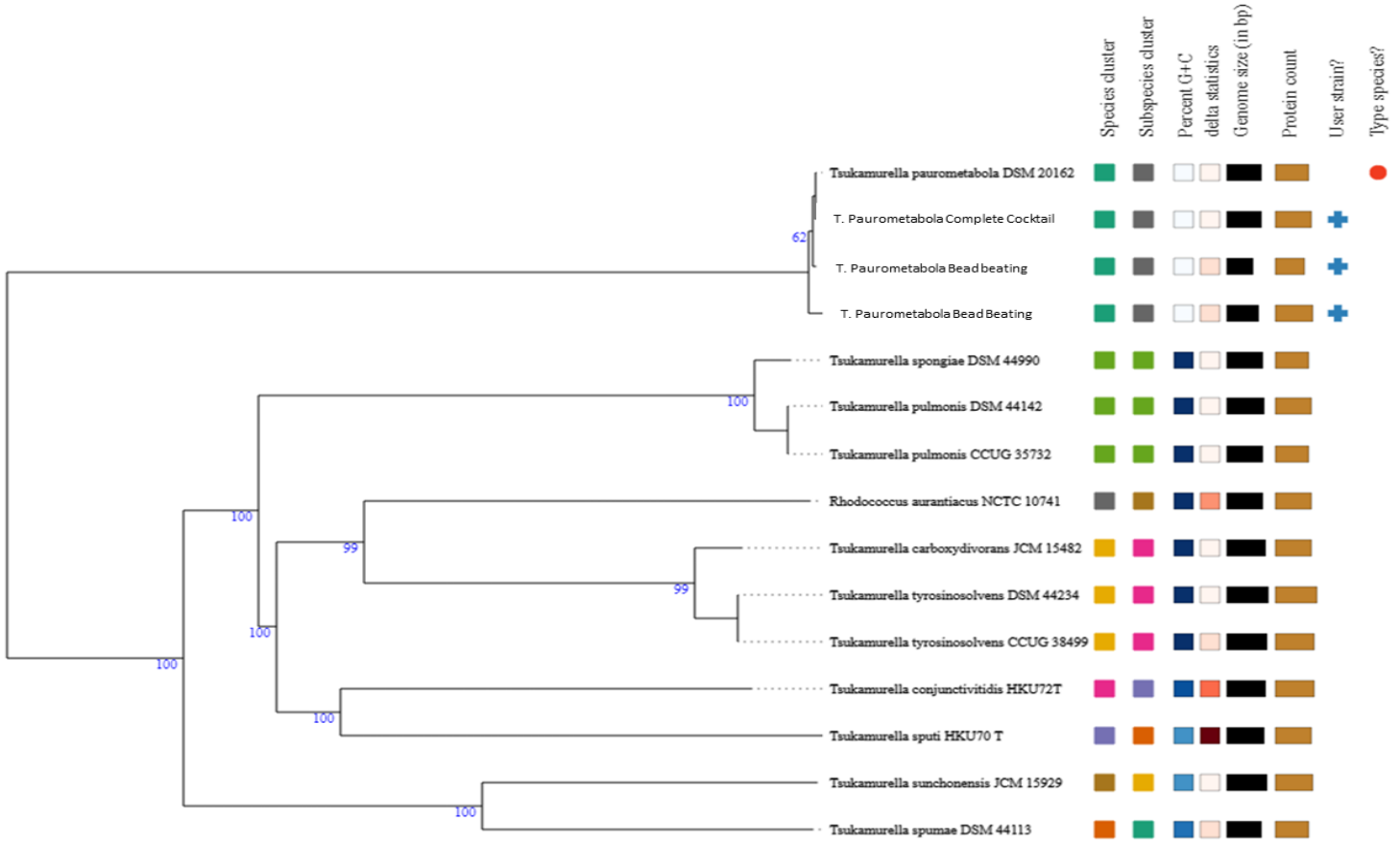


Figure 5.8 Phylogenetic tree of *T. paurometabola de novo* Assemblies. Phylogenetic tree analysed and constructed in TYGS of the *de novo* assemblies of Pooled barcodes of the three DNA extraction conditions tested.

5.4 Discussion

In this Chapter we have presented an extraction protocol that generated DNA that can be rapidly sequence, using ONT technology and construct a *de novo* sequence identify down to a species level consistently using DNA extracted using protocols outlined in **Chapter 4**.

During this study we have shown that the DNA obtained from the enzymatic lysis reagent described in **Chapter 4** provides greater sequencing quality across 4 different species, with consistently higher coverage of genomes than bead beating and higher than No Novel Enzymes with the exceptions of *M. marinum,* when analysing single barcodes.

Although not a method that provides HMW DNA and ultra long read sequencing, the combination of the Novel DNA extraction protocol use of Nanopore sequencing and *de novo* assembly protocol, is a pipeline that is accessible, cheap and does not require extensive knowledge of molecular biology or microbiology. With enough scope to adapt the protocol to the needs of the individual, such as increasing the power of the sequencing.

The results have shown that using the Complete Cocktail extraction in conjunction with GenElute™ Bacterial Genomic DNA Kits and ONT sequencing can provide DNA reads that can be used to produce *de novo* assemblies which have consistently >80% alignment to references genomes of the species tested in this study. Although No Novel enzyme extraction DNA can provide high coverage from a single barcode, as in the case of *M. marinum*, the consistency of the genome assemblies across the 4 studied species is less than the Complete Cocktail.

This study has corroborated a wide range of studies that show the sequencing of DNA extracted via Bead beating is of poor and inconsistent quality. While Bead beating can produce some genome constructs of okay quality and length, such as obtained from *M. smegematis.* The consistence of results is poor as in the cases of *N. farcinica* and *T. paurometabola* in which not all barcodes were able to be constructed into draft genomes in FLYE.

Overall, the sequenced quality of the DNA in this study aligns to studies based on similar species using ONT long read sequencing, with Bouso & Planet, (2019) obtaining full genomes from mean read length of <2 kbp. The power of the pipeline used in this study

is the ease of access, outside of FLYE and NanoStat, with NanoStat being optional and not an essential aspect of *de novo* Assemblies, as it only details basic statists of read quality. All other tools used (QUAST, PATRIC and TYGS) are free web-based tools, allowing for their use on machines that do not support Linux system which is typically used for WGS and genome analysis. This can enable a more accessible route to WGS and species identification, as well as use in the field, although not a replacement for more sophisticated methods such as those outlined by Chen et al. (2021).

For phylogenetic analysis a combination of TYGS and MinHash is appropriate as the limitations of such services are the databases which they pull from, as displayed in the analysis of *M. smegmatis* in which TYGS could only identify two species without the strain. Many difficulties could have been overcome with higher power software to fully determine plasmids and taxonomy in an automated manner, such as the use of PlasmidSeeker (Roosaare et al., 2018) for the identification of plasmids and Kraken2 (Wood et al., 2019) for taxonomic identification.

5.5 Future work

Although we have shown the effectiveness of the mycobacterial lysis reagents DNA extraction across 4 species, the future work could begin to optimise the use of the lysis reagent for DNA sequencing. This can be done in several ways such as using ligation during the library prep could aid in increasing the quality of the reads (Player et al., 2022), which although can add to the quality of the read adds to the preparation time, especially if the extraction is performed in conditions where sheering on the DNA is more common than in a molecular biology lab.

Although the webtools used allow for an efficient and easy access pipeline, limitations can occur. A potential way to bypass this while still staying on windows-based computing is developing a Google Batch workflow, which allows programmes to be run through Google Cloud computing. Eliminating the maintenance of hardware as well as access to data from any location and using more powerful tools while not requiring an increase in hardware capability.

Chapter 6. Discussion

The multi-layered impermeable nature of the mycobacterial cell wall, has made the genus important in human health, owing to difficulties in diagnosis and treatment (Saxena et al., 2021). In addition to a difficulty in treatment, the cell wall makes DNA extraction from the cells challenging (Ryu et al., 2016). The cell wall of mycobacterium contains an arabinogalactan domain which is present among several members of the Mycobacteriales order. The biosynthesis of this domain has been studied and review in great detail (Birch et al., 2008; McNeil et al., 1987; Seidel et al., 2007), however little is known about the breakdown and recycling of this domain, recently GlfH1 has been identified which can potentially recycle the D-Gal*f* domain (Shen et al., 2020). However, the breakdown of arabinose has remained unknown.

Recent work by Al-Jourani et al. (2023) identified PUL 42 of *D. gadei*, which contained D-arabinofuranosidases. During this study we have expanded the understanding of this PUL 42, characterising DG02470 (ascension HMPREF9455_02470) via biochemical and structural analysis (**Chapter 3)**. Identified as an exo-β-D-arabinofuranosidase with structural analysis via X-ray crystallography determining the structure to be monomeric with a N-terminal β-sandwich roll and a C-terminal (α/α)6-barrel fold. Previously only one β-D- arabinofuranosidase has been identified by that of ExoMA2$_{GH116}$ characterised by Shimokawa et al. (2023), however through structural and sequence analysis of the two showed very low homology, indicating that although the activity is shared, they do not pertain to the same GH family, with DG02470 being the founding member of a new GH family using SSN.

The characterisation of DG02470 provides further insight into the degradation of the arabinose domain of AG, as well as LM, LAM and ManLAM. Added to this are the recent identification of D-arabinofuranosidases; such as the enzymes located in PUL42 of *D. gadei* and DUF4185 containing proteins (Mab$_{GH4185}$, Phage$_{GH4185}$, Myxo$_{GH4185}$, MSMEG_2107, and Rv3707c) by Al-Jourani et al. (2023) and ExoMA2$_{GH116}$ from *Microbacterium arabinogalactanolyticum* characterised by Shimokawa et al. 2023. A broad range of enzymes have recently been identified which target the varying bonds which are present within mycobacterial arabinose oligosaccharides. Furthermore, galcatofuranosidase such

as those characterised by Al-Jourani et al. (2023) from PUL 37 of *D. gadei* and GlfH1 from *M. tuberculosis* (Shen et al., 2020) have been identified. These characterisations provide more solid foundation for the advancement of research into the mechanisms by which arabinogalactan is broken down within the order. These discoveries allow for a basis for homologous searches as well as new protocols for their identification. Potentially leading to a greater understanding of the breakdown and recycling of mycobacterial cell walls which can enable insights into the order, as well as the potential to open new therapeutic targets, allowing for highly targeted antimicrobials, especially as NTMs become increasingly clinically important and antibiotic resistance.

During this investigation we have also showed novel protocols which allow for the characterisation of β-arabinofuranosidase with greater ease. Such as using Δ*aftb C. glutamicum* mutants (Raad et al., 2010) to purify β-capless AG to identify specificity. The basis provide in this study will allow for future work in this field to have more entrenched protocols.

In **Chapter 4** we showed, using DG02470 in conjunction with the D-arabinofuranosidases identified from *D. gadei* PUL42 and D-galactofuranosidase enzymes characterised from *B. finegoldii* (PUL37) (Al-Jourani et al., 2023), that adding 5 µM of each arabinogalactan degrading enzymes to a modified GenElute™ Bacterial Genomic DNA Kit protocol successfully increased the yield of DNA from *M. smegmatis* and *M. abscessus* when compared to the protocol without the novel enzymes and bead beating. In addition, we also obtained preliminary yields from *N. farcinica, M. bovis* BCG and *M. avium* subp. *paratuberculosis,* which all increased significantly. Further modifications could be made to improve yield such as the inclusion of lipase and lysozyme specific to mycobacterium, such as lysB which specifically targets mycobacterial mycolic acids and RipA or RipB which target mycobacterial lysozyme (Gil et al., 2010; Healy et al., 2020; Martinelli & Pavelka, 2016), however at the time of this investigation we were unable to purify these proteins to a consistent and usable quantity. The protocol presented allows an initial start to further investigate accessible protocols for mycobacterial gDNA extraction. Although not a replacement for more complex HMW DNA extraction techniques, this protocol provides a simple to follow method for the extraction of mycobacterial gDNA that can be used by

those with limited molecular biology experience with yield high enough for qPCR and PCR rapidly.

