Elucidation of alkane metabolism in the filamentous fungi *Ascocoryne sarcoides*

A thesis submitted to the Faculty of Science, Agriculture and Engineering of Newcastle University for the partial fulfilment for the degree of Doctor of Philosophy.

Joshua Loh

Abstract

Ascocoryne sarcoides has been reported to produce a variety of secondary metabolites such as linear and cyclic alkanes that are suitable for biofuel applications. Alkanes and alkenes are important as they are fully compatible with current fuel infrastructure. The genetic and biochemical basis for the biosynthesis of linear alkanes in fungi is not known and routes for cyclic alkane biosynthesis in any domain remains to be established. In this thesis, *A. sarcoides* was able to grow robustly in chemically-defined media in which linear, but not cyclic, alkanes are the sole carbon source, providing evidence for fungal degradation of alkanes. To establish alkane metabolic pathways in *A. sarcoides,* the genome and metabolome of six publicly available *A. sarcoides* isolates were examined. The genomes of all six isolates were sequenced, assembled and annotated. For each isolate, over 10, 000 gene products were identified by combining expression data with Hidden Markov machine learning. Each genome and predicted proteome achieved over 90% complete annotation against BUSCO's database, considered the threshold for a high-quality dataset. No homology to known alkane producing genes were detected in any *A. sarcoides* isolates. By integrating annotations, pathway mapping and gene ontology with comparative analysis, hypothetical pathways for alkane degradation (via ALK-like P450), linear alkane biosynthesis (via *fdc1*-mediated fatty acid decarboxylation/decarbonylation) and cyclic alkane biosynthesis (via lipid lyase route) are proposed. These findings provide candidate genes for downstream heterologous expression and have the potential to increase the available toolkit for advanced biofuel applications. Solvent extraction and stir bar sorptive methods coupled to GC/MS were used to screen for biogenic hydrocarbon metabolites. The solvent extraction method did not identified the presence of biogenic alkanes. Moreover, results from SBSE were inconclusive in establishing *A. sarcoides* as an alkane producer due to exogenic alkane contamination and will require further method development.

Acknowledgement

I would like to express my thanks to the BBSRC and Newcastle University for the opportunity to carry out this PhD research project.

My sincerest words here lacks the gravity in gratitude for my PhD supervisor, Dr Thomas Howard. I am thankful for his brilliant insights, sage-like patience, dedicated mentorship, and unwavering support through the last four years. It has been my pleasure to work with him. Anymore mellifluous words will only be mistaken for worship, of which I am sure Tom will be entirely uncomfortable. I would like express my thanks to my secondary supervisor, Dr Jem Stach, for his contributions throughout the years, particularly for brilliant insights during many meetings. Also to Dr Jon Marles-Wright, I am grateful for the support through the years and also for the resources to learn the dark arts of coding. To my panel advisors, Dr Ethan Hack and Dr Kristen Wolff, I thank them for their input and advice.

I would like to thank Dr Matthew Peake for his technical support and for his friendship. I also would like to thank Dr Camilla Liscio, from Anatune, and Dr Rachael Dack, from Newcastle University's School of Natural Science, for their analytical expertise and their support.

In the research group, I would like to give thanks to Dr Alice Banks and Dr Colette Whitfield. They have been exemplary in the last few years and have been fantastic company. I will no doubt miss them dearly. To my friends and colleagues, Christopher Azubuike, Alex Laverick, and Alis Prusokas, for their advice, company, and friendship.

To my friends, Alidi Kusuma, Natalia Hurtado, Niall Conboy, Ali Leverett, Jasmine Bird, Bradley Brown and Paulina Focht, I thank them for their wonderful company, unwavering support, and, for making post-graduate life bearable.

Last but not least, I dedicate this thesis to the people most important to me; To my mother and Yanhua.

List of abbreviations

AAR	acyl-ACP reductase
ACP	Acyl Carrier Protein
ACP	Acyl carrier protein
ADO	Aldehyde Deformylating Oxygenase
ADO	aldehyde decarboxylase oxygenase
AR	acyl reductase
ARE1	alkane responsive element 1
ATEX	Automated Tube Exchange
ATP	Adenosine Triphosphate
BCKD	branched-chain α-keto acid dehydrogenase
BGC	Biosynthetic Gene Clusters
BLAST	Basic Local Alignment Search Tool
BLAST	Basic local alignment search tool
BLASTP	Protein-Protein Blast
BLASTP	BLAST-Protein
BLASTX	Translated-Protein BLAST
BUSCO	Benchmarking Universal Single-Copy Orthologs
$CDCl_3$	Deuterated-Chloroform
CER1	ECERIFERUM1
CER3	ECERIFERUM3
CFU	Colony Forming Units
CPR	NADPH-cytochrome P450 reductase
CTP	Tricarboxylate Transport Protein
DMA	Defined Media Agar
DMAPP	Dimethylallyl Pyrophosphate
DMAPP	dimethylallyl pyrophosphate
DMAT	DNA Methyltransferase
DNA	Deoxyribonucleic Acid
DNA	Deoxyribonucleic Acid
e-Value	Expected Value
ER	Enoyl reductase
EU	European Union

EVs	Electric Vehicles
FA	Fatty Acids
FAD	Flavin Adenine Dinucleotide
FAE	Fatty Acid Elongation
FAME	Fatty Acid Methyl Ester
FAP	Fatty Acid Photodecarboxylase
FAR	Fatty Acid Reductase
FAS	Fatty acid synthase
FAS1	Fatty acid synthase 1 subunit alpha
FAS2	Fatty acid synthase 1 subunit beta
FFV	Flex Fuel Vehicles
FPP	farnesyl pyrophosphate
FPPS	farnesyl pyrophosphate synthase
GC-MS	Gas Chromatography-Mass Spectroscopy
GHG	Green House Gas
GMC	Glucose-Methanol-Choline
GO	Gene Ontology
GPP	geranyl pyrophosphate
GPPS	geranyl pyrophosphate synthase
HD	Hydroxyacyl dehydratase
HDAC	Histone Deacetylase
HGVs	Heavy Goods Vehicles
HMM	Hidden Markov Model
HPODE	hydroperoxyoctadecadienoic acid
HS-SPME	Headspace Solid Phase Micro-Extraction
IPP	Isopentenyl Pyrophosphate
IPP	isopentenyl pyrophosphate
ITS	Internal Transcribed Spacer
ITS	Internal Transcribed Spacer
KAAS	KEGG’s automated annotation server
KCS	β-ketoacyl-CoA synthase
KEGG	Kyoto Encyclopedia of Genes and Genomes
KR	Ketoacyl reductase

KS	Ketoacyl synthase
LOX	Lipoxygenase
MAFFT	Multiple Alignment Using Fast Fourier Transform
MBD	methyl-CpG-binding domain
MVA	mevalonate pathway
MtCO2e	Metric tons of CO$_2$ equivalent
MVA	Mevalonate Pathway
NADP	Nicotinamide adenine dinucleotide phosphate
NADPH	Nicotinamide Adenine Dinucleotide Phosphate
NGS	Next Generation Sequencing
NIST	National Institute of Standards and Technology
OAPEC	Organization of Arab Petroleum Exporting Countries
OD	Optical Density
OLC	Overlay-Layout-Consensus
OMA	Oat Meal Agar
ORFs	Open Reading Frames
PDA	Potato Dextrose Agar
PDB	Protein Data Bank
PDB	Potato Dextrose Broth
PDMS	Polydimethylsiloxane
PPO	psi-producing oxygenase
PSI-BLAST	Position-Specific Iterative BLAST
PT	Prenyltransferases
PTR-MS	Proton-Transfer Reaction Mass Spectroscopy
PTR-MS	Proton-Transfer Reaction Mass Spectroscopy
PUFA	polyunsaturated fatty acids
RBO	Reverse Beta-Oxidation
RT	Retention Time
SBSE	Stir Bar Sorptive Extraction
SNAP	Synonymous Non-Synonymous Analysis Program
SPME	Solid-Phase Microextraction
UK	United Kingdom
UPPS	undecaprenyl pyrophosphate synthase

USA	United State of America
VLC	Very Long Chain
VLCA	Very Long Chain Alkanes
VLCFA	Very Long Chain Fatty Acids
VOC	Volatile Organic Compounds

CHAPTER 1 GENERAL INTRODUCTION

1.1 Biofuels

The first use of biofuels for transportation dates back to early the 19th century. Internal combustion engine inventors, such as Samuel Morrey, Nikolaus August Otto and Rudolf Diesel, designed petrol and diesel engines that were powered by fuels such as alcohol and vegetable oils. In 1826, Morrey experimented with different fuel blends to power one of his engines, one such fuel mix was based on ethanol and turpentine. Otto, who advanced on Morrey's work, designed an engine to run on a high ethanol fuel blends (Ghobadian et al., 2004). During a demonstration in 1912, Diesel demonstrated the capabilities of one of his engine, which was powered exclusively on peanut oil (Harford, 2016). Historically, biofuel consumption increases during times of conflict or shortages. These early 20th century engines were able to operate on a mix of biofuel and fossil fuel during periods of fuel rationing. This reduces the cost of fuels and alleviates fuel supply stress in the market.

Biofuels remain appealing because of four major reasons. Firstly, sustainable biofuels are seen as part of the solution to address global issues such as anthropogenic green house gas (GHG) emissions. The United Nations Framework Convention on Climate Change, Kyoto Protocol and Paris Agreement are international treaties that commit and encourage countries to limit global temperature to 2 ℃ relative to preindustrial temperature (UNFCC, 1994, Kyoto Protocol, 1997, Paris Agreement, 2016). This is achieved by reducing anthropogenic (GHG) emissions. In the United Kingdom (UK), the transport sector is the largest GHG emitting sector (Figure 1.1A), at 126 MtCO2e (Metric tons of CO_2 equivalent) and it accounted for 28% of UK greenhouse gas (GHG) emissions in 2017 (Reducing UK emissions, 2018). The transport sector is the only sector that has seen an increase of 5% in GHG emissions in the period of 2012 to 2017 (Figure 1.1B) and is the sector with the slowest decrease (2%) for emissions in the period of 1990 - 2017 (UK Greenhouse Gas Emissions, 2017). This is in contrast to overall UK GHG emissions trends, where emissions have decreased by 43% in total (UK Greenhouse Gas Emissions, 2017). Road transportation like cars, heavy goods vehicles (HGVs) and vans account for 87% of transportation sector's emissions and has defied GHG emission reduction targets for the last few years (Reducing UK emissions, 2018). For the UK to meet its GHG targets, it is important to address the emission from road transport. Currently, the European Union (EU) has set out directives for GHG savings described in the EU's Renewable Energy Directive (2009/28/EC, 2009). This is a 60% total reduction in emissions for biofuels and bioliquids compared to fossil fuel counterparts (Edwards et al., 2017). This target is based on the process of producing and consuming biofuels against that of fossil fuels usage, extraction processes, and environmental impact (Directive (EU) 2018/2001, 2018). Production of biofuels in the EU will have to meet this sustainability target to meet the directive in order to reduce GHG emissions in the transport sector.

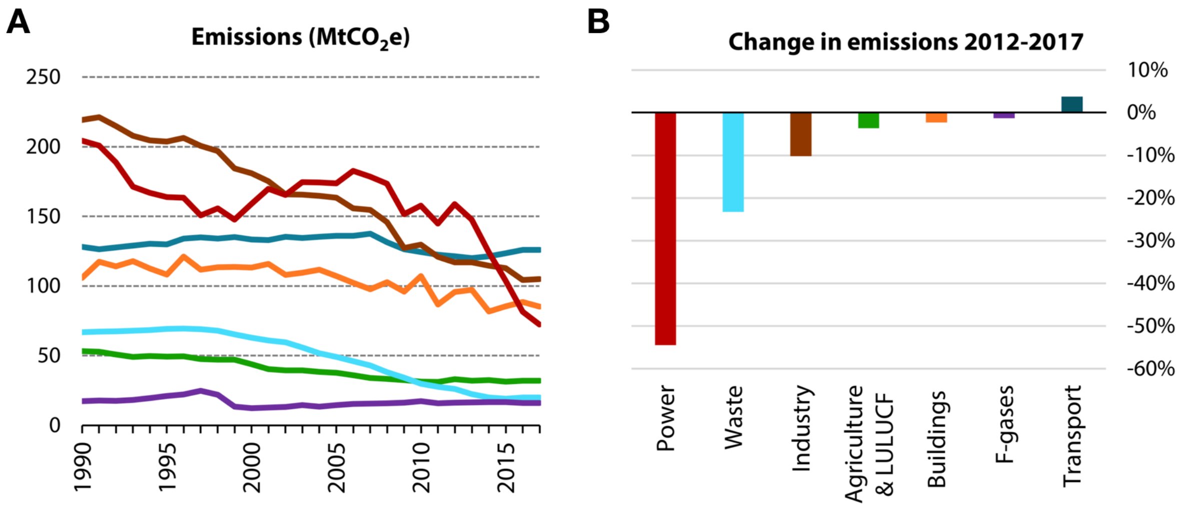


Figure 1.1. Emissions of CO_2 in metric tonne across eight major sectors in the UK. A) CO_2 emissions between 1990 - 2017 B) CO_2 emissions between 2012 - 2017. The UK transport sector is the only sector that has seen an increase in emission output. BEIS (2018) 2017 UK Greenhouse Gas Emissions, Provisional Figures; BEIS (2018) 2016 UK Greenhouse Gas Emissions, Final Figures. Figures are from "Reducing UK emissions – 2018 Progress Report to Parliament".

Secondly, unlike finite fossil fuel deposits, when appropriately sourced, biofuels can be a sustainable source of low GHG emission transport fuel. From a sustainability aspect, biofuels can be split into four categories. First-generation biofuels are dependent on crops such as sugar cane and corn for starch or vegetable oil. Crops produce sugar that is used for bioethanol fermentation and vegetable oil undergoes transesterification for the production of fatty acid methyl ester (FAME) fuel. Arable land and agricultural resources are used to grow fuel crops for the production of biofuel and this has its own complications regarding food securities, resource management and environmental impact. Second-generation biofuels utilise a wide range of biomass as a source of organic carbon and can include agricultural byproducts or non-food crops from non-arable lands. *Jatropha sp.*, switchgrasses, *Miscanthus* sp., and poplar are currently being considered or used for biofuel production (Samson et al., 2005, Heaton et al., 2008, Nahar et al., 2011, Dou et al., 2017). This has the distinct advantage of avoiding production competition with food crops. The main challenges of second-generation biofuel is the extraction of organic carbons from lignocellulosic biomass for the production of biofuel. Third-generation biofuels are algal-derived. Algal aquaculture does not directly compete for arable land usage and is a good source of renewable organic carbon as some species can accumulate up to 40% lipids to body its mass (Becker, 1994). The aquaculture can be extracted for its oil and subjected to a transesterification process to produce FAME biofuels. However, there are significant challenges associated with bioreactor scalability and cost-efficient extraction. Fourth-generation biofuels do not need biomass and use inorganic feedstocks such as water or sequestered CO_2 for the production of fuels and require a renewable energy source to drive the reaction. Hydrogen fuels can be produced by electrolysis and this reaction can be powered renewably by conventional methods or with a microbial fuel cell (Logan, 2009). Microbial electrosynthesis is also an alternative route for the production of biofuels. Organisms that are capable of electroautotrophic metabolism, such as *Cupriavidus necator* can be engineered to convert CO_2 to a hydrocarbon metabolite, the terpenes and alkanes under the presence of an electric current (Crépin et al., 2016, Krieg et al., 2018). An attractive prospect is to use algal or cyanobacterial “cell-factories” as these organisms are able to sequester CO_2, to undergo photosynthesis and be engineered with synthetic pathways to produce drop-in hydrocarbon fuels (Wijffels et al., 2013).

Thirdly, for any fuel to be viable for transportation usage energy density is an important consideration. Energy density in this context is defined as the amount of specific energy released by a fuel in a given mass. The energy density of current biofuels (e.g. ethanol, 26.8 MJ/kg (Thomas, 2000)) is comparable, albeit lower, to that of fossil fuels (e.g. petroleum, 44.4 MJ/kg (Thomas, 2000)) and higher than conventional lithium-ion polymer battery (0.2 - 0.7 MJ/kg (Gür, 2018)). Due to the energy density limitation of batteries, electric vehicles (EV) have limited range before necessitating a recharge. Furthermore, increasing the range of EVs by adding more batteries would be impractical, as diminishing returns due to increasing battery mass and a constant power output would be an impeding factor. This affects the carrying capacity of EVs and restricts the potential of EVs to transporting passengers in urban environments (Hall et al., 2017). For batteries, further consideration must be made for recharging spent energy. Refuelling, for most conventional vehicles, is limited by the capacity of the tank and the speed of the pump. In contrast, recharging for a high specification lithium-ion polymer EV battery can be up to 75 minutes with a specialised high-power direct current charge station for a new EV battery (Tesla, 2019). There are also practical concerns, as these batteries and charger are not yet commonly available infrastructure. The combination of range, carrying capacity, and recharge/refuel rate are important factors for operating vehicles such as heavy goods vehicles, freighter ships and aircraft.

Fourthly, there are legitimate political and economic impacts in producing and consuming biofuels, as it can reduce a country's dependencies from fossil fuel market and producers. In 1973, the Organization of Arab Petroleum Exporting Countries (OAPEC) mandated a decrease in oil production and embargo on exports to countries who offered political support to Israel for the Yom Kippur War (Maugeri, 2006). Through this action, the OAPEC was able to use oil production and export to gain political influence. This had an effect on global fuel supplies. In the United States of America (USA), fuel prices increased from US$3 to nearly US$12 per barrel (Potter, 2008), and created an oil crisis. Prior to the 1973 oil crisis, Brazil was a net importer of oil and ethanol fuels was only produce to supplement Brazil's fuel market when sugar prices are low. Although many countries in that era, including Brazil, had a politically distant stance from Arab-Israel events, the price of oil had impacted indiscriminately on countries around the world. This motivated the government of Brazil to reduce its reliance on imported oil by launching and supporting its National Alcohol Program (Soccol et al., 2005). The National Alcohol Program was designed to incentivise bioethanol production for the use of transportation fuel in Brazil by using sugarcane as a feedstock. To increase consumption of bioethanol fuels, investments and subsidies were also used to develop and increase adoption for vehicles that are capable of utilising high percentage hydrous ethanol as fuels. At its peak in 1986, these high percentage ethanol vehicles accounted for 72.6% of all light-vehicles in Brazil (ANFAVEA, 2012). By the 1980s, Brazil was able to fulfilled 20% of its fuel needs with ethanol fuels (Potter, 2008). The Brazilian fuel market today is able to take advantage of pricing differences in bioethanol or hydrocarbon fuels with the emergence of flex fuel vehicles. These vehicles are able to operate on different blends of bioethanol and hydrocarbon fuel (ANFAVEA, 2012). Brazil, with its sustainable bioethanol program, is able to exercise control over its energy policies and reduce dependency on fossil fuel imports, resulting in energy independence for the Brazilian transport sector.

The two most common biofuels are bioethanol and FAMES. These act as alternative fuels for petroleum and diesel respectively. They are widely consumed by the market and are compatible with conventional vehicles and infrastructure at low levels of blending with hydrocarbon fuels. Current infrastructure and non-FFV (Flex fuel vehicles) cannot tolerate high percentage mixes. This typically restricts blending limit to 5 to 15%. For high ethanol blends, the fuel blend contains a large proportion of oxygenated polar compounds with a relatively short alkyl chain, which makes the fuel mix hygroscopic. Without modifications to vehicles or the infrastructure, hydroscopic fuels can draw ambient moisture, leading to corrosion and eventually damage. This is more prevalent in high percentage ethanol fuels. There are also issues with long term stability during storage of FAMEs. All FAMEs are prone to oxidation over time. This oxidation leads to changes in pH, peroxide values and viscosity which can generate acidic species and leads to a loss of fuel quality and performance (Zuleta et al., 2012). Furthermore, acidic species within the fuel mix can damage incompatible material (e.g. metallic and polymeric components), leading to leaching, corrosion and degradation of components in conventional vehicles (Zuleta et al., 2012). To mitigate the effects of acidic degradation, specialist storage components are required to reduce oxidation and its effects, ultimately increasing the cost of biodiesel.