Due to the high variance and specificity among glycans, their presence, determined via biomarkers, is used as a method to detect bacteria, fungus and cancer (Campanero-Rhodes et al., 2020; Theel & Doern, 2013). Due to the highly unique polysaccharides present within mycobacteria, this technique has already been proposed to determine the presence of Mtb (Chan et al., 2015; Tong et al., 2005; Tra & Dube, 2014).

The development of the lysis regents presented in **Chapter 4** while having shown to increase gDNA yields, the lysis reagent may potentially be adapted to aid in glycan detection of D-Ara*f* and D-Gal*f* from arabinogalactan or D-Ara*f* LAM, as shown in **Figure 4.4**.. Due to the unique nature of the conformation of D-Ara*f* and D-Gal*f* within the order, the lysis reagent followed by more conventional means of antibody biomarkers such as those outlined by Chan et al., (2015) in which My2F12 which binds the α1,2- mannose linkages, present on pathogenic Mtb, could be used. The addition of the mycobacterial lysis reagent presented, to a technique such as this, could allow for increased sensitivity as lower CFU count would be detectable. Removal of certain components, also reducing gDNA yield, from the lysis reagents, may allow for its application across other fields such as surface stability assays, or transporter studies.

Although we have shown that the lysis reagent improves DNA extraction, there are several other fields in which this could be implemented, adapting it for the individual study.

Using the DNA extracted from the 3 methods detailed in **Chapter 4** we also showed that the extracted using the complete cocktail DNA can be consistently used to sequence and constructed using *de novo* assembly down to a species level using GridION^{TM}, within a workday of experiment time, not including sequencing time. While the length of the DNA does not change between complete cocktail and no novel enzyme, which is expected due to both being enzymatic lysis of the cells, the higher yield from complete cocktail makes sequencing more consistent.

The pipeline presented *de novo* assembly using FLYE2 (Kolmogorov et al., 2020) to assemble contigs, genome analyse performed in PATRIC with the possibility to align to a reference genome in QUAST (Gurevich et al., 2013) or if the genome is unknown

identification down to species level can be performed using TYGS (Meier-Kolthoff & Göker, 2019). This has been shown for *M. smegmatis* MC2 155, *N. farcinica* NCTC11134, *M. marinum* ATCC927 and *T. paurometabola* DSM2016. With individual barcodes from enzyme extraction having >90% alignment for *T. paurometaabola* and *M. smegmatis*, with the power increasing if barcodes are pooled. However using 700 µl sample to obtain this consistency is still impressive from such famously resilient bacteria. Future adaptations to the protocol in **Chapter 4** could be done to increase the quality of sequencing as well as the testing and optimisation for specific species, although we have presented a broad protocol.

Higher power pipelines do exist such as using Kraken2 (Wood et al., 2019) and PlasmidFinder (Carattoli et al., 2014), however these involve the extensive use and knowledge and use of Linux, an operating system that can be uncommon to be able to use. Additionally, these programmes involve the downloading of databases which can be limiting based on permissions to the computers available. Or the use of a server such as CLIMB-BIG-DATA (Connor et al., 2016), which allows for the analysis for genome through cloud-based computing through a virtual machine.

The use of our novel mycobacterial lysis regents with an adapted protocol in conjunction with the *de novo* assembly pipeline provides a low-cost method for mycobacterial DNA and sequencing. Although not a replacement for more complex techniques designed purely towards high molecular weight DNA extraction, we have provided a more accessible, method which reduces the need for toxic chemical such as phenol: chloroform or expensive equipment such as bead beating.

References

Abrahams, K. A., & Besra, G. S. (2018). Mycobacterial cell wall biosynthesis: A multifaceted antibiotic target. In *Parasitology* (Vol. 145, Issue 2). https://doi.org/10.1017/S0031182016002377

Alderwick, L. J., Harrison, J., Lloyd, G. S., & Birch, H. L. (2015). The mycobacterial cell wall—peptidoglycan and arabinogalactan. *Cold Spring Harbor Perspectives in Medicine*. https://doi.org/10.1101/cshperspect.a021113

Aliberti, S., Sotgiu, G., Castellotti, P., Ferrarese, M., Pancini, L., Pasat, A., Vanoni, N., Spotti, M., Mazzola, E., Gramegna, A., Saderi, L., Perno, C. F., van Ingen, J., Codecasa, L. R., & Blasi, F. (2020). Real-life evaluation of clinical outcomes in patients undergoing treatment for non-tuberculous mycobacteria lung disease: A ten-year cohort study. *Respiratory Medicine*, *164*. https://doi.org/10.1016/j.rmed.2020.105899

Al-Jourani, O., Benedict, S. T., Ross, J., Layton, A. J., van der Peet, P., Marando, V. M., Bailey, N. P., Heunis, T., Manion, J., Mensitieri, F., Franklin, A., Abellon-Ruiz, J., Oram, S. L., Parsons, L., Cartmell, A., Wright, G. S. A., Baslé, A., Trost, M., Henrissat, B., … Moynihan, P. J. (2023). Identification of d-arabinan-degrading enzymes in mycobacteria. *Nature Communications 2023 14:1*, *14*(1), 1–14. https://doi.org/10.1038/s41467-023-37839-5

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3). https://doi.org/10.1016/S0022-2836(05)80360-2

Amaro, A., Duarte, E., Amado, A., Ferronha, H., & Botelho, A. (2008). Comparison of three DNA extraction methods for Mycobacterium bovis, Mycobacterium tuberculosis and Mycobacterium avium subsp. avium. *Letters in Applied Microbiology*, *47*(1), 8–11. https://doi.org/10.1111/j.1472-765X.2008.02372.x

Ashkenazy, H., Abadi, S., Martz, E., Chay, O., Mayrose, I., Pupko, T., & Ben-Tal, N. (2016). ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Research*, *44*(W1). https://doi.org/10.1093/NAR/GKW408

Aşır, S., Türkmen, D., & Topçu, A. A. (2016). DNA Purification by Solid Phase Extraction (SPE) Methods. *Hacettepe Journal of Biology and Chemistry*, *3*(44). https://doi.org/10.15671/hjbc.20164420568

Assegehegn, G., Brito-de la Fuente, E., Franco, J. M., & Gallegos, C. (2019). The Importance of Understanding the Freezing Step and Its Impact on Freeze-Drying Process Performance. In *Journal of Pharmaceutical Sciences* (Vol. 108, Issue 4). https://doi.org/10.1016/j.xphs.2018.11.039

Aujoulat, F., Roger, F., Bourdier, A., Lotthé, A., Lamy, B., Marchandin, H., & Jumas-Bilak, E. (2012). From environment to man: Genome evolution and adaptation of human opportunistic bacterial pathogens. In *Genes* (Vol. 3, Issue 2). https://doi.org/10.3390/genes3020191

Bainomugisa, A., Duarte, T., Lavu, E., Pandey, S., Coulter, C., Marais, B. J., & Coin, L. M. (2018). A complete high-quality MinION nanopore assembly of an extensively drug-resistant

Mycobacterium tuberculosis Beijing lineage strain identifies novel variation in repetitive PE/PPE gene regions. *Microbial Genomics*, *4*(7). https://doi.org/10.1099/mgen.0.000188

Ballou, C. E., Vilkas, E., & Lederer, E. (1963). Structural studies on the myo-inositol phospholipids of Mycobacterium tuberculosis (var. bovis, strain BCG). *The Journal of Biological Chemistry*, *238*. https://doi.org/10.1016/s0021-9258(19)83963-7

BañUls, A. L., Sanou, A., Van Anh, N. T., & Godreuil, S. (2015). Mycobacterium tuberculosis: Ecology and evolution of a human bacterium. In *Journal of Medical Microbiology* (Vol. 64, Issue 11). https://doi.org/10.1099/jmm.0.000171

Baolin Liu, & Xinli Zhou. (2014). Freeze-Drying of Proteins. *Cryopreservation and Freeze-Drying Protocols*, 459–476.

Barkan, D., Hedhli, D., Yan, H. G., Huygen, K., & Glickman, M. S. (2012). Mycobacterium tuberculosis lacking all mycolic acid cyclopropanation is viable but highly attenuated and hyperinflammatory in mice. *Infection and Immunity*, *80*(6). https://doi.org/10.1128/IAI.00021-12

Batt, S. M., Burke, C. E., Moorey, A. R., & Besra, G. S. (2020). Antibiotics and resistance: the two-sided coin of the mycobacterial cell wall. In *Cell Surface* (Vol. 6). https://doi.org/10.1016/j.tcsw.2020.100044

Bhatnagar, B., & Tchessalov, S. (2020). Advances in freeze drying of biologics and future challenges and opportunities. In *Drying Technologies for Biotechnology and Pharmaceutical Applications*. https://doi.org/10.1002/9783527802104.ch6

Birch, H. L., Alderwick, L. J., Bhatt, A., Rittmann, D., Krumbach, K., Singh, A., Bai, Y., Lowary, T. L., Eggeling, L., & Besra, G. S. (2008). Biosynthesis of mycobacterial arabinogalactan: Identification of a novel α(1→3) arabinofuranosyltransferase. *Molecular Microbiology*, *69*(5). https://doi.org/10.1111/j.1365-2958.2008.06354.x

Bjursell, M. K., Martens, E. C., & Gordon, J. I. (2006). Functional genomic and metabolic studies of the adaptations of a prominent adult human gut symbiont, Bacteroides thetaiotaomicron, to the suckling period. *Journal of Biological Chemistry*. https://doi.org/10.1074/jbc.M606509200

Bouso, J. M., & Planet, P. J. (2019). Complete nontuberculous mycobacteria whole genomes using an optimized DNA extraction protocol for long-read sequencing. *BMC Genomics*, *20*(1), 793. https://doi.org/10.1186/S12864-019-6134-Y/FIGURES/3

Brennan, P. J. (2003). Structure, function, and biogenesis of the cell wall of Mycobacterium tuberculosis. *Tuberculosis*, *83*(1–3). https://doi.org/10.1016/S1472-9792(02)00089-6

Brennan, P. J., & Nikaido, H. (1995). The envelope of mycobacteria. In *Annual Review of Biochemistry*. https://doi.org/10.1146/annurev.bi.64.070195.000333

Brown-Elliott, B. A., Vasireddy, S., Vasireddy, R., Iakhiaeva, E., Howard, S. T., Nash, K., Parodi, N., Strong, A., Gee, M., Smith, T., & Wallace, R. J. (2015). Utility of sequencing the erm(41) gene in isolates of Mycobacterium abscessus subsp. abscessus with low and intermediate clarithromycin MICs. *Journal of Clinical Microbiology*, *53*(4). https://doi.org/10.1128/JCM.02950-14

Bryant, J. M., Brown, K. P., Burbaud, S., Everall, I., Belardinelli, J. M., Rodriguez-Rincon, D., Grogono, D. M., Peterson, C. M., Verma, D., Evans, I. E., Ruis, C., Weimann, A., Arora, D., Malhotra, S., Bannerman, B., Passemar, C., Templeton, K., MacGregor, G., Jiwa, K., … Floto, R. A. (2021). Stepwise pathogenic evolution of Mycobacterium abscessus. *Science*, *372*(6541). https://doi.org/10.1126/science.abb8699

Burkovski, A. (2013). Cell Envelope of Corynebacteria: Structure and Influence on Pathogenicity. *ISRN Microbiology*, *2013*, 1–11. https://doi.org/10.1155/2013/935736

Campanero-Rhodes, M. A., Palma, A. S., Menéndez, M., & Solís, D. (2020). Microarray Strategies for Exploring Bacterial Surface Glycans and Their Interactions With Glycan-Binding Proteins. In *Frontiers in Microbiology* (Vol. 10). https://doi.org/10.3389/fmicb.2019.02909

Carattoli, A., Zankari, E., Garciá-Fernández, A., Larsen, M. V., Lund, O., Villa, L., Aarestrup, F. M., & Hasman, H. (2014). In Silico detection and typing of plasmids using plasmidfinder and plasmid multilocus sequence typing. *Antimicrobial Agents and Chemotherapy*, *58*(7). https://doi.org/10.1128/AAC.02412-14

Carpenter, J. F., Chang, B. S., Garzon-Rodriguez, W., & Randolph, T. W. (2002). Rational design of stable lyophilized protein formulations: theory and practice. In *Pharmaceutical biotechnology* (Vol. 13). https://doi.org/10.1007/978-1-4615-0557-0_5

CDC. (2016). *Signs and Symptoms | Nocardiosis | CDC*. https://www.cdc.gov/nocardiosis/symptoms/index.html

Chan, C. E., Götze, S., Seah, G. T., Seeberger, P. H., Tukvadze, N., Wenk, M. R., Hanson, B. J., & MacAry, P. A. (2015). The diagnostic targeting of a carbohydrate virulence factor from M.Tuberculosis. *Scientific Reports*, *5*. https://doi.org/10.1038/srep10281

Chen, Y., Nie, F., Xie, S. Q., Zheng, Y. F., Dai, Q., Bray, T., Wang, Y. X., Xing, J. F., Huang, Z. J., Wang, D. P., He, L. J., Luo, F., Wang, J. X., Liu, Y. Z., & Xiao, C. Le. (2021). Efficient assembly of nanopore reads via highly accurate and intact error correction. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-020-20236-7

Cherf, G. M., Lieberman, K. R., Rashid, H., Lam, C. E., Karplus, K., Akeson, M., & Biotechnol, N. (2012). Automated Forward and Reverse Ratcheting of DNA in a Nanopore at Five Angstrom Precision 1 HHS Public Access Author manuscript. *Nat Biotechnol*.

Ciancia, M., Fernández, P. V., & Leliaert, F. (2020). Diversity of Sulfated Polysaccharides From Cell Walls of Coenocytic Green Algae and Their Structural Relationships in View of Green Algal Evolution. In *Frontiers in Plant Science* (Vol. 11). https://doi.org/10.3389/fpls.2020.554585

Cole, S. T., Eiglmeier, K., Parkhill, J., James, K. D., Thomson, N. R., Wheeler, P. R., Honoré, N., Garnier, T., Churcher, C., Harris, D., Mungall, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R. M., Devlin, K., Duthoy, S., Feltwell, T., … Barrell, B. G. (2001). Massive gene decay in the leprosy bacillus. *Nature*. https://doi.org/10.1038/35059006

Comín, J., Cebollada, A., Iglesias, M. J., Ibarz, D., Viñuelas, J., Torres, L., Sahagún, J., Lafoz, M. C., Esteban de Juanas, F., Malo, M. C., & Samper, S. (2022). Estimation of the mutation rate of Mycobacterium tuberculosis in cases with recurrent tuberculosis using whole genome sequencing. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-21144-0

Connor, T. R., Loman, N. J., Thompson, S., Smith, A., Southgate, J., Poplawski, R., Bull, M. J., Richardson, E., Ismail, M., Thompson, S. E., Kitchen, C., Guest, M., Bakke, M., Sheppard, S. K., & Pallen, M. J. (2016). CLIMB (the Cloud Infrastructure for Microbial Bioinformatics): an online resource for the medical microbiology community. *Microbial Genomics*, *2*(9). https://doi.org/10.1099/mgen.0.000086

Crooks, G. E., Hon, G., Chandonia, J. M., & Brenner, S. E. (2004). WebLogo: A sequence logo generator. *Genome Research*, *14*(6). https://doi.org/10.1101/gr.849004

Daffe, M., Brennan, P. J., & McNeil, M. (1990). Predominant structural features of the cell wall arabinogalactan of Mycobacterium tuberculosis as revealed through characterization of oligoglycosyl alditol fragments by gas chromatography/mass spectrometry and by 1H and 13C NMR analyses. *Journal of Biological Chemistry*, *265*(12). https://doi.org/10.1016/s0021-9258(19)39211-7

Daffe, M., McNeil, M., & Brennan, P. J. (1993). Major structural features of the cell wall arabinogalactans of Mycobacterium, Rhodococcus, and Nocardia spp. *Carbohydrate Research*, *249*(2). https://doi.org/10.1016/0008-6215(93)84102-C

Daher, W., Leclercq, L. D., Johansen, M. D., Hamela, C., Karam, J., Trivelli, X., Nigou, J., Guérardel, Y., & Kremer, L. (2022). Glycopeptidolipid glycosylation controls surface properties and pathogenicity in Mycobacterium abscessus. *Cell Chemical Biology*, *29*(5). https://doi.org/10.1016/j.chembiol.2022.03.008

Daher, W., Leclercq, L. D., Viljoen, A., Karam, J., Dufrêne, Y. F., Guérardel, Y., & Kremer, L. (2020). O-Methylation of the Glycopeptidolipid Acyl Chain Defines Surface Hydrophobicity of Mycobacterium abscessus and Macrophage Invasion. *ACS Infectious Diseases*, *6*(10). https://doi.org/10.1021/acsinfecdis.0c00490

Dahl, V. N., Mølhave, M., Fløe, A., van Ingen, J., Schön, T., Lillebaek, T., Andersen, A. B., & Wejse, C. (2022). Global trends of pulmonary infections with nontuberculous mycobacteria: a systematic review. *International Journal of Infectious Diseases*, *125*, 120–131. https://doi.org/10.1016/J.IJID.2022.10.013

Daito, H., Kikuchi, T., Sakakibara, T., Gomi, K., Damayanti, T., Zaini, J., Tode, N., Kanehira, M., Koyama, S., Fujimura, S., Ebina, M., Ishii, K. J., Akira, S., Takai, T., Watanabe, A., & Nukiwa, T. (2011). Mycobacterial hypersensitivity pneumonitis requires TLR9-MyD88 in lung CD11b+ CD11c+ cells. *European Respiratory Journal*, *38*(3). https://doi.org/10.1183/09031936.00177110

Davies, G., & Henrissat, B. (1995). Structures and mechanisms of glycosyl hydrolases. *Structure*. https://doi.org/10.1016/S0969-2126(01)00220-9

Davies, G. J., & Sinnott, M. L. (2008). Sorting the diverse: the sequence-based classifications of carbohydrate-active enzymes. *Biochemical Journal*. https://doi.org/10.1042/bj20080382

Davies, G. J., Wilson, K. S., & Henrissat, B. (1997). Nomenclature for sugar-binding subsites in glycosyl hydrolases. In *Biochemical Journal*. https://doi.org/10.1042/bj3210557

De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M., & Van Broeckhoven, C. (2018). NanoPack: Visualizing and processing long-read sequencing data. *Bioinformatics*, *34*(15). https://doi.org/10.1093/bioinformatics/bty149

Delahaye, C., & Nicolas, J. (2021). Sequencing DNA with nanopores: Troubles and biases. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0257521

Di Marco, F., Spitaleri, A., Battaglia, S., Batignani, V., Cabibbe, A. M., & Cirillo, D. M. (2023). Advantages of long- and short-reads sequencing for the hybrid investigation of the Mycobacterium tuberculosis genome. *Frontiers in Microbiology*, *14*, 1104456. https://doi.org/10.3389/FMICB.2023.1104456/BIBTEX

Dippenaar, A., Goossens, S. N., Grobbelaar, M., Oostvogels, S., Cuypers, B., Laukens, K., Meehan, C. J., Warren, R. M., & van Rie, A. (2022). Nanopore Sequencing for Mycobacterium tuberculosis : a Critical Review of the Literature, New Developments, and Future Opportunities. In *Journal of Clinical Microbiology* (Vol. 60, Issue 1). https://doi.org/10.1128/JCM.00646-21

Dokic, A., Peterson, E., Arrieta-Ortiz, M. L., Pan, M., Di Maio, A., Baliga, N., & Bhatt, A. (2021). Mycobacterium abscessus biofilms produce an extracellular matrix and have a distinct mycolic acid profile. *Cell Surface (Amsterdam, Netherlands)*, *7*. https://doi.org/10.1016/J.TCSW.2021.100051

Drmanac, R., Drmanac, S., Chui, G., Diaz, R., Hou, A., Jin, H., Jin, P., Kwon, S., Lacy, S., Moeur, B., Shafto, J., Swanson, D., Ukrainczyk, T., Xu, C., & Little, D. (2002). Sequencing by hybridization (SBH): advantages, achievements, and opportunities. In *Advances in biochemical engineering/biotechnology* (Vol. 77). https://doi.org/10.1007/3-540-45713-5_5

Duggal, S. D., & Chugh, T. Das. (2020). Nocardiosis: A Neglected Disease. In *Medical Principles and Practice* (Vol. 29, Issue 6). https://doi.org/10.1159/000508717

Dulberger, C. L., Rubin, E. J., & Boutte, C. C. (2020). The mycobacterial cell envelope — a moving target. In *Nature Reviews Microbiology*. https://doi.org/10.1038/s41579-019-0273-7

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5). https://doi.org/10.1093/nar/gkh340

Ellis, M., Egelund, J., Schultz, C. J., & Bacic, A. (2010). Arabinogalactan-proteins: Key regulators at the cell surface? *Plant Physiology*, *153*(2). https://doi.org/10.1104/pp.110.156000

Epperson, L. E., & Strong, M. (2020). *A scalable, efficient, and safe method to prepare high quality DNA from mycobacteria and other challenging cells*. https://doi.org/10.1016/j.jctube.2020.100150

Fatahi-Bafghi, M. (2018). Nocardiosis from 1888 to 2017. In *Microbial Pathogenesis* (Vol. 114). https://doi.org/10.1016/j.micpath.2017.11.012

Forbes, B. A., Hall, G. S., Miller, M. B., Novak, S. M., Rowlinson, M. C., Salfinger, M., Somoskövi, A., Warshauer, D. M., & Wilson, M. L. (2018). Practice guidelines for clinical microbiology laboratories: Mycobacteria. *Clinical Microbiology Reviews*, *31*(2). https://doi.org/10.1128/CMR.00038-17/FORMAT/EPUB

Ford, C. B., Lin, P. L., Chase, M. R., Shah, R. R., Iartchouk, O., Galagan, J., Mohaideen, N., Ioerger, T. R., Sacchettini, J. C., Lipsitch, M., Flynn, J. L., & Fortune, S. M. (2011). Use of whole genome sequencing to estimate the mutation rate of Mycobacterium tuberculosis during latent infection. *Nature Genetics*, *43*(5). https://doi.org/10.1038/ng.811