Countries with established ethanol and FAME infrastructures and FFVs that are able to accommodate high percentage biofuels are able to mitigate associated disadvantages. In the case of vehicles, specialised components are required for FFVs to deal with leaching of plastics into the fuel and corrosion of unprotected components. Most of these components cannot be retrofitted to existing non-FFV. Moreover, infrastructure changes can be impractical as they are both expensive and time consuming. This combination of obstacles means that the deployment of biofuels is limited to a relatively low blend, typically around 5% to 10%, which is also known as the blend wall. While it is possible to establish a market that has a high consumption of biofuels along with high biofuel blends, infrastructure and vehicular implementation requires decades of political impetus and costly subsidisation (Potter, 2008). This makes it highly challenging for countries to adopt biofuels as the fuel of choice. For example, regardless of the high number of FFVs in the USA, the lack of infrastructure proved to be a barrier for the distribution of E85 fuels, reducing availability and, by extension, consumption of E85 fuel (Bredehoeft et al., 2014). The lack of infrastructure (such as refilling stations and fuel depot), limits consumer choice to conventional fuels. In the USA, despite initiating a similar ethanol program to Brazil in the 1970s, complemented by years of supportive policy and investments, the ethanol biofuel market has not reached its intended objective and has failed to diversify USA from petroleum-based fuels (Potter, 2008). For example in 2010, the USA bioethanol market only achieved 10% displacement of gasoline market (US Dept of Agriculture, 2010). In contrast and in 2008, Brazil is able to achieved 50% displacement of gasoline with ethanol (Agência Brasil, 2008). This highlights the difficulty of establishing a biofuel market within a country.

To maximise biofuel adoption and consumption, biofuels must be compatible with existing infrastructures and vehicles while minimising cost and disruption. Ideally, such biofuels have identical chemistry and physical properties to their fossil fuel counterparts. There is no blend wall to hydrocarbon-based biofuels (also known as drop-in fuels) and these biofuels can fully displace conventional fossil fuels (Table 1.1). Moreover, an advantage with hydrocarbon biofuel is the ability to blend different fuel molecules to achieve the desired specification. Importantly, specifications are heavily dependent on factors such as temperature, weather, and humidity, as these can affect the property of a given fuel. The properties of the fuel can be altered by changing the fuel composition with a range of hydrocarbon compounds and fuel additive compounds. To fully replicate existing fossil fuel composition, there is a clear need to discover unique pathways. To achieve this, the genetics and biochemistry of linear, branched and cyclic hydrocarbon compounds of varying lengths must be elucidated to replicate the full range of compounds found in petroleum, diesel and aviation fuel.

Table 1.1. The composition and properties of various fuel type. (Table adapted from Peralta-Yahya et al., 2012).

Fuel Type	Major components	Properties
Gasoline	C4 to C12 hydrocarbons. Linear, branched, cyclic, and aromatics	Octane rating: 87 to 91 Favoured for energy content
Diesel	C9 to C23 hydrocarbons. Linear, branched, cyclic, and aromatics	Cetane Rating: 40 to 60
Jet fuel	C8 to C16 hydrocarbons. Linear, branched, cyclic, and aromatics	Heat density Low gelling temperature

1.2 Biosynthesis of hydrocarbons

Hydrocarbons are simple compounds that consist only of carbon and hydrogen. Their biosynthesis is widespread and has been observed in different biological domains. These compounds have diverse functions in biology. The hydrophobicity of hydrocarbons can serve as a protective layer to minimise water loss or to act as a barrier in plants and insects (Bernard et al., 2012a, Qiu et al., 2012). Due to the volatility of smaller hydrocarbons, they are able to function as semiochemicals including pheromones and repellants (Howard et al., 1982, Reed et al., 1994, Qiu et al., 2012). For cyanobacteria, hydrocarbons function as an integral component to lipid membranes leading to changes in membrane properties (Berla et al., 2015, Lea-Smith et al., 2016).

Biosynthesis of hydrocarbons in biological systems are dependent on three distinct pathways. Alkanes and terminal olefins can be generated by fatty acid (FAs)

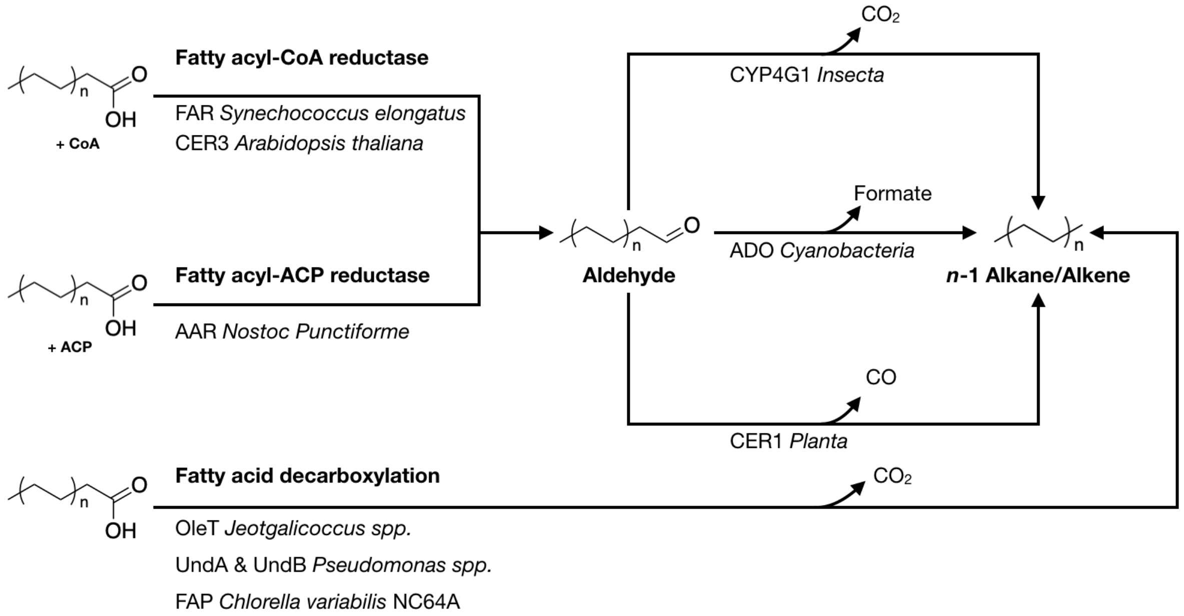


Figure 1.2. Biosynthetic pathways for alkane biosynthesis found in natural systems. FAR, CER3, and AAR are naturally occurring fatty acid reductases that are able to covert fatty acids to aldehyde. ADO, CER1, and CYP4G1 are able to decarboxylate aldehydes to form an n-1 alkane product. There are also 1-step pathway, like OleT, UndA, and FAP, that are able to catalyse fatty acids to alkenes/alkanes.

decarbonylation or decarboxylation, which generates hydrocarbon product by a $C_{(n-1)}$ cleavage of fatty acids (Figure 1.2). With polyolefenic hydrocarbons, biosynthesis is achieved by head to head condensation of olefins. Lastly, the isoprenoid pathway is capable of hydrocarbon biosynthesis by condensing repeating units of isopentenyl pyrophosphate (IPP) and dimethylallyl pyrophosphate (DMAPP) to form long chain isoprenes that can be further condense to form hydrocarbon terpenes. These will be covered in detail in the following sections.

1.2.1 Two step fatty acid derived hydrocarbons

A two-step enzymatic reaction facilitates the conversion of FAs to alkane. A feature of this pathway is the formation of a fatty aldehyde intermediate. The aldehyde intermediate is subjected to C-C scission, forming a $C_{(n-1)}$ alkane product and a C_1 byproduct. Two major enzymes play a key role in this pathway. The first enzyme is a fatty acid reductase and it is able to reduce fatty acids to an equivalent length fatty aldehyde. Secondly, an aldehyde decarbonylase cleaves the terminal aldehyde group to form an alkane and a C_1 byproduct. This is the typical route for alkane/alkene production in terrestrial plants, algae, insects and certain prokaryotes.

Linear alkane biosynthesis in planta

The cuticular wax of *Arabidopsis thaliana* is composed of alkanes, alcohols, aldehydes and ketones with a length around C20 to C36 (Bourdenx et al., 2011). Alkanes are also the most abundant hydrophobic compound in the cuticular wax layer (Bourdenx et al., 2011). The first indication of a genetic basis for alkane in *A. thaliana* came from eceriferum mutant lines, mutants that are incapable of producing wax (Jenks et al., 2002). Two mutant lines have been crucial in the genetics that underpin *cer* phenotype. The *cer*1 mutant was characterised as having decreased alkane, secondary alcohol and ketone levels with a slight increase in aldehydes in the cuticular wax. In contrast, *cer*3 mutant showed a dramatic reduction in aldehydes, alkanes, secondary alcohols and ketones (Chen et al., 2003, Kurata et al., 2003). This indicated the formation of aldehyde may be an intermediate product to the formation of waxy metabolites.

Radio-labelled feeding assays indicate that fatty acid reduction is required in the formation of an equivalent length aldehyde in *Pisum sativum* (Cheesbrough et al., 1984). This highlights the necessity of the fatty elongation pathways for generating

VLC (very long chain) fatty acids and, by extension, VLC aldehydes. Further evidence indicates that higher plants have the biochemical potential for multiple fatty acid elongation systems; 21 putative β-ketoacyl-CoA synthase (KCS) have been annotated in the *Arabidopsis* genome (Joubès et al., 2008). KCS initiates the first step of a four step fatty acid elongation reaction. The presence of multiple genes that encode KCS indicates distinct chain length preferences in the elongation step to biosynthesise VLC fatty acids. These VLC fatty acids are utilised for the production of very long chain alkanes (VLCA) (Bernard et al., 2012a). Once elongated, VLCFAs are catalysed to VLC alkanes by down stream acyl-CoA reduction and decarbonylation.

The first confirmation of aldehyde and alkane formation came from extracts of *Pisum sativum*, which was capable of catalysing the conversion of octadecanal to heptadecane (Cheesbrough et al., 1984). It was reported that a two step reaction is facilitated through the interaction of ECERIFERUM1 (CER1) with ECERIFERUM3 (CER3) and endoplasmic reticulum localised cytochrome b5 isoforms (CYTB5s) in *Arabidopsis* (Bourdenx et al., 2011, Bernard et al., 2012a). This was proven when the *cer1, cer3,* and *cytb5* were reconstituted in *S. cerevisiae* (Bernard et al., 2012a). Heterologous expression of each gene was able to define the function of each protein. CER3 was found to convert VLCFA to very-long chain aldehyde as an intermediate of the pathway. The aldehyde is then converted to a $C_{(n-1)}$ alkane product by CER1 (Bernard et al., 2012a). Moreover, confocal microscopy of fluorescent tagged CER1, CER3 and CYTB5 in *A. thaliana* seedlings indicated that CER1 and CER3 was found to function as a multi-enzyme complex with CYTB5 facilitating redox reaction to drive the reaction (Bernard et al., 2012a). This revealed a two-step catalytic conversion of fatty acyl-CoA to an alkane, with CER3 acting as a fatty acyl reductase and CER1 acting as an aldehyde decarbonylase. Site-directed mutagenesis of the histidine residues in CER1 abolished alkane biosynthesis, demonstrating that they are essential for alkane biosynthesis (Bernard et al., 2012a). This study points toward the importance of histidine-iron interaction for alkane biosynthesis. Such decarbonylation yields odd chain alkanes due to the presence of even numbered fatty acids. To date, no protein structure nor molecular mechanism have been elucidated from the CER1/CER3 complex.

Linear alkane biosynthesis in insecta

Insects also have the capabilities to produce VLCA. The cuticular wax layer found on *Drosophila melanogaster* contains complex blends of long-chain alkanes and alkenes (C21 - C37) that serve to waterproof the insect and as semiochemicals (Howard et al., 1982, Qiu et al., 2012). The first indication of fatty acidsconversion to cuticular alkane biosynthesis came from *Periplaneta spp* (American cockroaches). *Periplaneta* extracts convert radio-labelled fatty acids to alkanes, while long chain ketones and secondary alcohols, when added to the extract, were not incorporated into the final alkane product (Major et al., 1978). This was the first biochemical evidence that indicate a *n-1* fatty acids decarboxylation. Much like in plants, VLCA was proposed to use elongated fatty acids as a substrate. Evidence came from microsomal studies of the integument tissues of *Periplaneta spp*, which were observed to elongate stearoyl-CoA up to C26 fatty acid and linoleoyl-CoA up to C28 fatty acid (Vaz et al., 1988). This coincides with the observation of pentacosane, a C25 alkane, and C27 alkenes, respectively, in *Periplaneta spp* (Vaz et al., 1988). This, again, highlights the key role of fatty acid elongation (FAE) pathways in facilitating cuticular wax formation in insects.

In *D. melanogaster*, the P450 CYP4G subfamily was investigated as a potential cuticular alkane biosynthesis gene because it had at least one ortholog in all insect genomes (Qiu et al., 2012) and expression data indicated the ortholog had the greatest level of expression relative to all 85 P450 present in *D. melanogaster* (Daborn et al., 2002). Immunohistochemistry of CYP4G1 was able to indicate a colocalised NADPH-cytochrome P450 reductase (CPR) as the redox partner required for alkane biosynthesis (Qiu et al., 2012). RNAi suppression of *Cyp4g* and *cpr* gene product led to a dramatic decrease in cuticular alkane/alkene content in surviving mutant flies (Qiu et al., 2012). Moreover, coexpression of *Cyp4g1* and cytochrome P450 fatty acid reductase (FAR) in *S. cerevisiae* led to production of alkanes and assay of the transformant purified lysate was observed to convert trideuterated-octadecanal to trideuterated-heptadecane (Qiu et al., 2012). This was the first time an alkane biosynthesis enzyme was identified in *insecta*. Further biochemical assays in the presence of NADPH and O_2, show that CYP4G1 and cytochrome p450 reductase catalyse the conversion of fatty aldehydes to alkanes and CO_2 as a byproduct (Reed et al., 1994, Qiu et al., 2012). This identified CYP4G1 as an aldehyde decarbonylase oxygenase in *D. melanogaster*. The first hint of the molecular mechanism came from a biochemical assay with microsome of *Musca* *domestica,* which indicated a P450 was responsible for the conversion of fatty acids to alkane and CO_2 byproduct via an intermediate aldehyde route (Reed et al., 1994). It is thought that under oxygenating conditions, the enzyme is able to perform heterolytic cleavage of the O-O bond of the Fe-peroxy intermediate, which resulted in aldehyde conversion to alkane. Currently, there is no protein structure of CYP4G1, which has limited the exploration of the underpinning molecular mechanism. Unlike CER1/CER3, *insecta* linear alkane biosynthesis does not seem to have a specific fatty acyl reductase partner and evidence points toward many unspecific aldehyde generating pathways in *D. melanogaster* (Qiu et al., 2012).

Linear alkane biosynthesis in Cyanobacteria

Cyanobacteria have been observed to produce linear hydrocarbons ranging from C15 to C19 in chain length (Winters et al., 1969). Sampling of seawater has detected alkanes (such as pentadecane and heptadecane) at concentrations ranging from 2 to 130 pg/mL (Schwarzenbach et al., 1978, Gschwend et al., 1980). Initially, it was not clear whether the alkanes were of anthropogenic origin or of biological origin. The presence of obligate hydrocarbon-degrading bacteria, also referred to as hydrocarbonoclastic bacteria, in minimally polluted water suggested a sustainable, and perhaps a biological, source of hydrocarbons to maintain these hydrocarbonoclastic bacterial communities (Leahy et al., 1990, Yakimov et al., 2007, Lea-Smith et al., 2016). It is thought that the presence of cyanobacteria in these habitats plays a key role in contributing to the hydrocarbons observed in seawater and interacts with hydrocarbonoclastic bacteria to drive the hydrocarbon cycles in natural ecosystems. While these observations implicate alkane biosynthesis in cyanobacteria, the biosynthetic pathway for alkane has yet to be identified.

The breakthrough in elucidating the alkane biosynthesis pathway in cyanobacteria came when whole-genome sequencing became economically viable and maturation of bioinformatics tools to facilitate the analysis of the genomes (Schirmer et al., 2010). By combining phenotypic observations, it is possible to leverage these findings to investigate linear alkane biosynthesis in cyanobacteria by interrogating the genome of each cyanobacterium (Schirmer et al., 2010). Due to the limited coverage of the cyanobacteria phylum available at the time and the presence of a non-producing alkane strain, a comparative and subtractive genomics approach was chosen to elucidate the alkane biosynthetic pathway (Schirmer et al., 2010).

With this approach, two candidate genes stood out in terms of predicted function, gene loci adjacency, and conserved nature within all alkane producing cyanobacteria (Schirmer et al., 2010). The first gene, orf1594, was predicted to encode a short-chain dehydrogenase or reductase family protein. It was hypothesised that this gene may play a crucial role in the reduction of acyl carrier proteins or coenzyme A thioesters, potentially functioning as a fatty acyl reductase (Schirmer et al., 2010). The second gene, orf1593, was predicted to encode a ferritin-like or ribonucleotide reductase–like family protein, which was hypothesised to have similar radical chemistry that can enable decarbonylation reaction (Schirmer et al., 2010). Heterologous co-expression of orf1594 and orf1593 in *Escherichia coli*, was able to elicit the production of odd-chain alkanes and alkenes, as well as even-chain fatty aldehydes and fatty alcohols (Schirmer et al., 2010). Sole expression of orf1593 did not alter the hydrocarbon profile of *E. coli*. In contrasts, sole expression of orf1594 led to production of fatty alcohols, thought to be the result of the oxidation of fatty aldehyde by the *E. coli* host (Schirmer et al., 2010). In *S. elongatus*, the deletion of putative orf1594 and orf1593 abolished the presence of alkanes (Schirmer et al., 2010).

In vitro assays of the gene product of orf1594 demonstrated enzyme activity for the conversion of acyl-ACP and acyl-CoA to an equal length fatty aldehyde when incubated with NADP and Mg^{2+}. Further characterisation of the enzyme indicated a Michaelis-Menten constant for acyl-ACP (8 ± 2 μM) than acyl-CoA (130 ± 30 μM) (Schirmer et al., 2010). This indicates a preference for acyl-ACP and the enzyme was assigned as an acyl-ACP reductase (AAR). *In vitro* assays of *Nostoc punctiforme* ortholog orf1593 showed that it was able to decarbonylate octadecanal to heptadecane in the presence of ferredoxin, ferredoxin reductase, and NADPH (Schirmer et al., 2010). This demonstrated that alkane is formed via the decarbonylation of fatty aldehydes and, as a result, the enzyme was assigned as an aldehyde decarbonylase. Further studies with isotopic labelling assays identified molecular O_2 as an important co-substrate to drive the cleavage of aldehyde (Warui et al., 2011, Pandelia et al., 2013) with formate as a byproduct. With this evidence in mind, the enzyme was reassigned to be an aldehyde deformylating oxygenase (ADO) (Warui et al., 2011), though in literature it is commonly referred to as aldehyde decarbonylating oxygenase.

Homology analysis of orf1593 indicated that it was a good match for a publicly available crystal structure of another cyanobacterium, *Prochlorococcus marinus* (strain MIT9313; Joint Center for Structural Genomics; PDB ID: 20C5) (Schirmer et al., 2010). This initial comparison was able to give mechanistic insights to the ADO reaction and indicate that ADO belongs to ferritin-like or ribonucleotide reductase–like family of nonheme di-iron enzymes (Schirmer et al., 2010). Since then, the protein crystal structure for ADO with a bound aldehyde has been solved (Buer et al., 2014). The crystal structure revealed the active site to possess an α-helical structure with a di-iron centre housed in a ferritin-like four-helix bundle at the core of the protein (Khara et al., 2013, Buer et al., 2014). Access for the substrate to this active site is provided by a tunnel-like hydrophobic pocket (Khara et al., 2013, Buer et al., 2014). It was hypothesised that by introducing a larger residue, like tyrosine, to the tunnel-like hydrophobic pocket, this would hinder longer chain fatty aldehydes while maintaining access to shorter aldehydes and would lead to the generation of shorter chain alkanes (Bao et al., 2016). Ten different residues were identified, each generating a different alkane product for each mutation (C3 to C11) (Bao et al., 2016). Structural studies also revealed the di-iron in ADO is coordinated by the residue glutamine 144 and the conformational changes in the helix containing glutamine 144 further indicate different binding states of the di-iron center (Jia et al., 2015). With the residues tyrosine 39, arginine 62, glutamine 110, tyrosine 122, and aspartic acid 143 identified to be close to the di-iron centre of ADO from *Synechococcus elongates PCC7942* (Wang et al., 2017). Moreover, substitution of these polar residues to non-polar residues led to reduced activity or abolishment of ADO activity, indicating the importance of polar residues in coordinating aldehyde substrate to the di-iron active site.