Frieden, T. R., Sterling, T. R., Munsiff, S. S., Watt, C. J., & Dye, C. (2003). Tuberculosis. *Lancet*, *362*(9387). https://doi.org/10.1016/S0140-6736(03)14333-4

Gagneux, S. (2018). Ecology and evolution of Mycobacterium tuberculosis. In *Nature Reviews Microbiology* (Vol. 16, Issue 4). https://doi.org/10.1038/nrmicro.2018.8

Garde, S., Chodisetti, P. K., & Reddy, M. (2021). Peptidoglycan: Structure, Synthesis, and Regulation. *EcoSal Plus*, *9*(2). https://doi.org/10.1128/ecosalplus.esp-0010-2020

Gil, F., Grzegorzewicz, A. E., Catalão, M. J., Vital, J., McNeil, M. R., & Pimentel, M. (2010). Mycobacteriophage Ms6 LysB specifically targets the outer membrane of Mycobacterium smegmatis. *Microbiology*, *156*(5). https://doi.org/10.1099/mic.0.032821-0

Glassroth, J. (2008). Pulmonary disease due to nontuberculous mycobacteria. *Chest*, *133*(1). https://doi.org/10.1378/chest.07-0358

Gluckman, P., Beedle, A., Buklijas, T., Low, F., & Hanson, M. (2016). Coevolution, infection, and immunity. In *Principles of Evolutionary Medicine*. https://doi.org/10.1093/acprof:oso/9780199663927.003.0010

Gordon, R. E., & Barnett, D. A. (1977). Resistance to rifampin and lysozyme of strains of some species of Mycobacterium and Nocardia as a taxonomic tool. *International Journal of Systematic Bacteriology*, *27*(3). https://doi.org/10.1099/00207713-27-3-176

Goring, S. M., Wilson, J. B., Risebrough, N. R., Gallagher, J., Carroll, S., Heap, K. J., Obradovic, M., Loebinger, M. R., & Diel, R. (2018). The cost of Mycobacterium avium complex lung disease in Canada, France, Germany, and the United Kingdom: a nationally representative observational study. *BMC Health Services Research*. https://doi.org/10.1186/s12913-018-3489-8

Gray, D. A., White, J. B. R., Oluwole, A. O., Rath, P., Glenwright, A. J., Mazur, A., Zahn, M., Baslé, A., Morland, C., Evans, S. L., Cartmell, A., Robinson, C. V., Hiller, S., Ranson, N. A., Bolam, D. N., & van den Berg, B. (2021). Insights into SusCD-mediated glycan import by a prominent gut symbiont. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-020-20285-y

Gregg, K. J., Zandberg, W. F., Hehemann, J. H., Whitworth, G. E., Deng, L., Vocadlo, D. J., & Boraston, A. B. (2011). Analysis of a new family of widely distributed metal-independent α-mannosidases provides unique insight into the processing of N-linked glycans. *Journal of Biological Chemistry*, *286*(17). https://doi.org/10.1074/jbc.M111.223172

Griffith, D. E., Aksamit, T., Brown-Elliott, B. A., Catanzaro, A., Daley, C., Gordin, F., Holland, S. M., Horsburgh, R., Huitt, G., Iademarco, M. F., Iseman, M., Olivier, K., Ruoss, S., Von Reyn, C. F., Wallace, R. J., & Winthrop, K. (2007). An official ATS/IDSA statement: Diagnosis, treatment, and prevention of nontuberculous mycobacterial diseases. In *American Journal of Respiratory and Critical Care Medicine* (Vol. 175, Issue 4, pp. 367–416). American Thoracic Society. https://doi.org/10.1164/rccm.200604-571ST

Grondin, J. M., Tamura, K., Déjean, G., Abbott, D. W., & Brumer, H. (2017). Polysaccharide utilization loci: Fueling microbial communities. In *Journal of Bacteriology* (Vol. 199, Issue 15). https://doi.org/10.1128/JB.00860-16

Grootaert, H., van Landuyt, L., Hulpiau, P., & Callewaert, N. (2020). Functional exploration of the GH29 fucosidase family. *Glycobiology*, *30*(9). https://doi.org/10.1093/glycob/cwaa023

Gucwa, M., Lenkiewicz, J., Zheng, H., Cymborowski, M., Cooper, D. R., Murzyn, K., & Minor, W. (2023). CMM—An enhanced platform for interactive validation of metal binding sites. *Protein Science*, *32*(1). https://doi.org/10.1002/pro.4525

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: Quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8). https://doi.org/10.1093/bioinformatics/btt086

Gutiérrez, A. V., Viljoen, A., Ghigo, E., Herrmann, J. L., & Kremer, L. (2018). Glycopeptidolipids, a double-edged sword of the Mycobacterium abscessus complex. In *Frontiers in Microbiology* (Vol. 9, Issue JUN). https://doi.org/10.3389/fmicb.2018.01145

Hall, M. B., Rabodoarivelo, M. S., Koch, A., Dippenaar, A., George, S., Grobbelaar, M., Warren, R., Walker, T. M., Cox, H., Gagneux, S., Crook, D., Peto, T., Rakotosamimanana, N., Grandjean Lapierre, S., & Iqbal, Z. (2023). Evaluation of Nanopore sequencing for Mycobacterium tuberculosis drug susceptibility testing and outbreak investigation: a genomic analysis. *The Lancet Microbe*, *4*(2). https://doi.org/10.1016/S2666-5247(22)00301-9

Halla, F. F., Massawa, S. M., Joseph, E. K., Acharya, K., Sabai, S. M., Mgana, S. M., & Werner, D. (2022). Attenuation of bacterial hazard indicators in the subsurface of an informal settlement and their application in quantitative microbial risk assessment. *Environment International*, *167*. https://doi.org/10.1016/j.envint.2022.107429

Haworth, C. S., Banks, J., Capstick, T., Fisher, A. J., Gorsuch, T., Laurenson, I. F., Leitch, A., Loebinger, M. R., Milburn, H. J., Nightingale, M., Ormerod, P., Shingadia, D., Smith, D., Whitehead, N., Wilson, R., & Floto, R. A. (2017a). British Thoracic Society guidelines for the management of non-tuberculous mycobacterial pulmonary disease (NTM-PD). In *Thorax*. https://doi.org/10.1136/thoraxjnl-2017-210927

Haworth, C. S., Banks, J., Capstick, T., Fisher, A. J., Gorsuch, T., Laurenson, I. F., Leitch, A., Loebinger, M. R., Milburn, H. J., Nightingale, M., Ormerod, P., Shingadia, D., Smith, D., Whitehead, N., Wilson, R., & Floto, R. A. (2017b). British Thoracic Society guidelines for the management of non-tuberculous mycobacterial pulmonary disease (NTM-PD). In *Thorax*. https://doi.org/10.1136/thoraxjnl-2017-210927

Healy, C., Gouzy, A., & Ehrt, S. (2020). Peptidoglycan hydrolases RipA and Ami1 are critical for replication and persistence of Mycobacterium tuberculosis in the host. *MBio*, *11*(2). https://doi.org/10.1128/mBio.03315-19

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. In *Genomics* (Vol. 107, Issue 1). https://doi.org/10.1016/j.ygeno.2015.11.003

Hendrix, J., Epperson, L. E., Tong, E. I., Chan, Y. L., Hasan, N. A., Dawrs, S. N., Norton, G. J., Virdi, R., Crooks, J. L., Chan, E. D., Honda, J. R., & Strong, M. (2023). Complete genome assembly of Hawai'i environmental nontuberculous mycobacteria reveals unexpected co-isolation with methylobacteria. *PLoS ONE*, *18*(9 September). https://doi.org/10.1371/journal.pone.0291072

Henrissat, B. (1991). A classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochemical Journal*. https://doi.org/10.1042/bj2800309

Henrissat, B., & Davies, G. (1997). Structural and sequence-based classification of glycoside hydrolases. *Current Opinion in Structural Biology*, *7*(5). https://doi.org/10.1016/S0959-440X(97)80072-3

Hook, P. W., & Timp, W. (2023). Beyond assembly: the increasing flexibility of single-molecule sequencing technology. In *Nature Reviews Genetics* (Vol. 24, Issue 9). https://doi.org/10.1038/s41576-023-00600-1

Hou, K., Wu, Z. X., Chen, X. Y., Wang, J. Q., Zhang, D., Xiao, C., Zhu, D., Koya, J. B., Wei, L., Li, J., & Chen, Z. S. (2022). Microbiota in health and diseases. In *Signal Transduction and Targeted Therapy* (Vol. 7, Issue 1). https://doi.org/10.1038/s41392-022-00974-4

Huber, R., Langworthy, T. A., König, H., Thomm, M., Woese, C. R., Sleytr, U. B., & Stetter, K. O. (1986). Thermotoga maritima sp. nov. represents a new genus of unique extremely thermophilic eubacteria growing up to 90°C. *Archives of Microbiology*, *144*(4). https://doi.org/10.1007/BF00409880

Jain, M. (2023). *From kilobases to "whales": a short history of ultra-long reads and high-throughput genome sequencing*. https://nanoporetech.com/about-us/news/blog-kilobases-whales-short-history-ultra-long-reads-and-high-throughput-genome

Jia, B., Han, X., Kim, K. H., & Jeon, C. O. (2022). Discovery and mining of enzymes from the human gut microbiome. In *Trends in Biotechnology* (Vol. 40, Issue 2). https://doi.org/10.1016/j.tibtech.2021.06.008

Johansen, M. D., Herrmann, J. L., & Kremer, L. (2020). Non-tuberculous mycobacteria and the rise of Mycobacterium abscessus. In *Nature Reviews Microbiology* (Vol. 18, Issue 7). https://doi.org/10.1038/s41579-020-0331-1

Jou, W. M., Haegeman, G., Ysebaert, M., & Fiers, W. (1972). Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. *Nature*, *237*(5350). https://doi.org/10.1038/237082a0

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873). https://doi.org/10.1038/s41586-021-03819-2

Kam, J. Y., Hortle, E., Krogman, E., Warner, S. E., Wright, K., Luo, K., Cheng, T., Manuneedhi Cholan, P., Kikuchi, K., Triccas, J. A., Britton, W. J., Johansen, M. D., Kremer, L., & Oehlers, S. H. (2022). Rough and smooth variants of Mycobacterium abscessus are differentially controlled by host immunity during chronic infection of adult zebrafish. *Nature Communications*, *13*(1). https://doi.org/10.1038/s41467-022-28638-5

Kanetsuna, F. (1980). Effect of Lysozyme on Mycobacteria. *MICROBIOLOGY and IMMUNOLOGY*, *24*(12). https://doi.org/10.1111/j.1348-0421.1980.tb02920.x

Kaoutari, A. El, Armougom, F., Gordon, J. I., Raoult, D., & Henrissat, B. (2013). The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. *Nature Reviews Microbiology*, *11*(7). https://doi.org/10.1038/nrmicro3050

Käser, M., Ruf, M. T., Hauser, J., Marsollier, L., & Pluschke, G. (2009). Optimized method for preparation of DNA from pathogenic and environmental mycobacteria. *Applied and Environmental Microbiology*, *75*(2). https://doi.org/10.1128/AEM.01358-08

Käser, M., Ruf, M. T., Hauser, J., & Pluschke, G. (2010). Optimized DNA preparation from mycobacteria. *Cold Spring Harbor Protocols*, *5*(4). https://doi.org/10.1101/pdb.prot5408

Kaso, A. W., & Hailu, A. (2021). Costs and cost-effectiveness of Gene Xpert compared to smear microscopy for the diagnosis of pulmonary tuberculosis using real-world data from Arsi zone, Ethiopia. *PLoS ONE*, *16*(10 October). https://doi.org/10.1371/journal.pone.0259056

Khawbung, J. L., Nath, D., & Chakraborty, S. (2021). Drug resistant Tuberculosis: A review. In *Comparative Immunology, Microbiology and Infectious Diseases* (Vol. 74). https://doi.org/10.1016/j.cimid.2020.101574

Kim, H. Y., Kim, B. J., Kook, Y., Yun, Y. J., Shin, J. H., Kim, B. J., & Kook, Y. H. (2010). Mycobacterium massiliense is differentiated from mycobacterium abscessus and mycobacterium bolletii by erythromycin ribosome methyltransferase gene (erm) and clarithromycin susceptibility patterns. *Microbiology and Immunology*, *54*(6). https://doi.org/10.1111/j.1348-0421.2010.00221.x

Kolmogorov, M., Bickhart, D. M., Behsaz, B., Gurevich, A., Rayko, M., Shin, S. B., Kuhn, K., Yuan, J., Polevikov, E., Smith, T. P. L., & Pevzner, P. A. (2020). metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nature Methods*, *17*(11). https://doi.org/10.1038/s41592-020-00971-x

Korf, J., Stoltz, A., Verschoor, J., De Baetselier, P., & Grooten, J. (2005). The Mycobacterium tuberculosis cell wall component mycolic acid elicits pathogen-associated host innate immune responses. *European Journal of Immunology*, *35*(3). https://doi.org/10.1002/eji.200425332

Köser, C. U., Ellington, M. J., & Peacock, S. J. (2014). Whole-genome sequencing to control antimicrobial resistance. In *Trends in Genetics*. https://doi.org/10.1016/j.tig.2014.07.003

Kumar, P., Marathe, S., & Bhaskar, S. (2016). Isolation of Genomic DNA from Mycobacterium Species. *BIO-PROTOCOL*, *6*(5). https://doi.org/10.21769/bioprotoc.1751

Kus, J. V., Kelly, J., Tessier, L., Harvey, H., Cvitkovitch, D. G., & Burrows, L. L. (2008). Modification of Pseudomonas aeruginosa Pa5196 type IV pilins at multiple sites with D-Araf by a novel GT-C family arabinosyltransferase, TfpW. *Journal of Bacteriology*, *190*(22). https://doi.org/10.1128/JB.01075-08

Lafont, E., Conan, P. L., Rodriguez-Nava, V., & Lebeaux, D. (2020a). Invasive nocardiosis: Disease presentation, diagnosis and treatment – old questions, new answers? In *Infection and Drug Resistance* (Vol. 13). https://doi.org/10.2147/IDR.S249761

Lafont, E., Conan, P. L., Rodriguez-Nava, V., & Lebeaux, D. (2020b). Invasive nocardiosis: Disease presentation, diagnosis and treatment – old questions, new answers? In *Infection and Drug Resistance*. https://doi.org/10.2147/IDR.S249761

Laver, T., Harrison, J., O'Neill, P. A., Moore, K., Farbos, A., Paszkiewicz, K., & Studholme, D. J. (2015). Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomolecular Detection and Quantification*. https://doi.org/10.1016/j.bdq.2015.02.001

Lee, S. H. (2016). Tuberculosis infection and latent tuberculosis. In *Tuberculosis and Respiratory Diseases* (Vol. 79, Issue 4). https://doi.org/10.4046/trd.2016.79.4.201

Lee, Y. C., Wu, H. M., Chang, Y. N., Wang, W. C., & Hsu, W. H. (2007). The Central Cavity from the (Alpha/Alpha)6 Barrel Structure of Anabaena sp. CH1 N-Acetyl-d-glucosamine 2-Epimerase Contains Two Key Histidine Residues for Reversible Conversion. *Journal of Molecular Biology*, *367*(3). https://doi.org/10.1016/j.jmb.2006.11.001

Leshchiner, I., Alexa, K., Kelsey, P., Adzhubei, I., Austin-Tse, C. A., Cooney, J. D., Anderson, H., King, M. J., Stottmann, R. W., Garnaas, M. K., Ha, S., Drummond, I. A., Paw, B. H., North, T. E., Beier, D. R., Goessling, W., & Sunyaev, S. R. (2012). Mutation mapping and identification by whole-genome sequencing. *Genome Research*, *22*(8). https://doi.org/10.1101/gr.135541.111

Lopez-Marin, L. M., Gautier, N., Laneelle, M. A., Silve, G., & Daffe, M. (1994). Structures of the glycopeptidolipid antigens of Mycobacterium abscessus and Mycobacterium chelonae and possible chemical basis of the serological cross-reactions in the Mycobacterium fortuitum complex. *Microbiology*, *140*(5). https://doi.org/10.1099/13500872-140-5-1109

Maiga, M., Siddiqui, S., Diallo, S., Diarra, B., Traoré, B., Shea, Y. R., Zelazny, A. M., Dembele, B. P. P., Goita, D., Kassambara, H., Hammond, A. S., Polis, M. A., & Tounkara, A. (2012). Failure to Recognize Nontuberculous Mycobacteria Leads to Misdiagnosis of Chronic Pulmonary Tuberculosis. *PLoS ONE*, *7*(5), e36902. https://doi.org/10.1371/journal.pone.0036902

Margalit, I., Lebeaux, D., Tishler, O., Goldberg, E., Bishara, J., Yahav, D., & Coussement, J. (2021). How do I manage nocardiosis? In *Clinical Microbiology and Infection* (Vol. 27, Issue 4). https://doi.org/10.1016/j.cmi.2020.12.019

Marrakchi, H., Lanéelle, M. A., & Daffé, M. (2014). Mycolic acids: Structures, biosynthesis, and beyond. In *Chemistry and Biology* (Vol. 21, Issue 1, pp. 67–85). Cell Press. https://doi.org/10.1016/j.chembiol.2013.11.011

Martens, E. C., Koropatkin, N. M., Smith, T. J., & Gordon, J. I. (2009). Complex glycan catabolism by the human gut microbiota: The bacteroidetes sus-like paradigm. In *Journal of Biological Chemistry* (Vol. 284, Issue 37). https://doi.org/10.1074/jbc.R109.022848

Martens, E. C., Lowe, E. C., Chiang, H., Pudlo, N. A., Wu, M., McNulty, N. P., Abbott, D. W., Henrissat, B., Gilbert, H. J., Bolam, D. N., & Gordon, J. I. (2011). Recognition and degradation of plant cell wall polysaccharides by two human gut symbionts. *PLoS Biology*, *9*(12). https://doi.org/10.1371/journal.pbio.1001221

Martinelli, D. J., & Pavelka, M. S. (2016). The RipA and RipB peptidoglycan endopeptidases are individually nonessential to Mycobacterium smegmatis. *Journal of Bacteriology*, *198*(9). https://doi.org/10.1128/JB.00059-16

Matsumoto, Y., Kinjo, T., Motooka, D., Nabeya, D., Jung, N., Uechi, K., & Horii, T. (2019). Jiro Fujita & Shota Nakamura (2019) Comprehensive subspecies identification of 175 nontuberculous mycobacteria species based on 7547 genomic profiles. *Emerging Microbes & Infections*, *8*(1), 1043–1053. https://doi.org/10.1080/22221751.2019.1637702

McCarter, J. D., & Stephen Withers, G. (1994). Mechanisms of enzymatic glycoside hydrolysis. *Current Opinion in Structural Biology*, *4*(6). https://doi.org/10.1016/0959-440X(94)90271-2

McMurray, D. N. (1996). Chapter 33 Mycobacteria and Nocardia. *Medical Microbiology*.

McNeil, M., Wallner, S. J., Hunter, S. W., & Brennan, P. J. (1987). Demonstration that the galactosyl and arabinosyl residues in the cell-wall arabinogalactan of Mycobacterium leprae and Myobacterium tuberculosis are furanoid. *Carbohydrate Research*, *166*(2). https://doi.org/10.1016/0008-6215(87)80065-4

Meehan, C. J., Goig, G. A., Kohl, T. A., Verboven, L., Dippenaar, A., Ezewudo, M., Farhat, M. R., Guthrie, J. L., Laukens, K., Miotto, P., Ofori-Anyinam, B., Dreyer, V., Supply, P., Suresh, A., Utpatel, C., van Soolingen, D., Zhou, Y., Ashton, P. M., Brites, D., … Van Rie, A. (2019). Whole genome sequencing of Mycobacterium tuberculosis: current standards and open issues. In *Nature Reviews Microbiology* (Vol. 17, Issue 9). https://doi.org/10.1038/s41579-019-0214-5

Mehta, H. H., & Shamoo, Y. (2020). Pathogenic nocardia: A diverse genus of emerging pathogens or just poorly recognized? In *PLoS Pathogens* (Vol. 16, Issue 3). https://doi.org/10.1371/journal.ppat.1008280

Meier-Kolthoff, J. P., & Göker, M. (2019). TYGS is an automated high-throughput platform for state-of-the-art genome-based taxonomy. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-10210-3

Meng, E. C., Goddard, T. D., Pettersen, E. F., Couch, G. S., Pearson, Z. J., Morris, J. H., & Ferrin, T. E. (2023). UCSF ChimeraX: Tools for structure building and analysis. *Protein Science*, *32*(11). https://doi.org/10.1002/pro.4792

Mensink, M. A., Frijlink, H. W., van der Voort Maarschalk, K., & Hinrichs, W. L. J. (2017). How sugars protect proteins in the solid state and during drying (review): Mechanisms of stabilization in relation to stress conditions. In *European Journal of Pharmaceutics and Biopharmaceutics* (Vol. 114). https://doi.org/10.1016/j.ejpb.2017.01.024

MERCK. (2023). *GenElute™ Bacterial Genomic DNA Kit Protocol (NA2100, NA2110, NA2120)*. https://www.sigmaaldrich.com/GB/en/technical-documents/protocol/genomics/dna-and-rna-purification/genelute-bacterial-genomic-dna-kit

Mikušová, K., Mikuš, M., Besra, G. S., Hancock, I., & Brennan, P. J. (1996). Biosynthesis of the linkage region of the mycobacterial cell wall. *Journal of Biological Chemistry*, *271*(13). https://doi.org/10.1074/jbc.271.13.7820

Mishra, A. K., Driessen, N. N., Appelmelk, B. J., & Besra, G. S. (2011). Lipoarabinomannan and related glycoconjugates: Structure, biogenesis and role in Mycobacterium tuberculosis physiology and host-pathogen interaction. In *FEMS Microbiology Reviews* (Vol. 35, Issue 6, pp. 1126–1157). Oxford Academic. https://doi.org/10.1111/j.1574-6976.2011.00276.x

Moreno-Izquierdo, C., Zurita, J., Contreras-Yametti, F. I., & Jara-Palacios, M. A. (2020). Mycobacterium abscessus subspecies abscessus infection associated with cosmetic surgical procedures: Cases series. *IDCases*, *22*. https://doi.org/10.1016/j.idcr.2020.e00992

Nasiri, M. J., Dabiri, H., Fooladi, A. A. I., Amini, S., Hamzehloo, G., & Feizabadi, M. M. (2018). High rates of nontuberculous mycobacteria isolation from patients with presumptive tuberculosis in Iran. *New Microbes and New Infections*, *21*, 12–17. https://doi.org/10.1016/j.nmni.2017.08.008

Nataraj, V., Varela, C., Javid, A., Singh, A., Besra, G. S., & Bhatt, A. (2015). Mycolic acids: Deciphering and targeting the Achilles' heel of the tubercle bacillus. *Molecular Microbiology*, *98*(1), 7–16. https://doi.org/10.1111/MMI.13101

Odumeru, J., Gao, A., Chen, S., Raymond, M., & Mutharia, L. (2001). Use of the bead beater for preparation of Mycobacterium paratuberculosis template DNA in milk. *Canadian Journal of Veterinary Research*.