The molecular mechanism of ADO was also partly determined by radical clock probing with 2-(2-tetradecylcyclopropyl)acetaldehyde, a cyclopropyl aldehyde equivalent of octadecanal with a β positioned cyclopropyl motif to the carbonyl group (Paul et al., 2013). Compounds with cyclopropyl motifs have the capability to generate cyclopropylcarbinyl radicals, which undergo ring-opening reactions. The location of radical formation and the lifetime of a radical intermediate can be used to infer the molecular mechanism of ADO. Enzymatic assay of *N. punctiforme* ADO showed it was able to deformylate 2-(2-tetradecylcyclopropyl)acetaldehyde to 1-octadecene (Paul et al., 2013). This observation indicates that the cyclopropyl aldehyde underwent C-C scission radical cleavage at the α-carbon position, which rearranges cyclopropylcarbinyl to form 1-octadecene (Paul et al., 2013).

It is still unclear how molecular oxygen interacts with the di-iron site of the enzyme and the aldehyde substrate. A mechanism has been proposed which accounts for the molecular oxygen and was derived from quantum mechanical/molecular mechanical model and is based on the crystal structure of ADO from *Prochlorococcus marinus* (Wang et al., 2016). The ADO active site reacts with molecular oxygen to form an iron-peroxo group. This peroxo group initiates a nucleophilic attack on the α-carbon of the aldo group on the aldehyde substate, this destabilises the oxygen on the carbonyl group and forms a peroxyhemiacetal species with one of the molecular oxygen and lysing the peroxo group. Protonation on the carbonyl oxygen leads to heterolytic cleavage of the molecular oxygen and generates an instability on the α-carbon, leading to the formation of an alkyl radical with the α-carbon remaining attached as an oxo group. Protonation of the alkyl radical produce an alkane product. Formate is released as a byproduct when histone 160 activation of Fe^{2+} occurs concomitantly with breaking Fe_2–μ-hydroxo coordination with the formyl group. Hydration and reduction of the remaining hydroxyl group released the second molecular oxygen, which regenerates the state of active site for the next catalytic turnover.

1.2.2 Fatty Acid Decarboxylation pathway

There are pathways that decarboxylate fatty acids to form an $C_{(n-1)}$ alkane/alkene products that only require a single enzymatic step, thus do not form an aldehyde intermediate product. FAP from *Chlorella spp.,* OleT from *Jeotgalicoccus spp.* and UndA/UndB from *Pseudomonas spp.* have been demonstrated to produce linear terminal alkenes (known as olefins) from FAs.

Studies have observed the presence of alkanes with the length of C13 to C17, with pentadecane as the most abundant alkane, in green microalgal cultures of *C. variabillis* and *C. reinhardtii* (Sorigué et al., 2017)*.* Further microbial physiological studies showed that the formation of alkane was only detected in cultures incubated under light conditions (Sorigué et al., 2017). This result alluded to a light-dependent hydrocarbon pathway in alkane-producing *Chlorella spp*. Bioinformatic analysis of *C. variabillis* genome did not reveal homology to previously mentioned alkane synthases, which points toward a novel hydrocarbon pathway (Sorigué et al., 2017). As there was no known homology, purified cell lysate assays of *C. variabillis* was conducted. It was found that the purified fraction contained a glucose-methanol-choline (GMC) oxidoreductase was able to convert fatty acids to alkanes and alkenes (Sorigué et al., 2017). Heterologous expression in *E. coli* of the gene encoding for GMC oxiodoreductase resulted in the detection of alkane in the transformed *E. coli* cell lysate (Sorigué et al., 2017). As previous physiological studies indicated that light affected alkane production, an assay of the purified enzyme under different light conditions was undertaken. It indicated that fatty acid is required as a substrate and conversion to alkanes was only limited to lysates assayed under light conditions. Conversely, no conversion was detected in the assay in the absence of light. Furthermore, characterisation under different light regimes revealed the enzyme to be active in blue light (400 to 520 nm) and light intensity was linked to rate of reaction (Sorigué et al., 2017). This indicates that blue light is required to provide energy for redox reaction; Thusly, the enzyme was assigned as fatty acid photodecarboxylase (FAP). CvFAP have a wide range of substrate specificity, from C12 to C20. This enzymatic reaction results in the decarboxylation of fatty acids to produce a $C_{(n-1)}$ alkane product and a C_1 byproduct (Sorigué et al., 2017, Huijbers et al., 2018). Moreover, FAP requires FAD as a redox cofactor, although it remains unclear the exact mechanism in which FAD is reduced and oxidised under the presence of blue light. This is the first example of an alkane forming photoenzyme and, unlike other alkane biosynthetic pathway, does not require additional reductant to drive the reaction forward.

Jeotgalicoccus is a Gram-negative bacterial genus that is capable of producing terminal olefins. Olefin production was observed in *Jeotgalicoccus* cultures when it was supplemented with fatty acids. Interestingly, the olefins detected were $C_{(n-1)}$ to that of the fatty acid (Rude et al., 2011). This was further supported by cell lysate assays, which also elicit similar $C_{(n-1)}$ olefins (Rude et al., 2011)). Observations of both experiments supported a $C_{(n-1)}$ mechanism. The gene sequence was identified when a purified protein fraction was identified as having hydrocarbon biosynthesis activity (Rude et al., 2011). The protein identified is a P450 peroxygenase and is consistent with enzymatic assays, which required H_2O_2 to drive the reaction (Rude et al., 2011). It was noted that H_2O_2 is capable of damaging the heme and the protein (Hsieh et al., 2016). Later studies also indicate that OleT is capable of utilising other redox partners rather than solely H_2O_2 (Liu et al., 2014). Experiments with isotopic-labelled substrates provided further insights to the mechanism of OleT activity, revealeing that C1 carboxylate of a fatty acid is cleaved to form the CO_2 byproduct with no intermediates detected (Grant et al., 2015). The molecular mechanism was partially elucidated by utilising labelled substrates and analogous substrates and indicated that, when the substrate is docked in the active site, the hydroxyl-Fe^{3+} cofactor is able to abstract hydrogen from the substrate at β-carbon position. The substrate undergoes radicalisation leading to α-β carbon scission to form a CO_2 byproduct and the acyl radical stabilises the substrate by forming an alkene (Hsieh et al., 2017). Once abstracted, the hydroxyl-Fe^{3+} is oxidised to hydroxyl-Fe^{4+}, NADH+ and H_2O_2 reacts with the cofactor to regenerate it (Hsieh et al., 2017). Intriguingly, evidence also suggests that OleT can hydroxylate aliphatic substrates that do not contain a carboxylic group (Hsieh et al., 2017).

UndA and UndB are two similar olefin-producing enzymes from the *Pseudomonas* genus. These enzymes produce olefins via the decarboxylation of fatty acids. *P. aeruginosa* cultures that were fed with labelled substrate led to the observation of labelled olefin products (Rui et al., 2014). The *undA* gene was discovered using heterologous expression of a library containing over a thousand gene candidates generated through comparative bioinformatics (Rui et al., 2014). Furthermore, deletion of the candidate gene in wildtype *P. aeruginosa* generated a mutant that was olefin deficient (Rui et al., 2014). Insights from *in vitro* enzymatic assays indicate that UndA requires a Fe^{2+} cofactor to convert lauric acid to 1-undecene. Interestingly, UndA with a Fe^{3+} cofactor did not result in olefin production. This led to the assignment of UndA as a non-heme iron (II)-dependent enzyme (Rui et al., 2014). The redox partner remains elusive (Rui et al., 2014). With the crystal structure of UndA, insights from structural analysis indicate the presence of a hydrophobic pocket from the surface to the centre of the enzyme (Rui et al., 2014). It is hypothesised that the depth of the pocket is an important factor for determining substrate length (Rui et al., 2014), with residues around the active site coordinating substrate binding to the iron-complex, once the fatty acid substrate complexes with the Fe^{2+} cofactor in the active site and the cofactor is able to abstract the β-hydrogen of the substrate. This generates a radical that cascade to the terminal carboxyl group, forming a CO_2 and an olefin product. A reductant restores the active site cofactor to regenerate the Fe^{3+} to Fe^{2+} (Rui et al., 2014).

UndB was also discovered in a *Pseudomonas spp.* containing UndA. It was noted during hydrocarbon screening that there is a two-order of magnitude discrepancy in 1-undecene content between different *Pseudomonas* that could not be explained by transcriptional level and substrate availability (Rui et al., 2015). This led to the hypothesis for an alternative 1-undecene biosynthetic route. UndB was identified using a combination of transformation-associated recombination cloning of the *Pseudomonas* genome and substrate feeding of transformants (Rui et al., 2015). The gene identified had no sequence homology with *undA*, thus was a unique olefin-encoding gene. This gene was named *undB* and bioinformatic analysis indicate that it belongs to a membrane-bound fatty acid desaturase family (Rui et al., 2015). Enzymatic assays indicate that UndB catalyse olefin biosynthesis via a $C_{(n-1)}$ decarboxylation of lauric acid similar to that of OleT and UndA (Rui et al., 2015). Moreover, isotopic-labelling alludes to a β-hydrogen abstraction mechanism responsible for carbon-carbon scission (Rui et al., 2015). Genomic analysis of the *Pseudomonas* genus revealed that UndB homologs are present in other species of *Pseudomonas* but are not as widespread as UndA (Rui et al., 2015).

1.2.3 Head to Head Condensation

Polyolefenic hydrocarbons with a range of C25 to C33 were observed in Gram-positive bacteria such as *Sarcina lutea* and *Micrococcus luteus* (Albro et al., 1969, Beller et al., 2010). The biosynthetic genes that encodes carbon biosynthesis were identified in *M. luteus.* It was observed that the draft genome of *M. luteus* contained additional genes involved in fatty acid condensation, such as a homolog of *fabF* and two homologs of *fabH* (Beller et al., 2010). When these candidate genes were heterologously expressed in *E. coli*, one homolog of FabH was able to elicit a different metabolic profile (Beller et al., 2010). This *fabH* has the lowest homology to known Gram-negative *fabH*. Enzymatic assays of the gene product and tetradecanoyl-CoA resulted in the formation of 27:2 monoketone (Beller et al., 2010). This indicated that a long chain monoketone is formed by condensing two units of tetradecanoyl-CoA and the gene was reassigned as *oleA* to reflect the activity it encodes. Two other genes, *oleBC* and *oleD,* were found in the proximity of *oleA*, suggesting the presence of a gene cluster. The expression of all three genes in *E. coli* resulted in the production of polyolefins and individual expression of these genes did not result in polyolefins, thus confirming the importance of the cluster to polyolefin biosynthesis, which is known as the *oleABCD* cluster (Beller et al., 2010).

The *oleABCD* cluster in *Shewanella spp.* is responsible for the biosynthesis of alkenes that are highly unsaturated, containing nine internal double bonds (Sugihara et al., 2010, Sukovich et al., 2010). The cluster in *Shewanella spp*. exists as a four gene operon and examination of the genes concluded that oleBC is a fusion protein in *M. luteus* (Sukovich et al., 2010). Biochemical characterisation of OleA, OleB, OleC and OleD was able to elucidate each protein's activity and indicated that *oleA* encodes a thiolase, *oleB* encodes an alpha/beta hydrolase, *oleC* encodes an AMP-dependent ligase/synthase, and *oleD* encodes short-chain dehydrogenase/reductase (Sukovich et al., 2010). The biosynthetic pathway starts with OleA thiolase. In this step, a non-decarboxylative Claisen condensation of two fatty acyl-CoA molecules, such as tetradecanoyl-CoA (C14), results in the production of a C27 β-keto acid, 2-myristoylmyristic acid (Frias et al., 2011). The next step is catalysed by OleD, an NADPH-dependent 2-alkyl-3-ketoalkanoic acid reductase, resulting in the production of a hydroxyl alkanoic acid (Bonnett et al., 2011). In the last step, OleC generates a thermally labile β-lactone that is able to spontaneously decarboxylate to an alkene. However, it is currently believed that this final reaction is catalysed *in vivo* by OleB (Kancharla et al., 2016, Christenson et al., 2017b). Recent evidence indicates that OleB, OleC, and OleD enzymes assemble into a large, multiprotein complex (Christenson et al., 2017a).

1.2.4 Terpene-derived hydrocarbons

Terpenoid pathways are capable of producing structurally diverse hydrocarbons, known as terpenes. Two isoprene isomers, IPP and DMAPP, are key substrates for terpene biosynthesis. Isoprenes and terpenes are found universally in all organisms. Plants and most prokaryotes produce IPP and DMAPP via the non-mevalonate pathway. Contrastingly, other eukaryotes, archea and fungi relies on the mevalonate (MVA) pathway for the generation of isoprenes.

Hydrocarbon terpenes can be abundant in plant materials. For the terpene myrcene, this can be up to 39.1% of the total mass in wild thyme (Widén et al., 1977). The terpene-derived compounds farnesane and bisabolene have been previously commercialised for petrochemical and biofuel usage by forming saturated-derivatives (Renninger et al., 2008, Renninger et al., 2010, Peralta-Yahya et al., 2012). Both compounds are classified as sesquiterpenes as they are synthesised from three isoprene monomers and with a molecular formula of $C_{15}H_{24}$. Farnesene is an unsaturated repetitively branched hydrocarbon. The (E,E)-α-Farnesene isomer is found naturally on the surface of fruits and is partly responsible for the scent of apple (Huelin et al., 1966). In insects, farnesene has a role as a semiochemical. With termites it acts as an alarm pheromone (Šobotník et al., 2010) and for moths it acts as a food attractant (Hern et al., 2004). The terpene bisabolene is an unsaturated hydrocarbon containing a cyclohexyl motif with a branched unsaturated acyl group. Bisabolene is detected in the pheromones of insects such as fruit flies and stink bugs (Lu et al., 2001, Aldrich et al., 2007). The production of farnesene and bisabolene in production hosts such as *S. cerevisiae* and *E. coli* only produces unsaturated terpenes, which have limited applications. For the formation of aliphatic biofuels, farnesene, and bisabolene requires chemical hydrogenation to produce farnasene and bisabolene respectively (Renninger et al., 2008, Renninger et al., 2010, Peralta-Yahya et al., 2012).

1.2.5 Engineering Advanced Biofuels

While many organisms are capable of producing alkanes, not many organisms are suitable for viable production at an economic scale. Importantly, wild type producers usually produce a limited range of compounds and if not addressed, will limit blending options. With a rationale bioengineering approach, synthetic pathways can incorporate alkane biosynthetic elements to make the production process more efficient and to produce the range of hydrocarbons required. As fatty acids are the sole substrate for alkane/alkene production, the engineering of fatty acid metabolism has received much attention.

As mentioned in Section 1.2.1, the length of the fatty acid determines the length of alkane produced. Furthermore, it indicated the important role of FAE pathways in producing VLCA. To produce alkanes that are relevant to fuels, production host must be capable of producing alkanes with a product length of C8 to C14. By expressing specific acyl-ACP thioesterases, it is possible to terminate fatty acid elongation to alter fatty acid profile by length (Choi et al., 2013, Howard et al., 2013). This is because acyl-ACP thioesterases are able to cleave acyl-ACP of a specific length. By expressing *fat*B1, a C14 specific acyl-ACP thioesterase from *C. camphora,* in *ado-*

expressing *E. coli*, it is possible to shift the fatty acid profile in favour of tetradecanoic acid. Furthermore, the decarboxylation of tetradecanoic acid led to the production of tridecane in the alkane producing *E. coli* (Howard et al., 2013). Similarly, the engineering acyl-ACP thioesterase can also lead to the alteration of fatty acid length. TesA, an acyl-ACP thioesterases that cleaves long-chain fatty acids, was engineered ('TesA) to accept and hydrolyse a mixture of shorter chain acyl-ACP. When expressed heterologously, 'TesA was able to produce a range of short chain fatty acid, which shifts the fatty acid profile from C16 to C10 (Choi et al., 2013). This led to the production of nonane in *E. coli* expressing both *'tesA* and *cer*1 (Choi et al., 2013). These examples can be considered fatty acid synthase-based pathway in controlling fatty acid length.

Alternatively, it is possible to control the length of fatty acid by a reverse beta-oxidation (RBO) approach. For the production of shorter chain alkanes (C3 to C5), a reverse beta-oxidation approach was demonstrated to be effective in generating short chain fatty acids and, therefore, short chain alkanes (Sheppard et al., 2016). This required the addition of a four-gene module containing thiolase (BktB), reductase (PhaB), dehydratase (PhaJ), and enoyl-CoA reductase (Ter) (Sheppard et al., 2016). In an ADO-expressing *E. coli* host, this was able to elicit propane and pentane (Sheppard et al., 2016).

$C_{(n-1)}$ cleavage leads to odd chain alkane formation because the majority of endogenous fatty acids are even-chain. This limits the alkane biosynthesis to odd chain alkanes. Therefore, it is important to replicate even-chain alkanes to expand the profile of advanced biofuel. FabH2, a 3-oxoacyl-ACP synthase 3 protein 2, from *B. subtilis* has a wider substrate specificity than native FabH of *E. coli* and was able to initiate the condensation of the C3 propionyl-CoA instead of C2 acetyl-CoA in *E. coli*. As downstream fatty acid elongation extends fatty acids by C2, the odd-numbered chain length is preserved (Harger et al., 2012). In an *E. coli* expressing ADO, the addition of FabH2 was able to lead to an accumulation of odd chain fatty acids and was able to produce even chain tetradecane and hexadecane (Harger et al., 2012). The yield of even chain alkane was not as high as odd chain alkane, although the yield discrepancy can be reduced by the supplementation of propanoate to the production host (Harger et al., 2012).

Branched chain alkanes are important fuel molecules for the blending of advanced biofuels. The biosynthesis of 1-methyl alkanes can be engineered into *E. coli.* By introducing β-ketoacyl acyl-carrier-protein synthase III (KASIII) and branched-chain α-keto acid dehydrogenase (BCKD) from *B. subtilis*, *E. coli* was engineered to endogenously incorporate branched-chain amino acids (isoleucine, leucine, and valine) to the elongation of fatty acids (Howard et al., 2013). This was able to induce the *de novo* production of 1-methyltetradecanoic acid and 1-methylhexadecanoic acid, which was further decarbonylated to produce branched alkanes of the corresponding length, 1-methyltridecane and 1-methylpentadecane (Howard et al., 2013). This showed that by altering the profile of fatty acids, by adding branched elements or unsaturated bonds, it is possible to alter the hydrocarbon profile of the production host.

1.3 Fungi and Secondary Metabolites

Fungal secondary metabolites are chemically diverse compounds. It is this large chemical diversity that gives many secondary metabolites interesting functions and activities. Some fungal secondary metabolites have bioactivities suitable for a variety of applications such as antimicrobial agents, cholesterol modulation, or cancer treatment (Hoffmeister et al., 2007).

Typically, secondary metabolite genes in fungi are arranged in gene clusters (Keller et al., 2005). Biosynthetic gene clusters (BGC) are a collection of genes relating to a specific metabolite arranged spatially adjacent in the genome of a fungus. Gene clusters contain genes encodes metabolite-specific pathways, genes encoding resistance and may contain pathway-specific regulatory genes or element (Keller et al., 2005, Regueira et al., 2011). Tools such as AntiSMASH and SMURF incorporates machine learning algorithms for the detection of the biosynthetic genes and associated clusters (Khaldi et al., 2010, Medema et al., 2011). Once a biosynthetic gene or genes have been identified, a detailed search can be performed around the location of the gene in an attempt to identify other potential genes. A cluster can then be defined if the region contain one or more biosynthetic genes.