Ojha, A. K., Baughn, A. D., Sambandan, D., Hsu, T., Trivelli, X., Guerardel, Y., Alahari, A., Kremer, L., Jacobs, W. R., & Hatfull, G. F. (2008). Growth of Mycobacterium tuberculosis biofilms containing free mycolic acids and harbouring drug-tolerant bacteria. *Molecular Microbiology*, *69*(1), 164–174. https://doi.org/10.1111/J.1365-2958.2008.06274.X

Olson, R. D., Assaf, R., Brettin, T., Conrad, N., Cucinell, C., Davis, J. J., Dempsey, D. M., Dickerman, A., Dietrich, E. M., Kenyon, R. W., Kuscuoglu, M., Lefkowitz, E. J., Lu, J., Machi, D., Macken, C., Mao, C., Niewiadomska, A., Nguyen, M., Olsen, G. J., … Stevens, R. L. (2023). Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): a resource combining PATRIC, IRD and ViPR. *Nucleic Acids Research*, *51*(1 D). https://doi.org/10.1093/nar/gkac1003

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., & Phillippy, A. M. (2016). Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biology*, *17*(1). https://doi.org/10.1186/s13059-016-0997-x

Ozen, M., & Dinleyici, E. C. (2015). The history of probiotics: The untold story. In *Beneficial Microbes* (Vol. 6, Issue 2). https://doi.org/10.3920/BM2014.0103

Pasini, M. E., Intra, J., & Pavesi, G. (2008). Expression study of an α-l-fucosidase gene in the Drosophilidae family. *Gene*, *420*(1). https://doi.org/10.1016/j.gene.2008.04.021

Pervez, M. T., Hasnain, M. J. U., Abbas, S. H., Moustafa, M. F., Aslam, N., & Shah, S. S. M. (2022). A Comprehensive Review of Performance of Next-Generation Sequencing Platforms. In *BioMed Research International* (Vol. 2022). https://doi.org/10.1155/2022/3457806

Player, R., Verratti, K., Staab, A., Forsyth, E., Ernlund, A., Joshi, M. S., Dunning, R., Rozak, D., Grady, S., Goodwin, B., & Sozhamannan, S. (2022). Optimization of Oxford Nanopore Technology Sequencing Workflow for Detection of Amplicons in Real Time Using ONT-DART Tool. *Genes*, *13*(10). https://doi.org/10.3390/genes13101785

Prevots, D. R., & Marras, T. K. (2015). Epidemiology of human pulmonary infection with nontuberculous mycobacteria a review. In *Clinics in Chest Medicine* (Vol. 36, Issue 1). https://doi.org/10.1016/j.ccm.2014.10.002

Prieto, M. D., Alam, M. E., Franciosi, A. N., & Quon, B. S. (2023). Global burden of nontuberculous mycobacteria in the cystic fibrosis population: a systematic review and meta-analysis. *ERJ Open Research*, *9*(1). https://doi.org/10.1183/23120541.00336-2022

Purushothaman, S., Meola, M., & Egli, A. (2022). Combination of Whole Genome Sequencing and Metagenomics for Microbiological Diagnostics. In *International Journal of Molecular Sciences* (Vol. 23, Issue 17). https://doi.org/10.3390/ijms23179834

Quick, J., & Loman, N. J. (2018). Nanopore sequencing book: DNA extraction and purification methods. In *Nanopore Sequencing: An Introduction*.

Raad, R. B., Méniche, X., De Sousa-D'Auria, C., Chami, M., Salmeron, C., Tropis, M., Labarre, C., Daffé, M., Houssin, C., & Bayan, N. (2010). A deficiency in arabinogalactan biosynthesis affects Corynebacterium glutamicum mycolate outer membrane stability. *Journal of Bacteriology*, *192*(11). https://doi.org/10.1128/JB.00009-10

Ratnatunga, C. N., Lutzky, V. P., Kupz, A., Doolan, D. L., Reid, D. W., Field, M., Bell, S. C., Thomson, R. M., & Miles, J. J. (2020). The Rise of Non-Tuberculosis Mycobacterial Lung Disease. In *Frontiers in Immunology*. https://doi.org/10.3389/fimmu.2020.00303

Raymond, J. B., Mahapatra, S., Crick, D. C., & Pavelka, M. S. (2005). Identification of the namH gene, encoding the hydroxylase responsible for the N-glycolylation of the mycobacterial peptidoglycan. *Journal of Biological Chemistry*, *280*(1). https://doi.org/10.1074/jbc.M411006200

Reynolds, J., Moyes, R. B., & Breakwell, D. P. (2009). Differential Staining of Bacteria: Acid Fast Stain. *Current Protocols in Microbiology*, *15*(1). https://doi.org/10.1002/9780471729259.mca03hs15

Rhoads, A., & Au, K. F. (2015). PacBio Sequencing and Its Applications. In *Genomics, Proteomics and Bioinformatics* (Vol. 13, Issue 5). https://doi.org/10.1016/j.gpb.2015.08.002

Rinninella, E., Raoul, P., Cintoni, M., Franceschi, F., Miggiano, G. A. D., Gasbarrini, A., & Mele, M. C. (2019). What is the healthy gut microbiota composition? A changing ecosystem across age, environment, diet, and diseases. *Microorganisms*, *7*(1). https://doi.org/10.3390/microorganisms7010014

Rizal, N. S. M., Neoh, H. M., Ramli, R., Periyasamy, P. R. A. L. K., Hanafiah, A., Samat, M. N. A., Tan, T. L., Wong, K. K., Nathan, S., Chieng, S., Saw, S. H., & Khor, B. Y. (2020). Advantages and limitations of 16S rRNA next-generation sequencing for pathogen identification in the diagnostic microbiology laboratory: perspectives from a middle-income country. In *Diagnostics* (Vol. 10, Issue 10). https://doi.org/10.3390/diagnostics10100816

Roe, K. D., & Labuza, T. P. (2005). Glass transition and crystallization of amorphous trehalose-sucrose mixtures. *International Journal of Food Properties*, *8*(3). https://doi.org/10.1080/10942910500269824

Rogowski, A., Briggs, J. A., Mortimer, J. C., Tryfona, T., Terrapon, N., Lowe, E. C., Baslé, A., Morland, C., Day, A. M., Zheng, H., Rogers, T. E., Thompson, P., Hawkins, A. R., Yadav, M. P., Henrissat, B., Martens, E. C., Dupree, P., Gilbert, H. J., & Bolam, D. N. (2015). Glycan complexity dictates microbial resource allocation in the large intestine. *Nature Communications*, *6*. https://doi.org/10.1038/ncomms8481

Roosaare, M., Puustusmaa, M., Möls, M., Vaher, M., & Remm, M. (2018). PlasmidSeeker: Identification of known plasmids from bacterial whole genome sequencing reads. *PeerJ*, *2018*(4). https://doi.org/10.7717/peerj.4588

Roux, A. L., Viljoen, A., Bah, A., Simeone, R., Bernut, A., Laencina, L., Deramaudt, T., Rottman, M., Gaillard, J. L., Majlessi, L., Brosch, R., Girard-Misguich, F., Vergne, I., de Chastellier, C., Kremer, L., & Herrmann, J. L. (2016). The distinct fate of smooth and rough Mycobacterium abscessus variants inside macrophages. *Open Biology*, *6*(11). https://doi.org/10.1098/rsob.160185

Rouzaud, C., Rodriguez-Nava, V., Catherinot, E., Méchai, F., Bergeron, E., Farfour, E., Scemla, A., Poirée, S., Delavaud, C., Mathieu, D., Durupt, S., Larosa, F., Lengele, J. P., Christophe, J. L., Suarez, F., Lortholary, O., & Lebeauxa, D. (2018). Clinical assessment of a nocardia PCR-Based assay for diagnosis of nocardiosis. *Journal of Clinical Microbiology*, *56*(6). https://doi.org/10.1128/JCM.00002-18

Roy, M. L., Pikal, M. J., Rickard, E. C., & Maloney, A. M. (1991). The effects of formulation and moisture on the stability of a freeze-dried monoclonal antibody - Vinca conjugate: A test of the WLF glass transition theory. *Developments in Biological Standardization*, *74*.

Rüger, K., Hampel, A., Billig, S., Rücker, N., Suerbaum, S., & Bange, F. C. (2014). Characterization of rough and smooth morphotypes of mycobacterium abscessus isolates from clinical specimens. *Journal of Clinical Microbiology*, *52*(1). https://doi.org/10.1128/JCM.01249-13

Runyon, E. H. (1959). Anonymous mycobacteria in pulmonary disease. *The Medical Clinics of North America*. https://doi.org/10.1016/S0025-7125(16)34193-1

Ryu, Y. J., Koh, W. J., & Daley, C. L. (2016). Diagnosis and treatment of nontuberculous mycobacterial lung disease: Clinicians' perspectives. In *Tuberculosis and Respiratory Diseases* (Vol. 79, Issue 2). https://doi.org/10.4046/trd.2016.79.2.74

Sancho-Vaello, E., Albesa-Jové, D., Rodrigo-Unzueta, A., & Guerin, M. E. (2017). Structural basis of phosphatidyl-myo-inositol mannosides biosynthesis in mycobacteria. In *Biochimica et Biophysica Acta - Molecular and Cell Biology of Lipids* (Vol. 1862, Issue 11). https://doi.org/10.1016/j.bbalip.2016.11.002

Sanger, F., Coulson, A. R., Hong, G. F., Hill, D. F., & Petersen, G. B. (1982). Nucleotide sequence of bacteriophage λ DNA. *Journal of Molecular Biology*, *162*(4). https://doi.org/10.1016/0022-2836(82)90546-0

Satta, G., Lipman, M., Smith, G. P., Arnold, C., Kon, O. M., & McHugh, T. D. (2018). Mycobacterium tuberculosis and whole-genome sequencing: how close are we to unleashing its full potential? *Clinical Microbiology and Infection : The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases*, *24*(6), 604–609. https://doi.org/10.1016/J.CMI.2017.10.030

Saxena, S., Spaink, H. P., & Forn-Cuní, G. (2021). Drug resistance in nontuberculous mycobacteria: Mechanisms and models. In *Biology* (Vol. 10, Issue 2). https://doi.org/10.3390/biology10020096

Schwarze, K., Buchanan, J., Fermont, J. M., Dreau, H., Tilley, M. W., Taylor, J. M., Antoniou, P., Knight, S. J. L., Camps, C., Pentony, M. M., Kvikstad, E. M., Harris, S., Popitsch, N., Pagnamenta, A. T., Schuh, A., Taylor, J. C., & Wordsworth, S. (2019). The complete costs of genome sequencing: a microcosting study in cancer and rare diseases from a single center in the United Kingdom. *Genetics in Medicine 2019 22:1*, *22*(1), 85–94. https://doi.org/10.1038/s41436-019-0618-7