BGCs commonly contain biosynthetic genes that encodes polyketide synthases, non-ribosomal peptide synthetases, terpene cyclases and prenylation synthetases (Keller et al., 2005, Medema et al., 2011). Other BGCs may contain genes that confer resistance to BGC associated secondary metabolites. For example, these genes can encode anabolic enzymes that breakdown the metabolite or efflux pump for the removal of the associated intracellular metabolites (Regueira et al., 2011, Tang et al., 2015). The presence of resistance genes can be use to guide elucidation of specific BGC. Using this principle, previous work was able to identify IMP dehydrogenase, an enzyme that catalyses the breakdown of mycophenolic acid. By locating the gene that encodes IMP dehydrogenase, the BGC responsible for mycophenolic acid biosynthesis from *Penicillium brevicompactum* was elucidated (Regueira et al., 2011).

Regulation of BGCs is dependent on environmental conditions such as carbon and nitrogen source, nutrient availability, ambient temperature, light and pH (Bennett et al., 1983). The fungus *Aspergillus flavus*, for example, grows optimally at 37 °C and the biosynthesis of aflatoxins is activated at 30 °C (Ogundero, 1987). Alternatively, BGCs can be regulated by cluster-specific regulators including DNA-specific or ligand-specific binding (Schmitt et al., 2000, Chang et al., 2013). This can be observed in the biosynthesis of melanin, in which DNA-binding by zinc fingers activate the BGC for melanin (Tsuji et al., 2000). This pinpoints the need to understand microbial physiology of biochemically interesting organisms for the elucidation of BGCs, as the activation condition of one BGCs can vary.

There are challenges associated with discovering and characterising novel BGCs and its associated metabolites. Frequently, BGCs are transcriptionally dormant under standard laboratory conditions. It is estimated that only 10% to 20% of BGCs are characterised in well-studied fungal systems like *Aspergillus nidulans*, *Penicillium chrysogenum*, *Aspergillus terreus*, and *Saccharopolyspora erythraea* (Oliynyk et al., 2007, Van Den Berg et al., 2008, Yeh et al., 2016). This creates a dilemma for characterising novel and inactive BGCs. Are the predicted BGCs transcriptionally dormant or are they misidentified by genome mining pipelines? Even when a BGC is genuine, what are the conditions required to activate it?

Heterologous expression of BGCs in a production host has been successful in overcoming dormant BGCs. This was demonstrated in the expression of neothioviridamide BGC. The cluster was discovered during genome mining of *Streptomycete spp.* to search for thioviridamide-like BGCs, a previously characterised cluster (Kawahara et al., 2018). This led to the identification of a gene encoding a precursor peptide containing a VMAAAATVAFHC motif in the genome of *Streptomyces sp*. MSB090213SC12 (Kawahara et al., 2018). Different fermentation conditions were explored in an attempt to activate the putative BGC in *Streptomyces sp*. MSB090213SC12 but this did not induce the BGC (Kawahara et al., 2018). In an attempt to activate the putative BGC, a heterologous expression approach was favoured instead. A refactored putative BGC was cloned and expressed into *S. avermitilis* SUKA22 resulting in successful expression. Biochemical characterisation of the transformant's extract led to the discovery of a compound named neothioviridamide, a novel thioviridamide derivative (Kawahara et al., 2018). The heterologous expression approach itself has its associated disadvantages. It is impractical for cloning methods to accommodate large BGCs, leading to poorer transformation efficiency. Moreover, refactoring and redesigning fungal BGCs becomes more complex for larger clusters. This can include identifying and removing elements such as introns and intergenic regions to reduce overall size (Greunke et al., 2018, Duell et al., 2019). Lastly, production hosts may be incompatible with novel BGCs, leading to host toxicity or failed expression (Zhang et al., 2019).

For some cryptic BGCs, the usage of epigenetic modulators can lead to activation and alteration of metabolite profiles. The epigenome can be manipulated by two class of compounds, histone deacetylase (HDAC) and DNA methyltransferase (DMAT) inhibitors. The DNA uncoiling and coiling process is mediated by histone acetyl transferases and histone deacetylases respectively. When a region of DNA is acetylated it is uncoiled from the histone and the exposed DNA is accessible for binding by transcription-based protein thereby increasing gene expression. In contrast, deacetylation of DNA leads to an increase in binding with histone and this concealment leads to transcriptional deactivation. By supplementing HDAC inhibitors into the growth media, HDAC interrupts the coiling process and can lead to an increases gene expression. Lower expression has been linked to regions with a higher proportion of methylated DNA. There are two likely explanations for the transcriptional regulation as a result of methylation. Firstly, methylation physically interferes with binding of transcriptional proteins to the gene (Choy et al., 2010). Secondly, regions of methylated DNA may be bound by proteins known as methyl-CpG-binding domain proteins (MBDs) (Nan et al., 1993). It is thought that binding to MBDs recruits histone and histone-associated proteins that leads to the physical binding of the methylated region (Zemach et al., 2003). Therefore, DMAT inhibitors can interfere with this regulation process. This was the case in *Cladosporium cladosporioides* and the addition of HDAC and DMAT inhibitors stimulated the production of several oxylipins, cladochromes and of calphostin B not produced under standard conditions (Williams et al., 2008).

1.3.1 Ascocoryne sarcoides as an alkane producer

The fungal endophyte *Ascocoryne sarcoides* NRRL 50072 was isolated from Patagonian forest (Stinson et al., 2003) and is capable of producing a diverse set of secondary metabolite volatile organic compounds (VOC) (Strobel et al., 2010b). This includes linear and cyclic alkanes that are suitable for transport fuel and petrochemical applications. The *A. sarcoides* fungus can be found widely distributed across temperate forests of Europe, North America, and Patagonia (Roll-hansen et al., 1979, Metzler, 1997, Stinson et al., 2003). Sightings of the fungi are common in healthy living stems of *Picea spp.* (Roll-hansen et al., 1979, Metzler, 1997), indicating the endophytic role of the fungus. This distribution is reflected in culture collections, of the six publicly available isolates of *A. sarcoides,* five were isolated from European forests and a singular isolate collected from Canada. This is further expanded upon in Table 1.2. When a fungus enters the sexual phase of its life cycle, it produces a fruiting body for the propagation of spores. The fruiting body of *A. sarcoides* resembles a purplish jelly-like disc that protrudes from a tree bark. To date, there no studies are able to to elicit fruiting body formation.

Table 1.2. Origin and host of *Ascocoryne sarcoides* isolates. Six different isolates that were studied in this thesis. (*) = denotes culture not in public collection.

Isolate ID	Origin	Host	Date of isolate
170.56	UK	*Nothofagus* sp.	1913
171.56	Germany	*Pinus contorta*	1913
246.80	Norway	*Picea abies*	1966
309.71	Switzerland	Quercus wood	1967
44013	Norway	*Malus domestica*	1966
64019	Canada	*Betula alleghaniensis*	Unknown
NRRL 50072 *	Pantagonia	*Eucryphia cordifolia*	2003

The initial taxonomical identity of NRRL 50072 was unclear. Early attempts to resolve the identity by phylogenetics was confounded by the lack of genomic databases for comparison (Strobel et al., 2010a). Morphological studies were used instead, which identified NRRL 50072 as an isolate of *Gliocadeum roseum* (Strobel et al., 2010b). Later phylogenetic analysis reassigned isolate NRRL 50072 to be a species of *Ascocoryne sarcoides* after homology analysis of the internal transcribed spacer (ITS) indicates a significant alignment to an ITS of *A. sarcoides* (Griffin et al., 2010).

When *A. sarcoides* NRRL 50072 was cultured and sampled by solid-phase microextraction (SPME) coupled with proton-transfer reaction mass spectroscopy

(PTR-MS), a diverse set of hydrocarbons were detected that included linear and cyclic alkanes. Moreover, the volatile profile was altered when *A. sarcoides* was cultured in different rich growth media (Strobel et al., 2010b). After NRRL 50072 was reassigned to *A. sarcoides*, it led to further questions concerning the metabolic potential of the *Ascocoryne* genus. Three isolates of *A. cylichnium*, one isolate of *A. solitaria* and seven isolates of *A. sarcoides* were sampled (Griffin et al., 2010). Furthermore, comparable metabolites were detected in other *Ascocoryne* species and isolates. Linear hydrocarbons were detected across the genus. However, cyclic alkanes were only detected in the isolate of *A. sarcoides* 309.71 and NRRL 50072, indicating cyclic alkane biosynthesis may be restricted to a species level to *A. sarcoides* (Griffin et al., 2010). Across multiple studies, linear alkanes and alkenes with the length of C5 to C9 and long chain alcohols were detected in each isolate of *Ascocoryne*. Moreover, cyclic alkane was only observed in *A. sarcoides* isolates 309.71, 44013 and NRRL 50072. Across these studies, the VOCs profile is highly varied in hydrocarbon composition (Stinson et al., 2003, Griffin et al., 2010, Strobel et al., 2010b, Gianoulis et al., 2012, Mallette et al., 2012, Mallette et al., 2014). Since then, a multitude of studies have independently observed *A. sarcoides* ability to produce linear and cyclic alkanes. These include a study to establish an optimal chemically defined medium to culture NRRL 50072. Physiological experiments indicate the importance of substrate, growth stage, pH, temperature and medium composition in affecting the quantity and composition of hydrocarbons produced by *A. sarcoides* (Mallette et al., 2014). As an endophytic hydrocarbon-producing fungi, the cellulolytic potential was explored to establish *A. sarcoides* as a possible industrial hydrocarbon-producer. Secondary metabolite such as alkanes, alcohols, ketones, esters and various other hydrocarbons including derivatised benzenes, cycloalkanes when *A. sarcoides* NRRL 50072 was cultured in rich cellulose medium (Griffin et al., 2010) and, as well as, chemically-defined medium supplemented with cellulose substrate (Gianoulis et al., 2012, Mallette et al., 2014). While *A. sarcoides* is metabolically active in these conditions, these findings did not account for the presence of lignified-cellulose in cellulosic biomass.

The genome and the transcriptome of *A. sarcoides* NRRL 50072 was developed to aid discovery of hydrocarbon biosynthetic genes (Gianoulis et al., 2012). The transcriptome was based on three different growth conditions and two time points, yielding six different producing conditions which included the production of C7-C9 alkanes (Gianoulis et al., 2012). The two bioinformatic datasets enabled the prediction of an *in silico* inventory of the gene product (Gianoulis et al., 2012). By correlating the metabolome, transcriptome, and genome, it enabled the prediction of a hypothetical linear hydrocarbon pathway (Gianoulis et al., 2012). However, this proposed pathway deviates significantly from the C_{n-1} paradigm associated with known alkane biosynthetic genes. Furthermore, this proposed pathway predicts the biosynthesis of terminal olefins and does not account for the linear or cyclic alkane observed during SPME sampling. Additionally, genes annotated for the alkane forming steps have homology to well characterised fatty acid β-oxidation enzymes. The hypothetical pathway remains experimentally untested.

1.4 Aims and objectives

Previous studies have established *A. sarcoides* as a hydrocarbon producer. Observations of the volatilome allude to a unique metabolic profile for each *A. sarcoides* isolate. However, the hydrocarbon biosynthetic pathway has yet to be established. The aim of this project is to investigate alkane biosynthesis in *A. sarcoides* and to elucidate the key genes involved in the pathway to expand the toolkit for future advanced biofuel endeavours. This project will use six publicly available isolates of *A. sarcoides* to investigate the alkane producing potential. The fundamental microbiology of all six isolates will be established for metabolomic further investigations. Moreover, the *A. sarcoides* metabolome will be studied in detail for the sole purpose of identifying fuel compounds for biofuel applications. This will differ from previous studies by adopting methods that are specific to detect linear and cyclic alkanes. Each isolate will be sequence to develop detailed genomic and *in silico* proteome datasets. This will be used for bioinformatics interrogation and to explore the biochemical potential of *A. sarcoides*. The following data chapters will cover key objectives of the investigation in this thesis and are listed as follows:

Chapter 2: Materials and methods
Chapter 3: Understand the fundamental microbiology and husbandry of *A. sarcoides*
Chapter 4: Developing high quality bioinformatics dataset
Chapter 5: Exploring the biochemical potential of *A. sarcoides*
Chapter 6: Exploring the metabolome of *A. sarcoides* for biofuel application
Chapter 7: General discussion and conclusion

CHAPTER 2 MATERIALS AND METHODS

2.1 Microbiology

2.1.1 Isolate stocks

Inoculants of *Ascocoryne sarcoides* isolates 170.56, 171.56, 246.80, 309.71, were obtained from CBS-KNAW culture collection. Isolate 44013 and 64019 were obtained from ATCC. Each isolate was reconstituted according to the suppliers' instructions, with the CBS isolates cultured on oatmeal agar (OMA) and ATCC isolates cultured on potato dextrose agar (PDA). Mycelium were subcultured after 21 days of incubation at 25 °C.

Viable stocks were stored in sterile distilled water at 4 °C This was used for culture inoculation. Viable stocks were directly prepared from -80 °C long term stock and long term stocks were stored with a paper based method (see below in Section 2.2.4).

2.1.2 Medium composition

Potato dextrose (PD) broth was prepared to a concentration of 24 g/L (obtained from Formedium). Defined medium (DM) was prepared with 5 g/L Ammonium chloride, 2.75 g/L Magnesium sulphate, 0.86 g/L Sodium monophosphate, 0.28 g/L Calcium nitrate and trace salt solution. The trace salt solution is composed of 80.0 mg/L Potassium nitrate, 60.0 mg/L Potassium chloride, 5.0 mg/L Manganese chloride, 2.5 mg/L Zinc sulphate, 2.0 mg/L Iron chloride, 1.4 mg/L Boric acid and 0.7 mg/L Potassium iodide). For solid media, the corresponding medium is further supplemented with 15 g/L of agar. Each component was autoclaved at 121 °C for 15 minutes.

Defined medium were also supplemented with a carbon source (obtained from Sigma-Aldritch), 20 g/L glucose (≥99%, Sigma-Aldritch) was prepared by filter sterilisation with a 0.22 µm filter, 15 g/L monosodium glutamate (≥99%, Sigma-Aldritch), 20 g/L acetic acid and 20 g/L cellobiose (≥98%, Sigma-Aldritch) were prepared by autoclaving. Carbon sources such as 15 g/L cyclodecane (>90%, Sigma-Aldritch), 15 g/L tetradecanoic (≥99%, Sigma-Aldritch) acid and 15 g/L tetradecane (olefine free ≥99.0%, Sigma-Aldritch) were added directly into the media prior to inoculation. Components of the media were assembled combinatorially with a liquid handling robot from stock concentration to achieve the required final concentration.

2.1.3 Culture conditions

In experiments involving conical flasks (non-baffled), 20% of the flask's total volume (e.g. 20 ml of medium was used to fill 50 ml conical flasks). In experiments with 96 well microtitre plates, wells were filled with 200 µL of medium. For liquid cultures, the colvume of inoculant were adjusted to 1% (v/v) of total culture volume. While for solid cultures, fungal plugs were used for subculturing. Plugs were taken from distal fungal filaments to ensure fresh fungal growth. Additionally, plugs made from the corresponding agar for the required subcultures (e.g. PDA plugs were used for PDA subculturing). Lastly, all cultures were grown in temperature controlled incubators at 23 °C, with no light, and at a speed of 120 rpm for liquid cultures.

2.1.4 Measuring growth

Measuring cultures in flasks

For liquid flask cultures, growth were measured by colony forming units (CFUs). 100 uL samples were taken and sequentially diluted with 900 uL of Ringer's solution. All diluents were then plated on PDA and were incubated at 23 °C. Colonies were counted usually at 7 to 10 days or until colonies were observable.

Measuring cultures in microtitre plates

For liquid 96 well plate cultures, growth was measured with a Varioskan Lux spectrophotometer. Plates were shaken at 120 rpm for 10 seconds prior to measurements and OD_{600} absorbance measurement was recorded.

Measuring cultures on agar

For solid agar cultures, growth were measured by diameter at daily intervals.

2.1.5 Long term storage

Fungal material was harvested from each isolate cultured on PDA plates. To ensure minimum bias for a phenotype, fungi materials were selected from 4 different plates per isolates and 4 different samples from per plate for each storage protocol.

Storage by filter paper

Cellulose filter paper was cut into approximately 1 cm^2 squares. The filter paper was placed in a sealed bag to prevent dampness and was autoclaved for 121 °C for 15 minutes. Filter papers were then placed on PDA and followed by the addition of fungal plugs on the filter paper (Fong et al., 2000). The cultures were then incubated for 15 days at 23 °C to allow colonisation on the filter paper medium. Once the fungi proliferated, the paper bearing fungi material was separated from the underlying media and was then desiccated in a vacuum desiccator with silica gels for 14 days. The desiccated fungi material was kept in a sterilised plastic tube and stored in -80 °C.

Storage by glycerol and skimmed milk

The glycerol-milk solution was prepared with 17% skim milk and 20% glycerol was mixed 1:1 after autoclaving to achieve a final concentration of skim milk 8.5%, glycerol 10%. 2ml of the glycerol-milk mixed was aliquoted into sterile cryogenic tubes with internal tube threads and followed by the introduction of fungal material. The glycerol-milk with fungal material is then kept in long term storage at -20 °C.

Storage by sterile distilled water

The fungal mycelium was suspended in 20ml sterile distilled water in a sterile tube and was stored at 4 °C and away from light source.

Storage by agar slants

Slants were made from 20ml of PDA and left to cool at a 45° angle in a falcon tube. Fungi material was subcultured on the surface of the agar and deposited at half the depth of the agar. It is then left to incubate for 14 day at 23 °C. Slants were stored at 4 °C and away from light source.

2.2 Genome preparation

2.2.1 Samples preparation and sequence

All isolates were cultured on PDA and as stated in 2.2.2. Genomic extraction for isolates 170.56, 171.56, and 246.80 were performed according to Qiagen Genomic Extraction kit. Illumina sequencing was performed for each isolate using paired 100 nt Illumina HiSeq 2500.

Isolates 309.71, 44013 and 64019 were prepared according to the instructions of MicrobesNG. DNA genome extraction and sequencing were performed by MicrobesNG. Illumina sequencing was performed for each isolate using paired 250 nt Illumina HiSeq 2500.

2.2.2 Post-sequencing QC

Raw Illumina libraries were quality checked using FastQC (v0.11.8) (Andrews, 2010) and trimming was performed using Trimmomatic (v0.32) (Bolger et al., 2014) with a minimum Phred score of 28. Both software perfomed on a linux workstation. MAXINFO algorithm was used on all libraries.

2.2.3 Genome assembly

Trimmed Illumina libraries were assembled with SPAdes (v3.12) (Bankevich et al., 2012) on a linux workstation. These produce draft genomes for each assembly. Assembly were checked for quality with QUAST (v4.6.0) (Gurevich et al., 2013) on a linux workstation and completeness metric were analysed with BUSCO (v3) on a linux workstation with OrthoDB ascomycota (v9) (Simão et al., 2015) Augustus (v2.5.5), HMMer (v3.1b2), BLAST+ suite (v2.7.1) (Eddy, 1998, Stanke et al., 2004, Camacho et al., 2009). BUSCO was perfomed on genome mode and with Augustus HMM species training set to *Aspergillus nidulans*.

2.2.4 ORFs gene prediction

Ab initio gene prediction was achieved with MAKER2 (v2.31.9) on a linux workstation (Holt et al., 2011). Dependencies included Augustus (v2.5.5), HMMer (v3.1b2), BLAST+ suite (v2.7.1), RepeatMasker (v4.0.7), Exonerate (v2.2.0), and snap (v0.15.4) (Stanke et al., 2004, Johnson et al., 2008, Camacho et al., 2009, Tarailo-Graovac et al., 2009, Finn et al., 2011). mRNA seq data and protein evidence from NRRL 50072 were included in the MAKER2 to improve annotation. These were obtained from JGJ genome's database (Gianoulis et al., 2012) (NRRL50072 v1.0, https://genome.jgi.doe.gov/Ascsa1/Ascsa1.home.html). Species specific HMM training were achieved by using BUSCO's Augustus derived HMM model and SNAP HMM was derived from two separate round of MAKER. In total, two rounds of maker were used to produce the final annotation set for each isolate's draft genome. The number of predicted genes were derived from linux OS grep command.

2.2.4 ORFs gene annotation

ORFs were annotated with standalone BLAST+ suite (v2.7.1) (Camacho et al., 2009) and perfoemed on Newcastle University's Rocket linux HPC. A custom python script was use to split annotation set (>10,000) to 180 entry FASTA accession. Another custom bash script produce appropriate SLURM script based on the aforementioned fragment and submit the BLAST query job. This was to ensure parallelisation of BLAST jobs to take advantage of nodular architecture of a HPC. The e-value for each job was set to 1. Results were concatenated to their respective annotation set upon completion of BLAST jobs. Another python script was used to extract accession code with an e-value of 1e-1 from output format 7 of BLAST results. Number of queries and significant hits were derived from linux OS grep command.

Each annotation set was queried against database including: NCBI NR (accessed: 10/12/2018), NCBI fungi refseq (accessed: 10/12/2018), UniProt's SwissProt (accessed: 10/12/2018), UniProt's *Aspergillus nidulans* FGSC A4 (taxonomy: 227321, accessed: 16/11/2018) and UniProt's *Sacchromyces cerevisiae* S288C (taxonomy: 227321, accessed: 16/11/2018). These database were formatted by makeblastdb tool from BLAST+ suite (v2.7.1).

2.2.5 PANTHER ontology

Once an isolate's UniProt's accessions were identified, the identified accession was enriched by PANTHER (v14, http://pantherdb.org) classification (Mi et al., 2013). Functional classification analysis was ran against a PANTHER's reference *Emericella nidulans* and on default settings*.*

2.2.6 KEGG ontology

Each isolate MAKER2's predicted amino acid set was submitted to KEGG Automated Annotation Server (v2.1) (Moriya et al., 2007) and was set to bi-directional best hit method. BLAST was chosen as the search program and was set to an e-value of 1e-60. The annotation job was queried against KEGG's database containing fungal species from every major phyla: *S cerevisiae, N crassa, S sclerotiorum, A nidulans, P nodorum, T melanosporum, C neoformans var. neoformans JEC21, E cuniculi,* and *S pombe.* Moreover, two alkane producing species were also included: *A thaliana,* and *D melanogaster.*

2.2.7 Phylogeny

Phylogeny was based on single copy orthologs protein sequence recovered with BUSCO genome analysis. This ensured that only conserved genes were used. In addition to all isolate genomes of *A. sarcoides* (including NRRL 50072), we included *S. cerevisiae* S288C (GCA_000146045.2)*, A. nidulans* FGSC A4 (GCA_000011425.1) and *B. cineria* B05.10 *(GCA_000143535.4)* into the analysis and were obtained from European Nucleotide Archive. With a python script, 1049 complete single copy ortholog proteins were identify to be common to all genomic dataset and amino acids sequence were concatenated respective to their origin. This produce a FASTA file containing a concatenated sequence of approximately 700,000 predicted protein recovered by BUSCO per genome. The resulting file was analysed with MAFFT webserver Katoh, (Kuraku et al., 2013; Rozewicki, Yamada 2017). With strategy set to auto, gap alignment set to on, and with default parameters. Phylogenetic tree generation was made with memory-saving tree, with a JTT substitution model and was drawn on Phylo.io (v1.0.0).

2.3 Comparative analysis

2.3.1 Identifying orthologous cluster

A centralised database containing predicted genes and associated annotation evidences generated in Section 2.2 were compiled. Each predicted gene was given a unique accession number for auditing purposes.

Reciprocal network analysis was used to identify orthologous genes. To reduce computational work load, amino acid sequences were used. The proteome of each isolate, NRRL 50072, *A. nidulans* FGSCa4, and *S. cerevisiae* S288c were included for this comparative analysis. These databases were formatted by makeblastdb tool from BLAST+ suite (v2.7.1). To identify orthologs in an another isolate and/or organism, a reciprocal hit was used. In a reciprocal hit, the query must align significantly to the subject and the subject must also significantly aligns with the query. If two sequences aligned from another proteome, it is considered to be an ortholog. If two sequences aligned from within the proteome, it is considered to be a duplicated gene. This was achieved by cross comparing against other proteome by BLASTP tool from BLAST+ suite (v2.7.1). For the BLASTP analysis, BLASTP was set to produce a single match, single highest scoring pairs, and no cut-off score was set.

In post-processing a cut-off of 1 e-3 was set for the e-value and 90% was set for the percentage identification. To identify non-redundant orthologs, a custom python programme was built. This programme functions by networking filtered matches. For example, if A, B, and C forms an orthologous cluster if D is reciprocal to A, it is also considered to be reciprocal to B and C. Conversely, if E is significantly align to A but A is not significantly E, then it is not considered to be reciprocal and, by extension, not orthologous to A, B or C. This enabled the identification of cluster of orthologs across all isolate and other species. If an orthologous cluster contain a match with *S. cerevisiae* and *A. nidulans*, the orthologous cluster is considered not unique to *A. sarcoides.* Inversely, the cluster that contain only *A. sarcoides* are considered to be unique. Comparative analysis was also achieved by OrthoVenn2. Orthovenn2 also annotated genes against UniProt database. The non-redundant orthologous cluster can be examined for biological relevances.

2.4 Hydrocarbon screenings

2.4.1 Hydrocarbon screening by solvent extraction

A. sarcoides were cultured in 250 mL flask and with 50 ml of defined medium and incubated in the dark for 21 days at 23 °C and 120 rpm. Deuterated chloroform was added at a 1:1 ratio to cultures and was left incubated on a stirrer for 30 minutes. This mixture is then decanted into a separatory funnel, which is equipped with a PTFE-lined component. The organic layer is then tapped, evaporated and resuspended with 1 ml of deuterated chloroform. A total of six biological replicates were extracted and analysed.

2.4.2 GC-MS analysis for solvent extraction

Culture headspace samples were analysed using a GC-MS (Agilent 7890GC / 5975MSD). Compounds were injected onto the GC column at 240 °C splitless injection for 30 s with a 0.75 mm ID injection port liner, and were separated on a DB-WAX column (30 m × 0.25 mm ID × 0.50 μm film thickness; Phenomenex). The GC temperature programme was held at 30 °C for 2 min after injection, and increased to 220 °C at 7 °C min−1 with a constant flow rate of 1 ml min^{-1} Ultra High Purity Helium. 5t (EI) spectra were obtained from EI ionization at 70 eV (source temperature

150 °C) and data were collected over the mass range 50–650 Da. Results were analysed with Agilent's proprietary MassHunter.

2.4.3 Experimental set-up by SBSE

Amber glass vials of 20 mL capacity was chosen to avoid plastic leaching, reduce photoreactivity, and give ample headspace for aeration. To pair with the vials, PTFE-lined melamine screw caps were used. The screw caps and PTFE lining ensure cultures were properly sealed and prevent the release of metabolites and introduction of contaminants.

2.4.4 Cleaning glass vials

Glass vials and caps were washed and submerged overnight with 5% (v/v) of Chemgene HLD4 (clear and un-fragranced). Vials were washed and submerged overnight with a caustic solution of 6% (w/v) of KOH and 80% (v/v) of EtOH (99.8%, LC-MS grade Fischer Scientific) to remove any organic residues. Distilled water was used to rinse excess caustic solution. Glass vials were sealed with aluminium foil and baked dry at 130 °C in an oven.

2.4.5 Culturing for microextraction of secondary metabolites

Vials were sealed with aluminium foils and sterilised by autoclaving at 121 °C for 15 minutes. Caps were sterilised with 70% (v/v) EtOH (99.8%, Fischer Scientific), excess EtOH was rinsed with sterile distilled water and excess liquid was dried with sterile cotton in a Category 2 biosafety hood. Vials were unsealed in a Category 2 biosafety hood and were filled with liquid defined medium supplemented with 20 g/L of glucose and 0.01% (v/v) of Tween 80. Inoculation was achieved by transferring 1% (v/v) of mycelium water stock. Additionally, Twister stir bars were added to culture. Twister stir bars were magnetic bar coated with polydimethylsiloxane (PDMS), an inert polymer that enables hydrophobic metabolite adsorption. Every seven days, Twister bars were replaced with a clean bar. Removal was achieved with a pair of magnetised metal forceps. Forceps were washed with methanol, sterilised with 70% ethanol and air dried to minimise chemical and microbial cross contamination. Cultures were incubated in the dark for 21 days at 23 °C and 120 rpm.

2.4.6 Biological non-producing organisms

Aspergillus nidulans FGSC A4 and *Saccharomyces cerevisiae S288C* were chosen as a non-producing hydrocarbon fungus. Both organisms were cultured with defined medium as mentioned in Section 2.1.2 and supplemented with 20 g/L of glucose. The cultures were incubated at 23 °C and 120 rpm.

2.4.7 GC-MS analysis for SBSE

Twister stir bars were analysed with a GC-MS (Agilent 7890B GC and 5977B MSD) equipped with a Gerstel thermal desorption unit (TDU2). It is equipped with a HP-5MS Ultra inert column (30 m x 0.25 mm x 0.25 µm). SBSE were injected with pulsed splitless and GC set to 250 °C with a flow of 1 mL/min and 50 °C held for 2 min, 8 °C/min to 320 °C, held for 9.25 min with a total runtime of 45min. Spectra were obtained by electron ionisation mode at 300°C, Mass range 50-650 m/z. Results were analysed with Agilent's MassHunter and OpenChrome.

CHAPTER 3 HUSBANDRY AND MICROBIOLOGY OF *A. SARCOIDES*

3.1 Introduction

3.1.1 Ascocoryne sarcoides

Ascocoryne sarcoides is a fungus that occupies both endophytic and saprophytic ecological niches. Field observation reports of *A. sarcoides* indicate the fungus inhabits temperate deciduous forest in Europe and North America. Furthermore, field studies observed *Ascocoryne spp.* on 48% of 60 year old spruce (*Picea abies*) stems (Roll-hansen et al., 1979). Fungal endophytes are fungi that coexist in plants and proliferate within plant tissues. Many studies have described the chemical diversity of secondary metabolite of endophytes thus these group of fungi represents a rich source of undiscovered chemistry (Keller et al., 2005, Hoffmeister et al., 2007). Fungal saprophytes are fungi capable of scavenging nutrients from their immediate environment and play a key role in cycling nutrients by decomposing litter. These nutrients are released into the soil and are key to maintaining an ecosystem. It is thought that litter decomposition is facilitated by an array of hydrolytic enzymes that convert complex organic compounds to simpler inorganic compounds.

To leverage the biotechnological potential of *A. sarcoides,* an understanding of the underpinning microbiology of the fungus is essential. For non-model organisms, microbiological isolation and exploitation can be difficult. This means that some organisms behave in an inconsistent manner in laboratory environment. Further developments in the understanding of identification, isolation, cultivation, screening and recovery of *A. sarcoides* are required. One of the major challenges working with alkane metabolising organism is the presence of hydrocarbon based contaminants. These can also be present in culture medium. It is imperative to have a medium that minimises hydrocarbon and is traceable in its chemical composition. Previous studies also report unstable phenotypes in *A. sarcoides* AV-70 when subjected to serial passage, leading to a loss of pigmentation and the inability to form aerial hyphae (Strobel et al., 2010a). Thus it is important to accurately identify phenotypic changes and how to reduce phenotypic drift. Finally, secondary metabolite production is thought to be engaged when microbial cultures are subjected to specific conditions. It is important to establish a medium that can support growth of *A. sarcoides* and allow assay growth*.* Method development enables downstream studies such as enzymology, radioisotope labelling, and metabolite discovery. These studies aid in the elucidation of alkane biosynthesis.

3.1.2 Experimental aims

The aims of this chapter are to develop a comprehensive understanding microbiological husbandry of *A. sarcoides* and build the required foundation ofknowledge for the microbial manipulation. This knowledge will be leverage at later stages of the project, such as sampling for secondary metabolite production and exploration of the the biochemistry of the fungus. Here we outline the microbiology methods for the manipulation of *A. sarcoides*. The objectives are:

- Phenotypic understanding of *A. sarcoides*.
- Developing culturing techniques for *A. sarcoides.*
- Growth kinetics of *A. sarcoides.*
- Growth on different carbon sources.

3.2 Results

3.2.1 Description and Culturing of A. sarcoides

The six *A. sarcoides* isolates used in this study were obtained from public repositories. Isolates 170.56, 171.56, 246.80 and 30971 were from ATCC and were revived on potato dextrose agar (PDA). Additionally, 44013 and 64019 were from CBS and were revived on oat meal agar (OMA). Isolates were successfully revived after 14 days. PDA and OMA are media commonly used in the maintenance of filamentous fungi. At day 15, all isolates were subcultured and maintained on PDA. Furthermore, long term stocks of these were made and will be described in Section 2.1.5.

Isolates were cultured in potato dextrose broth (PDB) and growth was assayed by colony forming units (CFU) at a daily timepoint (Figure 3.1). CFU is a direct measurement of growth and a measurement of viability within liquid cultures. This is performed by plating a sample volume from liquid cultures onto solid medium. It assumes that each colony is formed by a single viable cell. By distinguishing the amount of cells from a given volume, we can estimate the total number of viable cells from a culture. An additional benefit to this technique is the observation of a culture's morphology thus the fitness and viability of a culture. All isolates have similar growth kinetics, with a mean time of 1 to 2 days of lag phase and a mean of 7 to 10 days to reach stationary phase in liquid cultures. Moreover, upon reaching day 10, we observed the liquid medium takes on a darker colouration and a strong odorous scent.

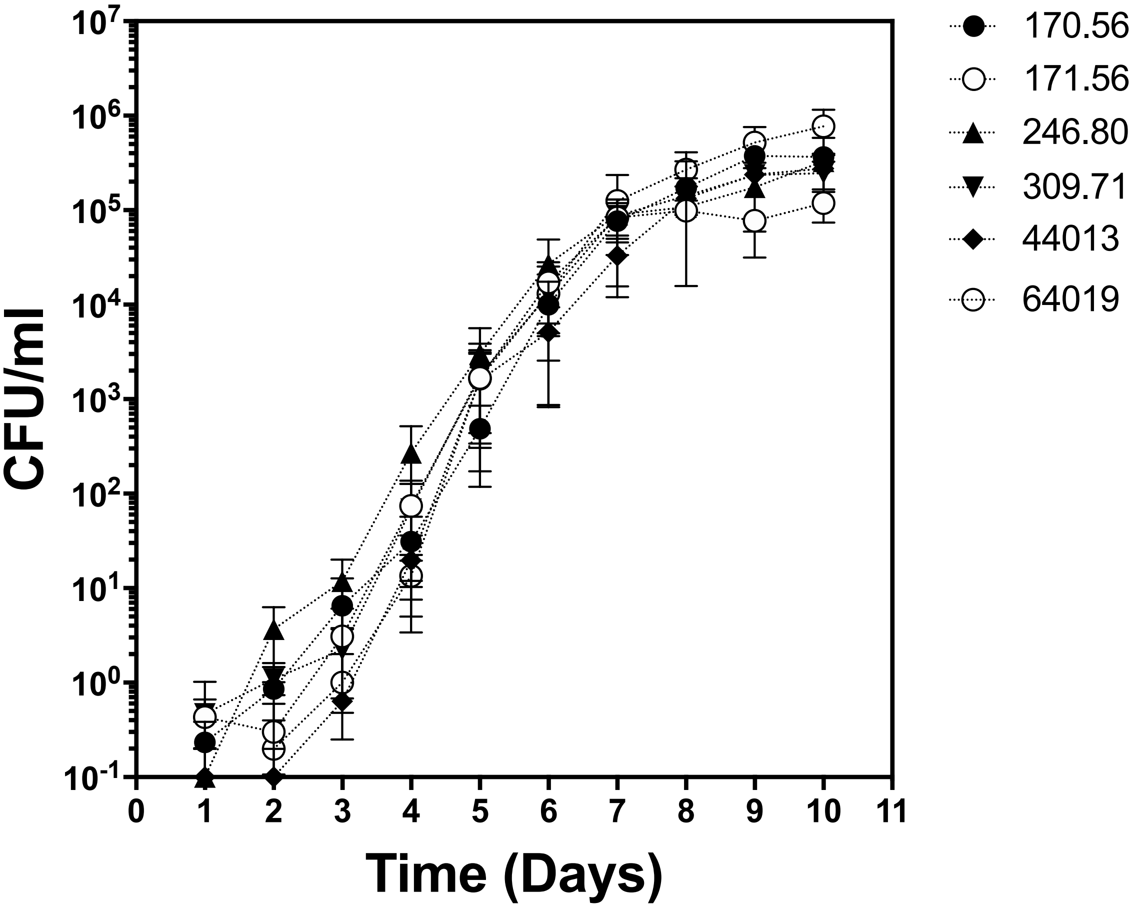


Figure 3.1. CFU growth kinetics of six *A. sarcoides* isolates. All six isolates of *A. sarcoides* were cultured in 50 ml potato dextrose broth in a 250 ml conical flask. Cultures were incubated at 23 °C and 120 rpm for 11 days. Samples of 100 μL were taken from cultures and plated on potato dextrose agar. Colony numbers were measured after 5 days and expressed as colony forming units CFU. Error bars are indicate the standard error of deviation. (n = 3).

When isolates were cultured on PDA (Figure 3.2), typically white filamentous colonies were observable and counted after 3 to 5 days. After 10 days, a dark brown/purple pigmentation was observed on fungal colonies. When cultured in solid media such as defined media agar (DMA), PDA and OMA, purple pigmentation was also observed after 7 to 10 days of culturing (Figure 3.2). The pigmentation is thought to be the antibiotic ascocorynin (Quack et al., 1980). In some cases, development of dark spots were observed, however observations were not consistent. Both observation persisted throughout the incubation of the old cultures. In both cases, development of pigmentation was also followed by a strong pungent odour. After 21 days of culturing on solid medium, aerial hyphae were observed in all *A. sarcoides* isolates grown in rich solid media and solid defined media glucose medium. Additionally, we did not observe formation of fruiting bodies in any cultures.