Seidel, M., Alderwick, L. J., Birch, H. L., Sahm, H., Eggeling, L., & Besra, G. S. (2007). Identification of a novel arabinofuranosyltransferase AftB involved in a terminal step of cell wall arabinan biosynthesis in Corynebacterianeae, such as Corynebacterium glutamicum and Mycobacterium tuberculosis. *Journal of Biological Chemistry*, *282*(20). https://doi.org/10.1074/jbc.M700271200

Sexton, P., & Harrison, A. C. (2008). Susceptibility to nontuberculous mycobacterial lung disease. In *European Respiratory Journal* (Vol. 31, Issue 6). https://doi.org/10.1183/09031936.00140007

Shen, L., Viljoen, A., Villaume, S., Joe, M., Halloum, I., Chêne, L., Méry, A., Fabre, E., Takegawa, K., Lowary, T. L., Vincent, S. P., Kremer, L., Guérardel, Y., & Mariller, C. (2020). The endogenous galactofuranosidase GlfH1 hydrolyzes mycobacterial arabinogalactan. *Journal of Biological Chemistry*, *295*(15), 5110–5123. https://doi.org/10.1074/jbc.RA119.011817

Shimokawa, M., Ishiwata, A., Kashima, T., Nakashima, C., Li, J., Fukushima, R., Sawai, N., Nakamori, M., Tanaka, Y., Kudo, A., Morikami, S., Iwanaga, N., Akai, G., Shimizu, N., Arakawa, T., Yamada, C., Kitahara, K., Tanaka, K., Ito, Y., … Fujita, K. (2023). Identification and characterization of endo-α-, exo-α-, and exo-β-d-arabinofuranosidases degrading lipoarabinomannan and arabinogalactan of mycobacteria. *Nature Communications*, *14*(1). https://doi.org/10.1038/s41467-023-41431-2

Silva, J., Ferraz, R., Dupree, P., Showalter, A. M., & Coimbra, S. (2020). Three Decades of Advances in Arabinogalactan-Protein Biosynthesis. In *Frontiers in Plant Science* (Vol. 11). https://doi.org/10.3389/fpls.2020.610377

Škovierová, H., Larrouy-Maumus, G., Zhang, J., Kaur, D., Barilone, N., Korduláková, J., Gilleron, M., Guadagnini, S., Belanová, M., Prevost, M. C., Gicquel, B., Puzo, G., Chatterjee, D., Brennan, P. J., Nigou, J., & Jackson, M. (2009). AftD, a novel essential arabinofuranosyltransferase from mycobacteria. *Glycobiology*, *19*(11). https://doi.org/10.1093/glycob/cwp116

Smith, C., Halse, T. A., Shea, J., Modestil, H., Fowler, R. C., Musser, K. A., Escuyer, V., & Lapierre, P. (2021). Assessing nanopore sequencing for clinical diagnostics: A comparison of Next-Generation Sequencing (NGS) methods for mycobacterium tuberculosis. *Journal of Clinical Microbiology*, *59*(1). https://doi.org/10.1128/JCM.00583-20

Sohn, J. Il, & Nam, J. W. (2018). The present and future of de novo whole-genome assembly. *Briefings in Bioinformatics*, *19*(1). https://doi.org/10.1093/bib/bbw096

Solanki, V., Krüger, K., Crawford, C. J., Pardo-Vargas, A., Danglad-Flores, J., Hoang, K. L. M., Klassen, L., Abbott, D. W., Seeberger, P. H., Amann, R. I., Teeling, H., & Hehemann, J. H. (2022). Glycoside hydrolase from the GH76 family indicates that marine Salegentibacter sp. Hel_I_6 consumes alpha-mannan from fungi. *ISME Journal*, *16*(7). https://doi.org/10.1038/s41396-022-01223-w

Starciuc, T., Malfait, B., Danede, F., Paccou, L., Guinet, Y., Correia, N. T., & Hedoux, A. (2020). Trehalose or Sucrose: Which of the Two Should be Used for Stabilizing Proteins in the Solid State? A Dilemma Investigated by In Situ Micro-Raman and Dielectric Relaxation Spectroscopies During and After Freeze-Drying. *Journal of Pharmaceutical Sciences*, *109*(1). https://doi.org/10.1016/j.xphs.2019.10.055

Su, M., Satola, S. W., & Read, T. D. (2019). Genome-based prediction of bacterial antibiotic resistance. In *Journal of Clinical Microbiology* (Vol. 57, Issue 3). https://doi.org/10.1128/JCM.01405-18

Sulzenbacher, G., Bignon, C., Nishimura, T., Tarling, C. A., Withers, S. G., Henrissat, B., & Bourne, Y. (2004). Crystal structure of Thermotoga maritima α-L-fucosidase: Insights into the

catalytic mechanism and the molecular basis for fucosidosis. *Journal of Biological Chemistry*, *279*(13). https://doi.org/10.1074/jbc.M313783200

Sundaramurthi, P., & Suryanarayanan, R. (2010). Trehalose crystallization during freeze-drying: Implications on lyoprotection. *Journal of Physical Chemistry Letters*, *1*(2). https://doi.org/10.1021/jz900338m

Surayot, U., Lee, J. H., Kanongnuch, C., Peerapornpisal, Y., Park, W. J., & You, S. G. (2016). Structural characterization of sulfated arabinans extracted from Cladophora glomerata Kützing and their macrophage activation. *Bioscience, Biotechnology and Biochemistry*, *80*(5). https://doi.org/10.1080/09168451.2015.1132149

Teufel, F., Almagro Armenteros, J. J., Johansen, A. R., Gíslason, M. H., Pihl, S. I., Tsirigos, K. D., Winther, O., Brunak, S., von Heijne, G., & Nielsen, H. (2022). SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nature Biotechnology*, *40*(7). https://doi.org/10.1038/s41587-021-01156-3

Theel, E. S., & Doern, C. D. (2013). β-D-Glucan testing is important for diagnosis of invasive fungal infections. *Journal of Clinical Microbiology*, *51*(11). https://doi.org/10.1128/JCM.01737-13

Tong, M., Jacobi, C. E., Van De Rijke, F. M., Kuijper, S., Van De Werken, S., Lowary, T. L., Hokke, C. H., Appelmelk, B. J., Nagelkerke, N. J. D., Tanke, H. J., Van Gijlswijk, R. P. M., Veuskens, J., Kolk, A. H. J., & Raap, A. K. (2005). A multiplexed and miniaturized serological tuberculosis assay identifies antigens that discriminate maximally between TB and non-TB sera. *Journal of Immunological Methods*, *301*(1–2). https://doi.org/10.1016/j.jim.2005.04.004

Tra, V. N., & Dube, D. H. (2014). Glycans in pathogenic bacteria - potential for targeted covalent therapeutics and imaging agents. *Chemical Communications*, *50*(36). https://doi.org/10.1039/c4cc00660g

Trigodet, F., Lolans, K., Fogarty, E., Shaiber, A., Morrison, H. G., Barreiro, L., Jabri, B., & Eren, A. M. (2022). High molecular weight DNA extraction strategies for long-read sequencing of complex metagenomes. *Molecular Ecology Resources*, *22*(5). https://doi.org/10.1111/1755-0998.13588

Turner, J., & Torrelles, J. B. (2018). Mannose-capped lipoarabinomannan in Mycobacterium tuberculosis pathogenesis. In *Pathogens and Disease* (Vol. 76, Issue 4). https://doi.org/10.1093/femspd/fty026

Tyler, A. D., Mataseje, L., Urfano, C. J., Schmidt, L., Antonation, K. S., Mulvey, M. R., & Corbett, C. R. (2018). Evaluation of Oxford Nanopore's MinION Sequencing Device for Microbial Whole Genome Sequencing Applications. *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-018-29334-5

Uttamchandani, R. B., Daikos, G. L., Kramer, M. R., Fischl, M. A., Dickinson, G. M., Yamaguchi, E., & Kramer, M. R. (1994). Nocardiosis in 30 patients with advanced human immunodeficiency virus infection: Clinical features and outcome. *Clinical Infectious Diseases*, *18*(3). https://doi.org/10.1093/clinids/18.3.348

Van Ingen, J., Aksamit, T., Andrejak, C., Böttger, E. C., Cambau, E., Daley, C. L., Griffith, D. E., Guglielmetti, L., Holland, S. M., Huitt, G. A., Koh, W. J., Lange, C., Leitman, P., Marras, T. K., Morimoto, K., Olivier, K. N., Santin, M., Stout, J. E., Thomson, R., … Wagner, D. (2018).

Treatment outcome definitions in nontuberculous mycobacterial pulmonary disease: An NTM-NET consensus statement. In *European Respiratory Journal* (Vol. 51, Issue 3). https://doi.org/10.1183/13993003.00170-2018

van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C. L. M., Söding, J., & Steinegger, M. (2023). Fast and accurate protein structure search with Foldseek. *Nature Biotechnology*. https://doi.org/10.1038/s41587-023-01773-0

Vollmer, W., Blanot, D., & De Pedro, M. A. (2008). Peptidoglycan structure and architecture. In *FEMS Microbiology Reviews* (Vol. 32, Issue 2, pp. 149–167). Oxford Academic. https://doi.org/10.1111/j.1574-6976.2007.00094.x

Votintseva, A. A., Pankhurst, L. J., Anson, L. W., Morgan, M. R., Gascoyne-Binzi, D., Walker, T. M., Quan, T. P., Wyllie, D. H., Del Ojo Elias, C., Wilcox, M., Walker, A. S., Peto, T. E. A., & Crook, D. W. (2015). Mycobacterial DNA extraction for whole-genome sequencing from early positive liquid (MGIT) cultures. *Journal of Clinical Microbiology*, *53*(4), 1137–1143. https://doi.org/10.1128/JCM.03073-14/SUPPL_FILE/ZJM999094142SO1.PDF

Wang, Y., Zhao, Y., Bollas, A., Wang, Y., & Au, K. F. (2021). Nanopore sequencing technology, bioinformatics and applications. In *Nature Biotechnology*. https://doi.org/10.1038/s41587-021-01108-x

Ward, K. R., & Matejtschuk, P. (2021). The principles of freeze-drying and application of analytical technologies. In *Methods in Molecular Biology* (Vol. 2180). https://doi.org/10.1007/978-1-0716-0783-1_3

Wardman, J. F., Bains, R. K., Rahfeld, P., & Withers, S. G. (2022). Carbohydrate-active enzymes (CAZymes) in the gut microbiome. In *Nature Reviews Microbiology* (Vol. 20, Issue 9). https://doi.org/10.1038/s41579-022-00712-1

Warren, R., De Kock, M., Engelke, E., Myburgh, R., Van Pittius, N. G., Victor, T., & Van Helden, P. (2006). Safe Mycobacterium tuberculosis DNA extraction method that does not compromise integrity. *Journal of Clinical Microbiology*, *44*(1). https://doi.org/10.1128/JCM.44.1.254-256.2006

Wattam, A. R., Davis, J. J., Assaf, R., Boisvert, S., Brettin, T., Bun, C., Conrad, N., Dietrich, E. M., Disz, T., Gabbard, J. L., Gerdes, S., Henry, C. S., Kenyon, R. W., Machi, D., Mao, C., Nordberg, E. K., Olsen, G. J., Murphy-Olson, D. E., Olson, R., … Stevens, R. L. (2017). Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center. *Nucleic Acids Research*, *45*(D1). https://doi.org/10.1093/nar/gkw1017