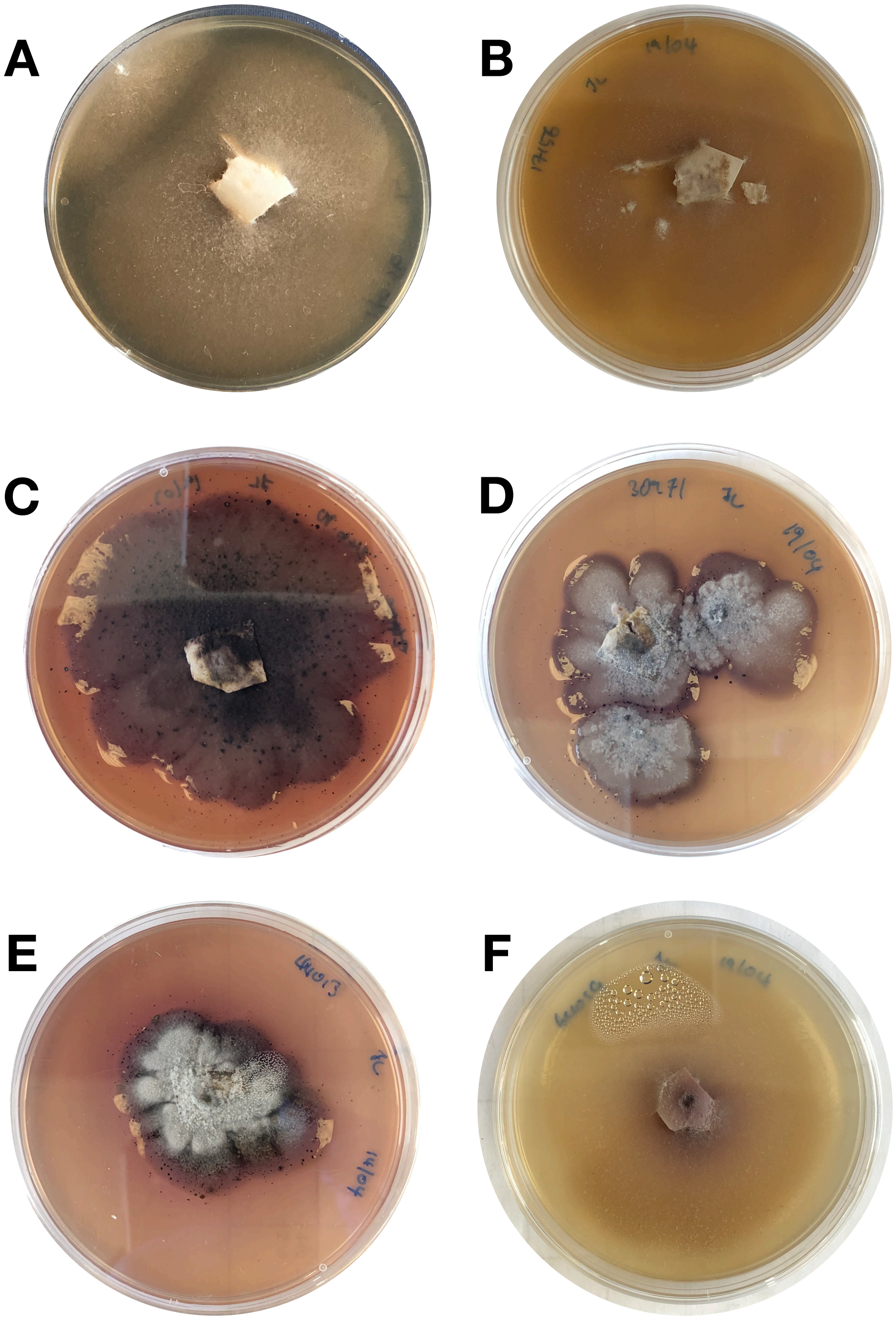


Figure 3.2. Six isolates of *A. sarcoides* cultured on Potato Dextrose Agar. Isolates were cultured on PDA medium. Cultures were incubated at 23 °C for 14 days. The designation represents the following isolates: A) 170.56, B) 171.56, C) 246.80, D)309.71, E) 44013, and F) 64019.

Most growth media, such as PDA or OMA, contain components derived from animal or plant matter. The inclusion of animal or plant matter in the medium facilitates growth by providing vital nutrition in the form of amino acids, lipids, vitamins, and minerals. Yet, with these media, it can be difficult to quantify chemical composition and they offers no traceability. In contrast, a defined medium is a growth medium in which all chemical components are quantifiable and traceable. Usually these chemical components are inorganic and/or highly purified. The use of defined media is necessary for metabolite discovery because it controls for contaminants and minimises variation in composition. This is crucial as alkanes are common environmental contaminants. The composition of defined medium was chosen to include essential inorganic nutrients, exclude chemical components from petrochemical sources and previous work demonstrating the medium's capacity to culture *A. sarcoides* NRRL 50072 (Mallette et al., 2014). We were able to culture all isolates in liquid defined medium supplemented with 15 g/L of glucose (Figure 3.3). When in liquid defined medium, kinetics were comparable to PDB cultures. It had a longer lag phase but was able to achieve a similar growth density during stationary phase. Once in the stationary phase, the clear and colourless cultures began to exhibit a slight purple pigmentation. Again, the colour change persist during culturing.

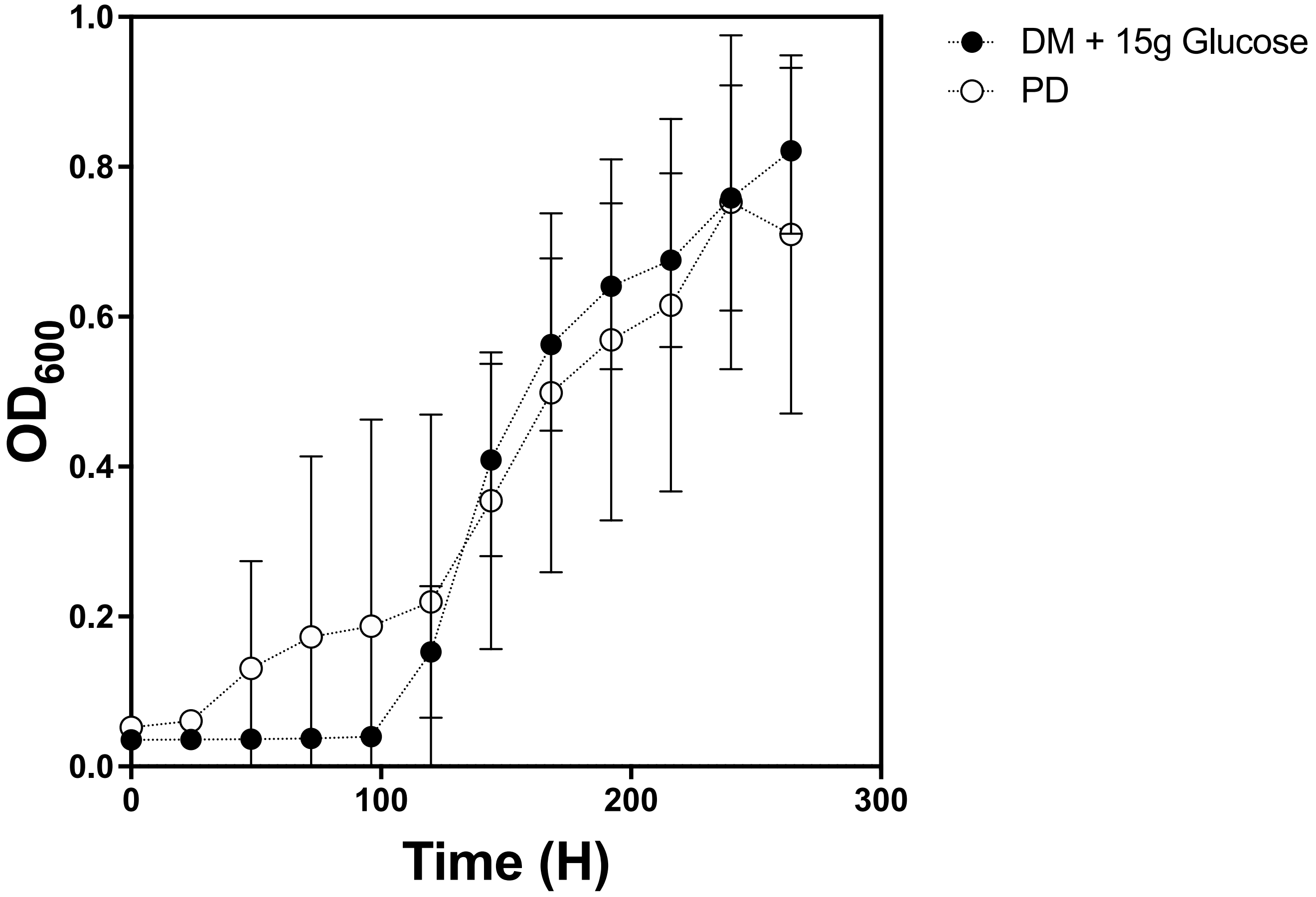


Figure 3.3. OD600 growth kinetics of *A. sarcoides* 64019. *A. sarcoides* 64019 were cultured with 200 µL potato dextrose broth and defined medium containing 15 g/L of glucose on a 96 well micro-titre. Cultures were incubated at 23 °C and 120 rpm for 11 days. Growth was measured daily and turbidity was measure with OD600. Error bar indicate standard error of deviation. (n = 30).

3.2.2 Screening for carbon source utilisation

Isotopic labelling with ^{13}C and ^{2}H isotopes is an integral technique for tracking assimilation and dissimilation pathways. Labelling achieves two objectives: Firstly, revealing nutrients that is able to assimilate labeled isotope and catabolised it to a metabolite. Thus by detecting a labelled metabolite, it is possible to show a metabolite is of a biological origin. Secondly, it allows for the detection of relevant intermediates and the potential enzymes involved in the pathways. The use of isotopic labelling, however, can be costly thus it can be limits usage.

In both solid and liquid culture defined media containing glucose have the highest growth rate and highest cell density relative to cultures grown on defined media with other carbon sources Figure 3.4A and 3.4B. There are significant differences in growth rates of liquid and solid culturing in relation to different carbon sources. For example, liquid cultures grown with glutamate ranked third in cell density when compared to other composition and it was able to achieve a similar cell density to defined media with glucose at the end of the experiment (Figure 3.4A). Conversely, in solid cultures, we observe signs of growth but glutamate ranked fifth among other solid cultures. We observed severely stunted growth and, in comparison, it achieved 10% growth diameter to solid cultures containing glucose. The opposite is true with solid cultures grown on cellulose, whereby liquid cultures grown in cellulose than those grown on solid medium. Solid cultures grown with tetradecanoic acids exhibit reduced or no growth relative to cultures grown with glucose. This may be a result of low dispersal of free fatty acids or oxidation of the fatty acids over the duration of incubation, thereby reducing the availability of fatty acids to the fungal culture. Interestingly, we were able to observe *A. sarcoides* utilising linear alkanes but not cyclic alkanes. Moreover, cultures grown on tetradecane exhibit spindly hyphae formation that are less dense to fungal hyphae cultured with glucose.

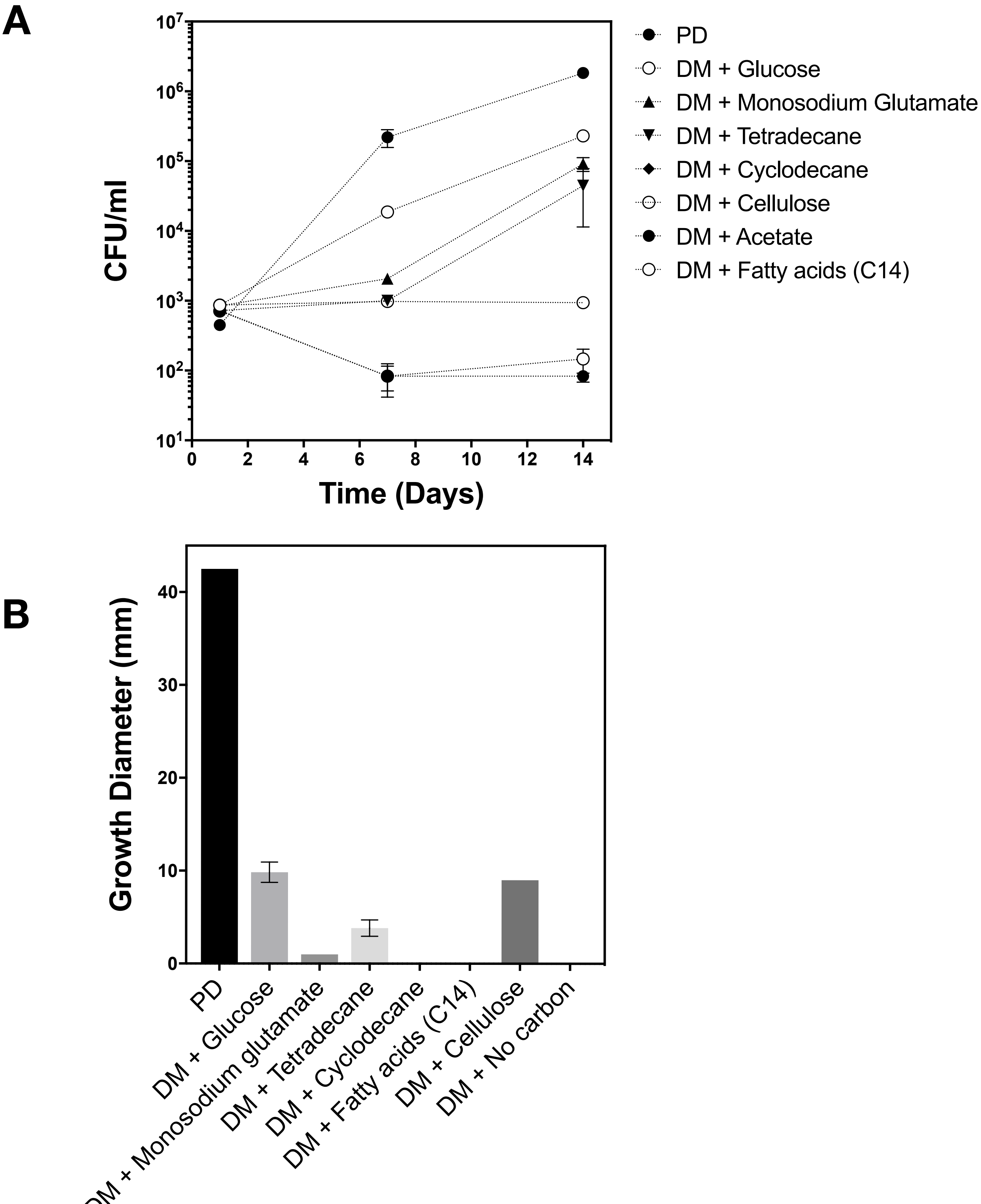

A
CFU/ml
10^1
10^2
10^3
10^4
10^5
10^6
10^7
0
2
4
6
8
10
12
14
Time (Days)
● PD
○ DM + Glucose
▲ DM + Monosodium Glutamate
▼ DM + Tetradecane
◆ DM + Cyclodecane
○ DM + Cellulose
● DM + Acetate
○ DM + Fatty acids (C14)
B
Growth Diameter (mm)
0
10
20
30
40
PD
DM + Glucose
DM + Monosodium glutamate
DM + Tetradecane
DM + Cyclodecane
DM + Fatty acids (C14)
DM + Cellulose
DM + No carbon

Figure 3.4 Growth of *A sarcoides* 64019 on different carbon sources 3A. Growth of *A. sarcoides* 64019 on potato dextrose broth and defined medium containing 15 g/L of different sole carbon source. Samples of 100 µL were taken from cultures and plated on potato dextrose agar. Colonies number were measured after 5 days and expressed as colony forming units CFU. Error bar indicate standard error of mean. (n = 3). Figure 3B. Growth of *A. sarcoides* 64019 potato dextrose agar and defined medium supplemented with different sole carbon sources. Growth was measured at the end of the experiment at day 30 and to measure diameter of the culture. PDA cultures achieved maximum plate coverage at the end of incubation. *n* = 3. Error bar indicate standard error of deviation. (n = 30).

3.2.3 Culturing with microtitre plates

Measuring growth by colony forming unit is precise. However, it is time and resource intensive. Additionally, it does not scale well when attempting to investigate different treatments. In contrast, measuring growth kinetics by optical density (OD) is a quick and simple procedure. Combined with the space-saving format of 96 well plates, we were able to overcome limitations relating to time, space and low replicate numbers when measuring the growth of *A. sarcoides*. For reliable growth measurement of a microbial culture, it needs to satisfy two criteria. First, the culture medium must be clear and not interfere with turbidity measurements. This is satisfied by using defined media. Second, the microbial cultures must be uniformly suspended in the medium. Initial attempts to measure turbidity with OD_{600nm} were confounded by agglomeration. The lack of homogeneity was thought to be associated with the clumping of filamentous fungal. Fungal agglomeration is a common occurrence for filamentous fungal in liquid cultures (Gibbs et al., 2000). This leads to unrepresentative sampling and bias in turbidity measurements.

Agglomeration must be overcome so that we can effectively measure growth by turbidity. It is thought that the supplementation of polysorbates, a non-ionic surfactant and an emulsifier, to a culture medium can reduce the adhesion of mycelium to itself and to the surface of the growth vessel (Meletiadis et al., 2001). Therefore, the addition of polysorbates can increase the homogeneity of fungal cultures for more precise turbidity measurements. Furthermore, polysorbates are able to solubilise hydrocarbon compounds and facilitate increase in alkane permeability across the cell membrane (Chan et al., 1991). We tested the effects of polysorbates on *A. sarcoides* incubation by supplementing cultures with 0.1% (v/v %) or 0.01% (v/v %) of either Tween 80 or Tween 20. Treatments with polysorbate were observed to have higher final OD and a reduction of standard error and standard deviation than cultures without Tween (Figure 3.5). Factors such as concentration of Tween and type of Tween have no measurable impact on OD (Figure 3.5). In addition, defined media supplemented with Tween 80 did not demonstrate any growth. This leads to the conclusion that *A. sarcoides* was unable to incorporate Tween 80 as a carbon source. The main effect of Tween is, therefore, to improve the homogeneity of *A. sarcoides* cultures and provide a more representative OD_{600} reading.

By having a more precise measurement of OD_{600}, we were able to downscale culture volume. We were able to fit *A. sarcoides* cultures on to a 96 well microtitre plate with a culture volume of 200 µL. Growth kinetics such as mean lag time and growth rate in 96 well plates (Figure 3.3) were comparable to growth curves in 50 mL liquid cultures grown in conical flasks and measured with CFUs (Figure 3.1). Thus we conclude that cultures at 200 µL measured at OD_{600nm} were representative of cultures grown at mL scale measured with CFUs. It is worth noting that local plate effects (well columns and well rows) did not contribute to growth associated biases. Although, wells on the edge of plates were observed to have increases in variation and a lower OD, this was thought to be associated with evaporation occurring at a higher rater on the edges of a microtitre plate. Thus, in all experiments, these wells were not used for data gathering purposes and contained water with volume equal to cultures.

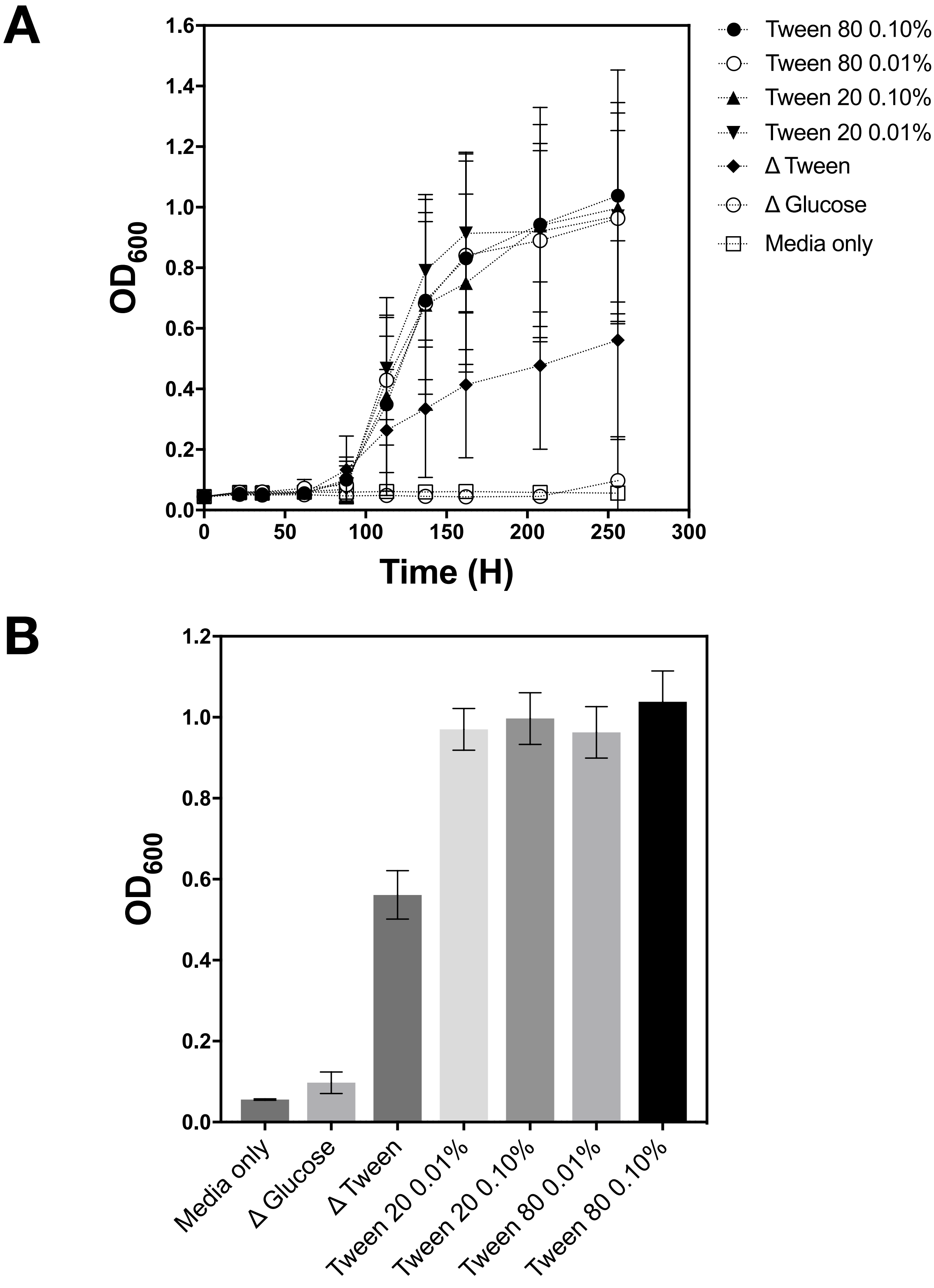

A
OD600
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
50
100
150
200
250
300
Time (H)
Tween 80 0.10%
Tween 80 0.01%
Tween 20 0.10%
Tween 20 0.01%
Δ Tween
Δ Glucose
Media only
B
OD600
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Media only
Δ Glucose
Δ Tween
Tween 20 0.01%
Tween 20 0.10%
Tween 80 0.01%
Tween 80 0.10%

Figure 3.5. OD_{600} growth kinetics of *A. sarcoides* 64019. A) Growth kinetics of *A. sarcoides.* B) End of experiment sampling. *A. sarcoides* 64019 were cultured with 200 µL defined medium containing 15 g/L of glucose and supplemented with different type and amount of Tween on a 96 well micro-titre. Cultures were incubated at 23 °C and 120 rpm for 11 days. Controls with no tween, no fungi and no glucose was also included. Growth was measured daily and to record turbidity by OD_{600}. Error bar indicate standard error of deviation. (n = 30).

3.2.4 Degradation of alkanes

Having established a high-throughput, small scale protocol for the culturing of *A. sarcoides*, we investigated the possibility of *A. sarcoides* as an alkane degrader. Organisms that occupy a saprophytic niche are required to be indiscriminate when assimilating nutrients. Plant wax composition often contains significant amounts of long chain alkanes, alcohols, aldehydes and ketones (Bernard et al., 2012b) and degradation of alkanes is a common trait among microorganisms (Van Beilen et al., 2007).

Results indicate that *A. sarcoides* is able to grow in defined media when supplemented with tetradecane as a sole carbon source but not with supplemented cyclodecane (Figure 3.6). It is important to note that our experimental setup does not directly measure the levels of alkanes and alkane degradation was only inferred from the growth of *A. sarcoides*. The rationale for this assay is that if an organism can proliferate on a sole carbon source, then it must be able to degrade and assimilate that carbon source for growth. Interestingly, cultures supplemented with tetradecane were comparable to cultures grown with glucose as a sole carbon source. Furthermore, we did not observe a dose response when the concentration of tetradecane was doubled. In cultures with higher levels of tetradecane (30 g/L), we observed an increase in the lag phase and reduced growth density relative to cultures subjected to lower tetradecane (15 g/L). Cultures containing higher levels of tetradecane may have been hampered by poor aeration due to the formation of hydrophobic layer on the surface of the medium. Interestingly, cultures incubated with cyclodecane as a sole carbon source were unable to grow. At the end of the experiment, cultures from this treatment and cultures with no carbon source were platted on PDA to check for viability. No viable colony was observed from cyclodecane treatment and, in contrast, cultures with no carbon source remain viable. This indicate that *A. sarcoides* may be toxic to cyclic alkanes at levels around 15 g/L. Thus, it can be assume that *A. sarcoides* can degrade linear alkanes but not cyclic alkanes for energy production.

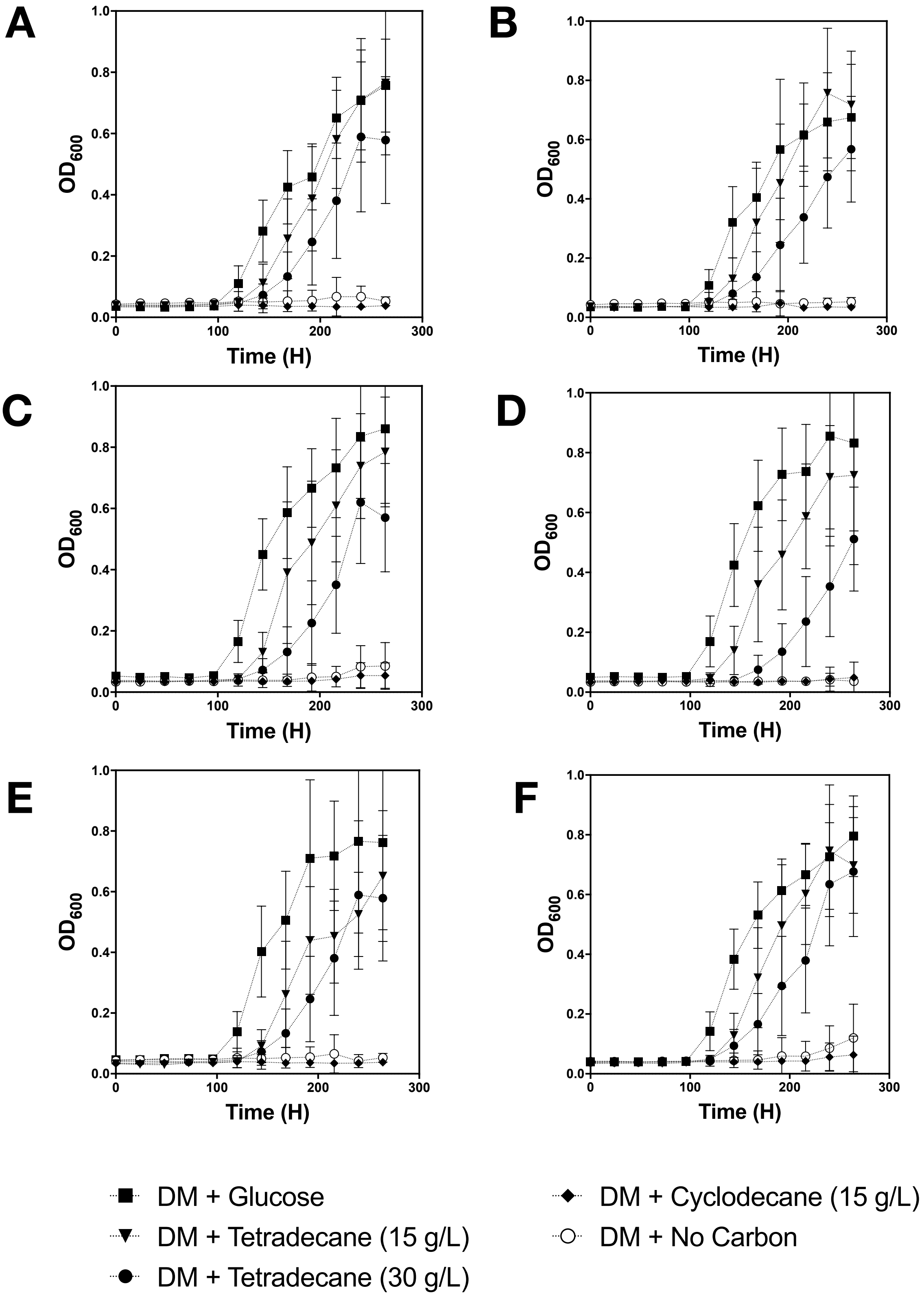

A
B
C
D
E
F
OD600
Time (H)
0.0
0.2
0.4
0.6
0.8
1.0
0
100
200
300
■ DM + Glucose
▼ DM + Tetradecane (15 g/L)
● DM + Tetradecane (30 g/L)
◆ DM + Cyclodecane (15 g/L)
○ DM + No Carbon

Figure 3.6. Alkane degradation with *A. sarcoides* 64019 in 96 well plates. Cultures were grown in defined medium supplemented with with 15 g/L and 30 g/L of Tetradecane, 15 g/L Cyclodecane, 20 g/L Glucose. Controls with no carbon source were also included. Cultures were incubated at 23 °C and 120 rpm for 11 days. Growth was measured daily and to record turbidity by OD_{600}. The designation represents the following isolate: A) 170.56, B) 171.56, C) 246.80, D)309.71, E) 44013, and F) 64019. Error bar indicate standard error of mean (n = 24).

3.2.5 Storage and work cultures

Key to the project success is to develop a reliable protocol for storing and maintaining *A. sarcoides*. The aim is to have a protocol with minimal changes in phenotype and acceptable recovery. Long term preservation of cultures aims to minimise activity while preserving viability. Activity within a cell can be suppressed when water is removed. Cellular activity is hampered on a molecular level in a multifaceted manner. An example of this are conformational changes in proteins, which affect protein function, and reduce solvation of biomolecules, thus yielding a reduction in bioavailability. A decrease in temperature also plays a key role in suppressing cellular activity. Again, the reduction of activity is on a molecular level; for example, reduced interactions between biomolecules and an increase in the activation energy required for enzymatic reactions. Reductions in water and temperature are favoured as the reduction in activity can be restored by rehydration and a return to favourable incubation temperature. These two factors are synergistic with regards to long term storage of microorganisms. Effectively, microbial cultures remain in long periods of inactivity and can be restored without removing viability. Considered as the gold standard, the sealing of lyophilised cultures have been the method of choice in many commercial culture collections. Lyophilised fungal cultures are able to maintain viability for an extended period (over 10 years), cannot be contaminated during storage, and require little space (Haynes et al., , 1955). Other storage methods also leverage dehydration and temperature. We stored and evaluated the effectiveness of each method.

Fungal cultures stored on solid medium and agar slants were sealed with parafilm. These fungal cultures were kept at 4 °C and examined periodically. With cultures in solid medium, we observed typical pigmentation at day 7 and aerial hyphae after day 14. Beyond day 60, we observe the loss of aerial hyphae. Cultures in agar slants were able to colonises the surface of the slants. No change in phenotype was observed. Subculturing slants indicate little variations in the rate of growth. However phenotypes were not consistent (e.g. lack of aerial hyphae and loss of colouration). However, after 12 months, slant cultures were no longer viable when plated on PDA.

The simplest method for mid term storage is suspending fungal material in sterile distilled water. Typically, this was used to store and suspend paper-harbouring mycelia at 4 °C and stocks were remade when depleted approximately every 6 to 12 months. Throughout three years, cultures from long-term stocks were revived to replenished microbial stocks. It is thought the lack of nutrition and the low temperature decreases cellular activity. When revived, fungal cultures were viable. Additionally, there were no observable phenotypic differences and rate of recovery for *A. sarcoides*.

Milk/glycerol mix was also used as a storage method. It is thought that both milk and glycerol act as cyroprotectant and decrease the damage caused by ice crystallisation. With this method we were able to revive all of our isolates successfully on a yearly basis. However, we noted, high variation in reviving time with 3 to 7 days. Moreover, we did not observe any phenotypic changes when revived.

Of all the methods tested, filter paper based method were the most similar to the lyophilisation method as it incorporates dehydration and low temperature. Sterilised filter paper were placed on PDA plates containing single colony of *A. sarcoides* isolate. This allows for the colonisation of the paper by the isolate, a process that took 21 days. When observed, mycelia on the filter paper developed pigmentation. The paper containing mycelia is then dehydrated in a bell jar containing dehydrated silica gels for over 14 days and were kept in plastic tubes at -80 °C. To revive an isolate, a single paper medium was placed face down on a PDA plate. Again, cultures were revived when necessary to replenish stocks. During the reviving phase, recovery required longer incubation time than the other methods thus affects the overall growth rate of the culture. All culture from this method showed consistency in phenotype and recovery rate.

Table 3.1. Different long term storage method and their efficacies. Cultures of *A. sarcoides* were stored with different methods over the course of 3 years and were checked for viability. All storage methods were checked yearly or the need to replenish existing stocks. Viability was performed by reviving mycelium stocks on a potato dextrose agar. Time taken to revive and phenotypic morphology were noted.

Storage Method	Viability	Time required for revive	Change in phenotype
PDA plates	3 months	1 day	Loss of aerial hyphae during storage
PDA slants	12 months	1 day	Production of pigments during storage
Water suspension	12 months	1 day	None noted
Milk and glycerol	Still viable	3 - 7 days	None noted
Lyophilising Filter paper	Still viable	3 days	None noted

3.3 Discussions

3.3.1 Husbandry of A. sarcoides

We were able to culture six different isolates of *A. sarcoides* within a rich medium and a chemically defined medium. Moreover, we were able to culture at different scales and to produce representative growth kinetics with two different type of measurement. When comparing growth kinetics of the two media, we note comparable growth kinetics and observed no differences in the phenotype and the morphological features of each isolate.

Viable growth in our defined medium indicates that *A. sarcoides* isolates are prototrophic and not auxotrophic. Thus, the isolates contain the required pathways to assimilate and synthesise key nutrients such as vitamins and amino acids from inorganic compounds encountered within the defined medium. Growth kinetics (Figure 3.3) suggest that the extended lag phase is necessary to activate necessary catabolic pathways. The assimilation of inorganic nitrogen, sulphur, and phosphorus requires specific pathways for uptake and catabolic pathways. For fungi to utilise inorganic nitrogen sources, the expression of specific permeases and reductases are required (Keller et al., 1997). Excessive ammonium is known to be toxic to fungi unless it is incorporated into non-toxic organic compounds (Temple et al., 1998). Radioisotope labelling in mycorrhizal fungi indicate that inorganic nitrogen is assimilated by pathways relating to the biosynthesis of asparagine, arginine, glutamine, and glutamate (Jin et al., 2005). Sulphur uptake in fungi is achieved by sulphate permease and S/H$^+$ co-transport. The uptake of sulphur is energy dependent (Marzluf, 1970). This intracellular sulphate is converted to ammonium persulphate by ATP sulphurylase. Further reduction of ammonium persulphate eventually leads to the generation of the amino acids cysteine and methionine. The cysteine is then incorporated into the fungi. Phosphate uptake is also energy dependent. It has been demonstrated that a P/H$^+$ co-transport is required (Bieleski et al., 1983). This is thought to overcome negative membrane polarisation due to imbalances in the electrochemical gradient in phosphate uptake (Bieleski et al., 1983). The phosphate ion directly interacts with intracellular processes in the form of nucleotide polyphosphate.

Furthermore, we were able to conduct high-throughput phenotyping of *A. sarcoides*. Firstly, by incorporating an automated liquid handling robot to execute different compositions of our defined medium. Secondly, with 96 well plates and turbidity measurements, we have a standardised method that is scalable for phenotyping and measuring growth. Finally, by reducing culture volumes we able to measure a higher number of samples concurrently.

The development of pink to dark burgundy pigment is consistent with descriptions from field observations and morphological studies of *A. sarcoides* fruiting bodies (Jordan, 2004). This coloured pigment is thought to be a bioactive secretion known as Ascocorynin, a terpenoid antibiotic compound active against gram-positive bacteria (Quack et al., 1980). Although it also reported that *A. sarcoides* in its natural habitat is capable of producing a smooth gelatinous purple fruiting body, this was never observed in any cultures.The strong odour release by the fungal culture lends support to the production of many volatile organic compounds reported previously in other studies (Stinson et al., 2003, Griffin et al., 2010, Strobel et al., 2010b).

Finally, we were successful in storing *A. sarcoides* on a long term basis with minimal phenotypic drift. The preferred method for medium term storage was suspending mycelium material in sterile distilled water as it is simple to execute and mycelia stored this way remain viable for a long time. The preferred method for long term storage was lyophilising mycelia on a filter paper. This method proved to be the most reliable and predictable for reviving the cultures. We note that we did not observe loss of pigmentation as reported in past study (Strobel et al., 2010b). In their case, repetitive sub-culturing eventually led to changes in the phenotype of *A. sarcoides* AV-70

3.3.2 Culturing on different carbon sources

It is unsurprising that an endophytic saprotrophic organism has pathways that are able to assimilate cellulose. This observation was also reported by previous work carried out on *A. sarcoides* (Griffin et al., 2010, Gianoulis et al., 2012, Mallette et al., 2014). This indicates that *A. sarcoides* has the required machinery for assimilating cellulose. Curiously, there are significant differences in growth when we compare liquid and solid cultures grown with cellulose. This is an odd observation as the bioavailability of cellulose should be higher in liquid medium. Therefore there is an unknown interaction that is enabling growth in solid medium or inhibition of growth in liquid medium. It may be that in solid cultures *A. sarcoides* is able to change its extracellular environment more effectively for the assimilation of cellobiose than in liquid cultures. *A. sarcoides* has been suggested as a biotechnological organism to produce hydrocarbon fuel by the conversion of cellulosic biomass (Griffin et al., 2010, Strobel et al., 2010b, Mallette et al., 2012). However, our result here disagrees with this assessment. Cellulose is converted to fermentable glucose in three enzymatic steps. First, endoglucanases act on amorphous cellulose and cleave between cellulose polymers. Second, cellobiohydrolases act on single polymer cellulose by cleaving 1-4 glycosidic bonds to produce tetrameric or dimeric cellobiose. The 1-4 glycosidic bonds in tetrameric and dimeric cellobiose are further reduced by beta-glucosidase resulting in monomeric glucose (Mohanram et al., 2013, Singhania et al., 2013). As beta-glucosidase conversion is the final rate-limiting step, growth with cellobiose infers the performance of the enzyme. When *A. sarcoides* is cultured exclusively with cellobiose, we see poor growth performance in relation to culturing with glucose. Moreover, agricultural cellulosic biomass includes impurities that may hamper cultivation of *A. sarcoides*. Furthermore, most secondary metabolite production occurs after the initial growth phase. Thus for *A. sarcoides* to be a tractable biofuel organism, it requires improve cellulase activity.

The inability to grow on fatty acids indicate that it is unable to uptake fatty acids. However, acetic acids were readily metabolised by *A. sarcoides*. Previous study indicates the ability of *Aspergillus* to grow on fatty acids of C14, C16, and C18 (Radwan and Soliman, 1988). They reported a similar growth performance to glucose when fatty acids were used as the sole carbon and energy source (Radwan et al., 1988). We proposed that the limitation in growth is due to the efficiency in emulsification of fatty acids. If this is true, we would be able to predict performances with shorter fatty acids, as they are more soluble than longer chain fatty acid.

Furthermore, glutamate in the form of monosodium glutamate was readily metabolised by *A. sarcoides*. As glutamate is a fairly polar compound, cells require specialised uptake proteins to facilitate the uptake of glutamate (Kinghorn et al., 1973). Furthermore, glutamate regulation is dependent on ammonium regulation in fungal cells (Kinghorn et al., 1973). Moreover, glutamate can act as a carbon source when the amino group is reduced to urea.

3.3.3 Degradation of linear and cyclic alkanes

From our results we were able to inferred *A. sarcoides* has the ability to degrade linear alkanes. Cyclodecane, however, was observed to be toxic to *A. sarcoides*. The ability to degrade alkane is fairly ubiquitous among microbial life. *A. sarcoides* is an endophyte and saprophyte (Roll-hansen et al., 1979, Pavlidis et al., 2005). Alkanes are commonly found as part of the waxy layer found in plants (Bourdenx et al., 2011, Bernard et al., 2012a). These wax layer are formed from 70% very long chain alkanes (C27-C33) which act as a barrier against water loss and may be involved in the defence against endophytes (Chen et al., 2003). Similarly, *Beauveria bassiana,* an entomopathological fungi that is able to degrade alkanes on insect's wax layer as part of its mode of action for pathogenesis. It is thought that *B. bassiana* is able to oxidised alkanes by through cytochrome P450 oxidative enzymes to a corresponding aldehyde (Pedrini et al., 2006). Alkane is taken up, oxidised, and further catabolised by the TCA cycle. To the best of our knowledge, this is the first description of linear alkane degradation and biosynthesis was observed and reported within a single fungus.