Weiss, C. H., & Glassroth, J. (2012). Pulmonary disease caused by nontuberculous mycobacteria. *Expert Review of Respiratory Medicine*, *6*(6), 597–613. https://doi.org/10.1586/ers.12.58

WHO. (2023a). *Global tuberculosis report 2023*. https://iris.who.int/bitstream/handle/10665/373828/9789240083851-eng.pdf?sequence=1

WHO. (2023b). *Leprosy*. https://www.who.int/news-room/fact-sheets/detail/leprosy

Wilkins, M. R., Gasteiger, E., Bairoch, A., Sanchez, J. C., Williams, K. L., Appel, R. D., & Hochstrasser, D. F. (1999). Protein identification and analysis tools in the ExPASy server. In

Methods in molecular biology (Clifton, N.J.) (Vol. 112). https://doi.org/10.1385/1-59259-584-7:531

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, *20*(1). https://doi.org/10.1186/s13059-019-1891-0

Xing, F., Lo, S. K. F., Ma, Y., Ip, J. D., Chan, W. M., Zhou, M., Gong, M., Lau, S. K. P., & Woo, P. C. Y. (2022). Rapid Diagnosis of Mycobacterium marinum Infection by Next-Generation Sequencing: A Case Report. *Frontiers in Medicine*, *9*. https://doi.org/10.3389/fmed.2022.824122

Zafar, H., & Saier, M. H. (2021). Gut Bacteroides species in health and disease. In *Gut Microbes* (Vol. 13, Issue 1). https://doi.org/10.1080/19490976.2020.1848158

Zallot, R., Oberg, N., & Gerlt, J. A. (2019). The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways. *Biochemistry*, *58*(41). https://doi.org/10.1021/acs.biochem.9b00735

Appendix A: Protocols

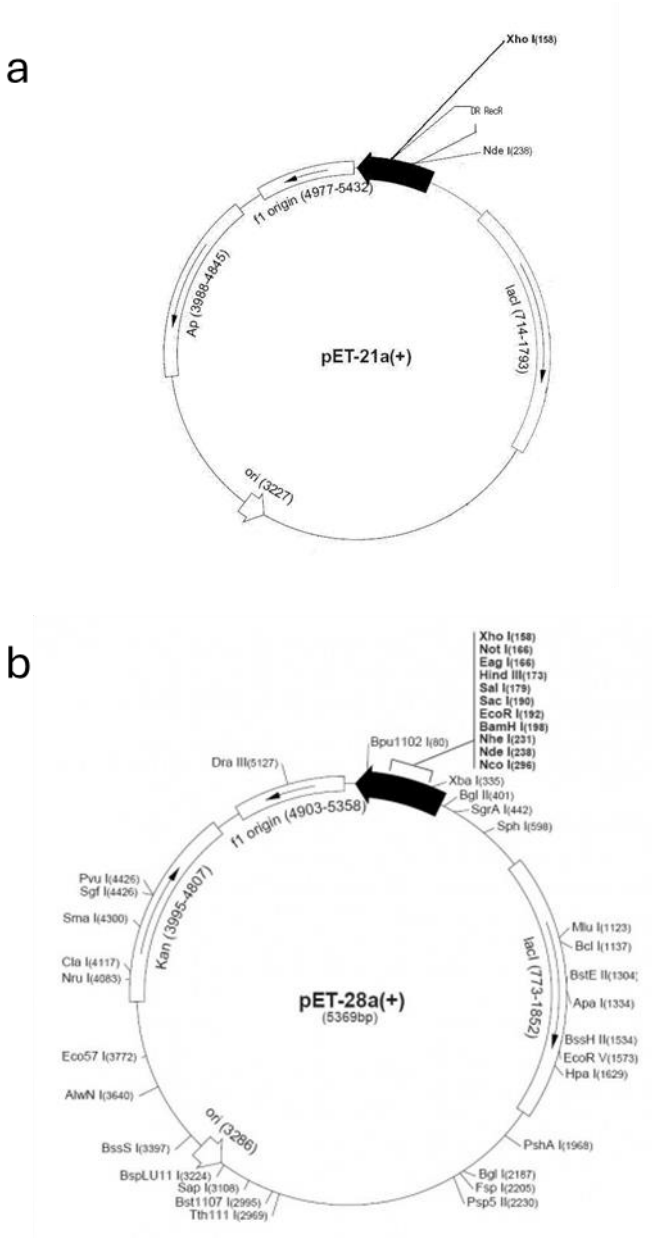


Figure 1: schematic representation of plasmids used. A. pET-21a and **b.** pET-28a

Competent cell growth protocol.

1. Set up a 5ml LB overnight of the cells.
2. Add 1ml of this to 100ml of LB in an unbaffled 1L flask.
3. Grow at 37oC/180rpm till OD600 get to around 0.4. This usually take about 2hr.
4. Transfer culture to 4 x 25ml sterilins.
5. Spin down cells at 5/6K for 5mins. This speed in our lab centrifuges equates to around 300g. It's possible that the centrifuge you are using may make a g value much higher than this and this could be the reason why your cells aren't competent because they've all burst!!
6. Pour off the LB supernatant and resuspend each of the four pellets in 3ml ice cold 0.1M MgCl2. Spin as above and pour of each of the supernatants.
7. Resuspend each pellet in 1ml ice cold 0.1M CaCl2. Pool each aliquot so you end up with 4ml of hopefully competent cells.
8. Leave on ice for at least 2 hours.
9. Store at -80 defrosting when needed.

Table 1 Typical PCR reaction set up. PCR components for a 50 µl PCR reaction.

Components and concentrations	Volume (µl)
Q5 DNA Polymerase	0.5
5x Reaction buffer	10
dNTPs	5
GC enhancer	5
Forward Primers (5 µM)	2.5
Reverse primers (5 µM)	2.5
Water	22.5
DNA	2

Table 2 qPCR reaction set up.

Regent	Volume (µl)
Luna® Universal qPCR Master Mix	5
Primer mix (500 nM; see Table 2.9)	2
DNA	2
Water	1

Table 3 SDM primer for DG02740

Mutation	Primer
H147A	CTT GTA TCA GGA GCT GTT ATC TAT GAG CGC
	GCG CTC ATA GAT AAC AGC TCC TGA TAC AAG
W270A	ATA TGG AAT CAG GCA ATG TTT AGT TCT TTG CGT
	ACG CAA AGA ACT AAA CAT TGC CTG ATT CCA TAT
W352A	TAC ACC ATG GAT GCC CGG AAA TCT
	AGA TTT CCG GGC ATC CAT GGT GTA TGG ATT TCT
E467A	CCT GTC GGA CGT TGG GCC GAT TTT GAG ACT TAC
	GTA AGT CTC AAA ATC GGC CCA ACG TCC GAC AGG
D468A	GTC GGA CGT TGG GAG GCC TTT GAG ACT TAC TGG -
	CCA GTA AGT CTC AAA GGC CTC CCA ACG TCC GAC –
E470A	CGT TGG GAG GAT TTT GCC ACT TAC TGG TCA TGT
	ACA TGA CCA GTA AGT GGC AAA ATC CTC CCA ACG
C475A	AG ACT TAC TGG TCA GCT TGC CGT CTG GGA ACA
	TGT TCC CAG ACG GCA AGC TGA CCA GTA AGT CTC
E530A	CAG CGT ACA CTA GAT GCC ATG CTT ATG ACA CAA
	TTG TGT CAT AAG CAT GGC ATC TAG TGT ACG CTG
D557A	GGT GTA ATG AAT GCT GCA GGG GAA TGG AAT GAT
	ATC ATT CCA TTC CCC TGC AGC ATT CAT TAC ACC
E559A	ATG AAT GCT GAC GGG GCC TGG AAT GAT TCT CGT
	ACG AGA ATC ATT CCA GGC CCC GTC AGC ATT CAT
D562A	GAC GGG GAA TGG AAT GCC TCT CGT CAA AGC CTT
	AAG GCT TTG ACG AGA GGC ATT CCA TTC CCC GTC
E628A	TAT GGC TTT ACG ATG GCC AAT TAT GGG CAT GAC

	GTC ATG CCC ATA ATT GGC CAT CGT AAA GCC ATA
H632A	TTT ACG ATG GAG AAT TAT GGG GCT GAC GGC
	GCC GTC AGC CCC ATA ATT CTC CAT CGT AAA
D633A	GAG AAT TAT GGG CAT GCG GGC CGG ACA AGT AAG
	CTT ACT TGT CCG GCC CGC ATG CCC ATA ATT CTC

Table 4 Library preparation for ONT Flow cell.

Reagent	Volume (µl)
Sequencing buffer (SQB)	34
Loading Beads (LB)	25.5
Nuclease-free water	4.5
DNA library	11
Total	75

Priming and loading the MinION and GridION Flow Cell

1. Open the MinION or GridION device lid and slide the flow cell under the clip. Press down firmly on the flow cell to ensure correct thermal and electrical contact.
2. Slide the flow cell priming port cover clockwise to open the priming port.
3. After opening the priming port, check for a small air bubble under the cover. Draw back a small volume to remove any bubbles:

    a. 1. Set a P1000 pipette to 200 µl
    b. 2. Insert the tip into the priming port
    c. 3. Turn the wheel until the dial shows 220-230 µl, to draw back 20-30 µl, or until you can see a small volume of buffer entering the pipette tip.
4. Complete the flow cell priming:

    a. 1. Gently lift the SpotON sample port cover to make the SpotON sample port accessible.

b. 2. Load 200 µl of the priming mix into the flow cell priming port (not the SpotON sample port),
c. avoiding the introduction of air bubbles.

5. Mix the prepared library gently by pipetting up and down just prior to loading.

    a. Add 75 µl of the prepared library to the flow cell via the SpotON sample port in a dropwise fashion. Ensure each drop flows into the port before adding the next.
    b. Gently replace the SpotON sample port cover, making sure the bung enters the SpotON port and close the priming port

6. Place the light shield onto the flow cell, as follows:
    a. Carefully place the leading edge of the light shield against the clip. Note: Do not force the light shield underneath the clip.
    b. Gently lower the light shield onto the flow cell. The light shield should sit around the SpotON cover, covering the entire top section of the flow cell.

FLYE assembly protocol

flye --genome-size --out-dir --threads 8 --nano-raw input_reads.fastq

--Genome-size specifies the size of the genome of the organism of interest.

--out-dir specificies the out directory to store FLYE results.

--Threads specifies the number of CPU core for the process.

--nano-raw input_reads.fastq Specifies the input flies are Nanopore reads FASTQ format.

1. **error correction.**

The process in which FLYE works is initially it begins by error correcting, by employing consensus-based approaches using overlapping reads to improve accuracy.

2. **Building contigs**

This is followed by graph building based on the overlaps between reads using a de Bruijin graph. Overlaps are constructed into contigs which allows for continuous rewards of long sections.

3. **Repeat graph construction**

The repeat graphs deal with repetitive regions of the genome which can cause misassembles. Each edge of a contig is classified as unique or repetitive. Each unique edge appears once in the traversal. This allows optimal construction of the assembly.

4. **Scaffolding**

If the contigs do not over lap Scaffolding will be using to orientate and link the contigs into larger sequences.

Genome comparison

The reference genomes for each of the target sequences as well as the constructed assemblies were aligned with each other allowing for sequence similarities to be identified as well as misassembles.

Genome annotation

PATRIC uses the RAST (Rapid Annotation using Subsystem Technology) toolkit to assign functions to genes and identify metabolic pathways.