There are also pathways such as the caprolactam degradation pathway (KEGG pathway 00930) that indicate the degradation of cyclohexane. Observations to date indicate that cyclic alkanes are oxidised to a cyclic alcohol and cyclic ketone and is further degraded and assimilated by caprolactam pathways (Stirling et al., 1977, Lee et al., 2008). To date, reports of cyclic alkane degradation are limited to prokaryotic organisms under anaerobic conditions (Musat et al., 2010, Varjani, 2017). Cylic alkanes are recalcitrant to degradation compared to similar length linear alkanes (Lee et al., 2008). This is due to the non-polarity of the compound. Degradation of cyclohexane by *Nocardia* led to an interesting observation of "intracytoplasmic membrane structures" not present in cultures not supplemented with cyclohexane, thus alluding to potential structural function (Stirling et al., 1977).

Further investigations are required to confirm *A. sarcoides* as an alkane degrader. Definitive proof of hydrocarbon degradation will require direct quantitative measurements of cultures incubated with alkanes and the possibility of measuring the direct product of alkane degradation.

3.4 Conclusions

A. sarcoides is a slow growing filamentous fungus with a mean time to stationary phase of 10-12 days in PD and DM. Growth characteristics were observed to be similar across all isolates. Due to the slow growing nature of the fungus and time consuming measuring growth by CFU, the culturing and measuring method was optimised for 96 well plates with OD_{600}. By using Tween 80 at 0.01%, we were able to increase the effectiveness of OD_{600} measurements to better represent the growth kinetics of *A. sarcoides*. With defined medium, it is possible to change the composition of the medium to test the fungi's ability to assimilate different carbon sources. While testing this, the results indicate that *A. sarcoides* was able to degrade and assimilate linear alkanes but not cyclic alkanes. Data generated here will inform experimental design choice in later experiments. The result here indicated the successful cultivation and manipulation of the fungi that will provide a foundational skill for future work.

CHAPTER 4 BUILDING A FUNCTIONAL GENOME FOR SIX ISOLATES OF *A. SARCOIDES*

4.1.0 Introduction

Having established the fundamental husbandry and microbiology for *A. sarcoides* (Chapter 3), we next wished to uncover the underpinning genetics for the reported alkane biosynthesis (Griffin et al., 2010). Current methodologies allow relatively cheap and easy access to whole genome sequencing. We therefore wished to undertake a subtractive genomics approach to aid pathway discovery. Here we describe the generation and analysis of the genomes of all six isolates.

4.1.1 Next generation sequencing

Contemporary next generation sequencing (NGS) has become so affordable that the current bottleneck has moved to the functional annotation of sequenced genomes. NGS methodologies, such as Illumina-Seq, distinguish individual base in a DNA fragment by using fluorescent dye chemistry to detect base changes in a sequence. This is achieved with cycles of complementary base binding of the target DNA sequence with fluorescent tagged nucleotide (Meyer et al., 2010). By limiting to one nucleotide binding, a distinct fluorescent signal is emitted for a given base. The Illumina flow cells facilitate this dye chemistry in a highly parallel manner during sequencing. The fluorescent signal is measured by an optical system to determine emission wavelength and emission intensity, which can be use to determined the base call and the amount of cycles also indicates the length of the target DNA sequence.

Depending on the applications, these methods are able to produce high number (> 4 million reads), short read libraries between 50 bp to 300 bp in read length. One of the advantages of high read number is the ability to provide redundancies for conflicting base calls, miscalled base events and error corrections, thus leading to high fidelity sequencing (Meyer et al., 2010). One of the issues with Illumina sequencing is the non-uniformity in coverage depth. This is due to the random nature of sequencing, a result of random fragmentation during library preparation) and short read length. Intrinsically, short reads are less likely to overlap with other reads, leading to the formation of gaps in the coverage of a genome. Illumina sequencing overcomes this issue with an overwhelming amount of reads. However, a high number of short reads can be computationally difficult due to the quantity of the data, size of raw read libraries, and the complexity associated with processing high read numbers (Aird et al., 2011). To overcome this requires specialised algorithms and significant computing resources to conduct post-sequencing processing and assemble read libraries.

4.1.2 Post-sequencing processing

Post-sequencing processing is the first step in handling genome sequencing data. It involves trimming the read libraries to improve the quality of the reads. Tools such as Trimmomatic (Bolger et al., 2014) and Trim Galore! (Krueger, 2015) are able to process NGS Illumina reads. Trimming has two primary purpose. First, it can be used to remove artefact sequences found in reads. Secondly, it also removes low quality bases within a sequence. Trimming yields reads that are truncated but higher in quality. In NGS read libraries, reads often contain leftover oligos and primers. Moreover, sequences derived from Illumina-Seq characteristically decreases in quality at the end of reads. As such, trimming is essential; First, to remove excessive and low quality data thus leading to a more manageable read library. And second, to increase precision of called bases therefore the fidelity of a read library. Excessive trimming can lead to shorter reads with an impact on coverage. In projects focusing on high quality *de novo* genome assembly and annotation, trimming must be balanced against the removal of sequencing artefacts, accuracy of the libraries and the coverage of libraries.

4.1.3 Assembly of Illumina-seq libraries

Once quality control is performed, assemblers are use to construct a draft genome. In the absence of a high quality mappable reference genome, *de novo* assemblers such as SPAdes (Bankevich et al., 2012) construct genomes by aligning reads that overlap from a library. These aggregate reads to form a sequence of DNA known as a contig and these contigs are ordered to a scaffold level. Assemblers treat the read alignment problem as a string reconstruction problem. This abstraction operates by identifying overlapping sequences to produce a sequence of DNA (in this case, a string of characters). Abstracting also allows for the use of De Bruijn approaches to solve *k-mer* graphs (all possible substrings of a given sequence), which leads to contig construction by identifying overlapping sequences (Compeau et al., 2011). Other popular algorithms include the Overlap/Layout/Consensus (OLC), which relies on overlap graphs as a method for constructing contig sequences and greedy algorithm, which functions by using sequences with the highest overlapping scores to construct contig sequences (Li et al., 2012). Assembly algorithms are challenged and confounded by factors such as low quality, low read count and short reads of read libraries. The lack of quality and lack of redundancies can confound assemblers as they cannot resolve conflicting base calls. This manifests as false negative and false positive overlapping sequences and culminates to an inaccurate genome assembly. Short reads lack coverage for the identification of overlapping sequence and yield assemblies that are gapped and fragmented into a high number of contigs. Crucially, these gaps undermine annotation efforts to generate a functional genome.

4.1.4 Producing functional genomes

Ontology, in a biological context, represents formal biological semantics for describing a gene. Thus, identifying ontological terms is important in annotation pipeline as they provide a comparable description of genes within an annotated genome. Successful annotation of ontologies can yield biologically meaningful data such as the function of a gene or the assembly of *de novo* pathways. This evidence may be integrated with metabolome, proteome, and transcriptome studies to allow for an ‘omics’ level investigation to identify new pathways.

In bioinformatics there has been an increase usage of machine learning to analysed large and complex datasets. One such machine learning algorithm is Hidden Markov Model (HMM). In biology, HMMs are use to model unique profiles related to a particular genotype but not limited to intron/exon, gene finding, motif finding, single-nucleotide and polymorphisms (Stanke et al., 2004, Johnson et al., 2008, Finn et al., 2011). In all of these cases, HMMs were applied to recover hidden data states, such as genes, from an initial un-obvious dataset, such as DNA sequences. This is the inherent advantage of using HMMs. They describe and construct a full probabilistic model specific to a given dataset. Once an HMM-derived model is completed, it is applicable to similar datasets to recover data states by observing the data.

It is common for *de novo* annotation tools to handle draft genomes with minimal reference source. This can be overcome when these tools incorporate HMM approaches into their workflow. Tools such as BUSCO (Benchmarking Universal

Single-Copy Orthologs) leverage Augustus HMMs to match orthologs independent of evolutionarily distance (Simão et al., 2015). An HMM profile of the orthologs is built which incorporates the characteristics of the genes within the dataset. The HMM profile is then used to interrogate other datasets for distant orthologs. By using single copy orthologs, datasets such as genomes, transcripts and predicted proteins can be scored by matching them to a curated database of highly conserved orthologs. Being able to measure a dataset is crucial in *de novo* pipeline as it can be difficult to gauge the data fidelity derived from such pipelines. Furthermore, by recovering sequences of orthologs it also allows for phylogenetic comparisons of highly conserved orthologous genes.

HMMs are commonly used in *ab initio* gene prediction to recover genes (the hidden state) from genomes (complex unobvious dataset). In eukaryotic genomes, *ab initio* gene prediction without transcriptomic data is often confounded by gene structures (e.g. Introns, exons, and slice sites). MAKER2 gene annotation pipeline deploys HMMs to model the structure of genes of a specific genome for the purpose of gene prediction (Holt et al., 2011). An HMM can be train with multiple heterogenous evidences to generate a gene model. This can include high quality reference dataset from closely related organisms or experimentally derived transcriptomic and proteomic data. These evidence are incorporated into an HMM model and is retrain to be specific to a given genome. Thus HMM can be use to conduct an exhaustive and specific gene prediction.

Lastly, the PANTHER ontology classification system contains an HMM module to conduct a more comprehensive search for genes that return no significant hit with BLAST algorithms (Thomas et al., 2003, Mi et al., 2013). A weakness associated with the BLAST algorithm is its inability to recognise a gene's characteristics in terms of position-specific information (This excludes PSI-BLAST, which includes HMMs of genes in its pipeline). And when BLAST is use for annotating distant homologous gene, it lacks the sensitivity required to return an appropriate annotation. The lack of sensitivity is due to the BLAST algorithm penalising mismatches or gaps, with no additional weighting for matching conserved motifs. When annotating a gene sequence that is too divergent, the BLAST algorithm is no longer able to resolve the identity of the match. In PANTHER, by contrast, an HMM is built for queries with no significant BLAST hits. These HMMs look for features within a sequence which then match these to the PANTHER ontology database. This approach balances computational resources with sensitivity as BLAST can be efficiently leverages for identifying close orthologs while resource intensive HMMs can be used only when required to match distant orthologs. Thus, pipelines like PANTHER are crucial to assign ontology semantics to distant orthologs from *de novo* annotations.

4.1.5 Aims and Objectives

In this Chapter the sequencing and annotation of six *A. sarcoides* genomes will be presented and discussed. In particular this data will provide the foundation for the subtractive analysis described in Chapter 6. Here we describe:

- Acquisition of NGS data
- Post-processing of NGS data in the form of trimming
- Assembly into draft genomes
- Predict the presence of ORFs
- Annotate predicted ORFs
- Identify biologically interesting features such as GO, synteny and phylogeny

4.2 Results

The generation of 6 *de novo* genome is outlined in Figure 3.1. This details the general pipeline used for sequencing, read quality control, genome assembly, and gene prediction. It also details downstream analysis and data mining approaches to recover biologically relevant data.

4.2.1 Trimming read libraries

Next generation sequencing (NGS) is a DNA sequencing method that produces a sequence library containing high number and short read length. This is both highly accurate and time efficient manner of sequencing compared to contemporary sequencing methods such as Sanger sequencing and Nanopore. However, large frequency reads require stringent quality control. Current methods employ trimming to remove low quality bases and sequencing artefacts from reads. This ensures that low quality bases are trimmed and low quality reads are discarded, thereby improving the quality of downstream processes such as assembly and gene annotations.

Two rounds of sequencing were undertaken to sequence six isolates of *A. sarcoides*. These were cultured on PDA, as described in Section 2.1.2. The 100 bp included isolates 170.56, 171.56, 246.80 and 309.71. The genomic DNA was isolated with QIAGEN Genomic DNA kits and sequenced with HiSeq 2000 system. This sequencing run produced libraries that were 100 bp in length. The 250 bp libraries included isolates 309.71, 44013 and 64019. Genomic DNA and sequencing for this set was performed by MicrobesNG. Reads were analysed with FastQC to assess the quality and features of the read libraries. This checks for symptoms of a failed sequencing. In such an event, trimming would not be able to rescue the sequence library and resequencing would be required. Each set was perfomed on the same sequencing run and so each library has similar qualities. Trimming decisions was made and applied on a set by set basis rather than specific to a library. These decisions were fed into Trimmomatic to produce the trimmed read libraries.

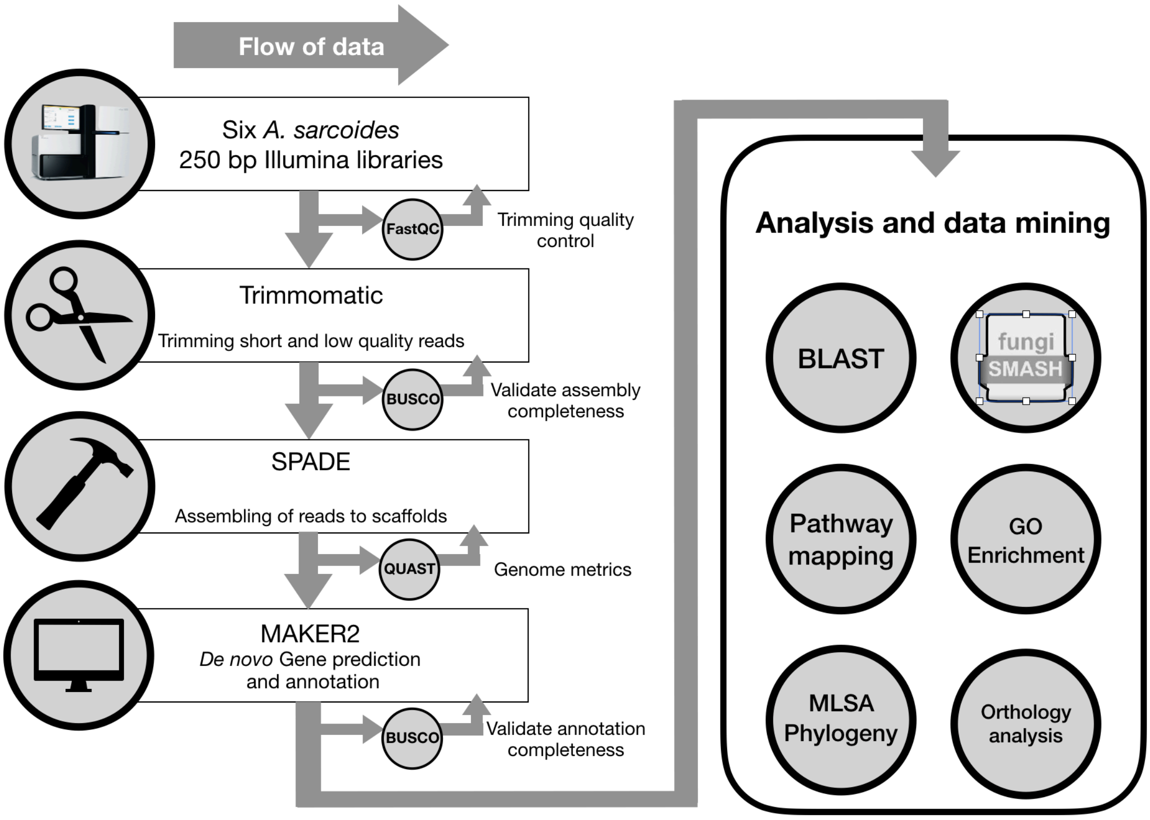


Figure 4.1 Overview of different bioinformatics pipelines and processes used for the production of six functional genomes of *A. sarcoides.*

Post-sequencing quality

In bioinformatics, base quality is measured according to Phred scale. The Phred scores is a metric that measures the accuracy of any given called base. Phred scores are expressed as a logarithmic probability of incorrectness. Therefore, the lower a Phred score the higher the probability a given base is incorrectly called and vice versa. In practice, Phred score allows us to estimate the error of a called base (Bokulich et al., 2013).

As a whole, reads derived from 100 bp libraries are of good quality for both the forward and reverse pair (Figure 4.2A, 4.2B and 4.3A, 4.3B). Although there are islands of low quality, these do not present a major issue for downstream trimming. The majority of reads from both forward and reverse libraries contain Phred scores of higher than 30 (Figure 4.2A and 4.3A) and indicate a normal distribution of quality within the dataset (Figure 4.2C and 4.3C). We did not observe abnormal quantity of N content in any of our reads; high N content in reads are undesirable as it signifies uncalled base and are functionally unknown bases.

Reads from the 250 bp libraries did not present any major sequencing failures. The forward libraries contain better sequence quality (Figure 4.5A and 4.5B) relative to the reverse libraries (Figure 4.6A and 4.6B). This can be observed in per base sequence quality plot (Figure 4.5A and 4.6A), where the drop in quality is more aggressive in the reverse libraries. Approximately 20% of all bases in the reverse libraries have a mean Phred score of under 20. Again, this can be observed in both libraries' per sequence quality score plots (Figure 4.5C and 4.6C), where the plot for the forward libraries peaked around Phred score of 38 while reverse libraries peaked at Phred score of 34. An issue like this manifests in higher base pair error rate and abnormal quantities of N content in reads. The difference in quality in the reverse libraries is attributed to the loss of quality in the sequencer flow cells. This can be observed in the Per Tile Sequence quality score plot (Figure 4.6B). It has to be noted that it is common for Illumina sequencers to suffer loss of quality at the end of their reads. Although this is a major issue, it can be resolved because the loss of quality is primarily located at the distal end of their reads. This mean that the low quality bases can be cropped from the reads without significantly reducing read length. Moreover, any resulting reads can be further subjected to quality trimming to further improve the quality of the reads.

Sequencing artefacts

Artefacts can be considered as sequences that are artificially introduced during sequencing. This can be deliberately introduced, in the case of adapters and oligos, as part of the sequencing workflow. Artefacts could also be highly erroneous and bias sequences do not reflect the fidelity of an organism's genome. It is not unusual for artefacts to manifest themselves in the final read libraries.

When checked for artefacts, the first 100 bp did not indicate the presence of any artefacts or anomalous sequence as indicated by Figure 4.1D and 4.2D. In contrast, in libraries from the 250 bp libraries, the per base sequence content indicate that 250 bp reads contained bias sequences, known as hexamers, in the first 15 bp of the reads (Figure 4.5D and 4.6D). Ideally, read sequences represent randomly sequenced portions of the genome thus such a plot should be parallel or converging. Deviations from the parallel or from divergence indicate the bias sequencing in a particular base pair. Such an anomaly can lead to incorrect bases in reads, resulting in poor assembly if enough bases are affected. Since there is no loss of quality in the 1-15 bp region (Figure 4.5A and 4.5B), this is likely to be caused by the presence of hexamers and adapters. A similar observation was made in all forward and reverse libraries within this set. Hypothetically, hexamer libraries contain a diverse set of sequences to avoid base pair biases. In practical terms, these diversities are still detectable during base detection. In such a case, adapter trimming and cropping of the first 15 bp can remove such biases from the read libraries. The advantage of trimming these sequences is to increase accuracy and efficiency during assembly.

Read libraries GC content

When plotting the GC content for each read sequence, it would be expected that the plot fits a bell shaped distribution. Deviation from a bell shaped distribution indicates bias in coverage during sequencing or potential inclusion of exogenous DNA fragment. All of our libraries indicates a normal distribution of GC content (Figure 4.2E, 4.3E, 4.5E and 4.6E). Moreover, the average GC content of all libraries is ~ 46%. An outlier to this was observed in the isolate library of 309.71, which had 43% GC content. Further observation also, indicate that the presence of a shoulder on the GC distribution plot in the library of isolate 309.71 in the first set, indicating the presence of contaminant DNA. As the nature of the contamination was unknown and the purpose of this endeavour is to find novel gene or genes, 309.71 was resequenced. This library was then removed from assembly and further analysis.

4.2.2 Quality control trimming

Read trimmings were undertaken to remove artefacts and improve the quality of reads. Read libraries containing distally located artefacts were trimmed by cropping the affected area. All reads were further subjected to Trimmomatic's MAXINFO quality trimming (Bolger et al., 2014). This algorithm balances both read quality and length. The trimming process discards portion of reads that do not surpass a user defined quality threshold and/or reads that are <60 bp. Short reads were discarded to improve assembly coverage. Trimming yielded 4 files: forward and reverse paired reads and forward and reverse unpaired reads. Surviving reads were then analysed by FastQC checklist until they achieve satisfactory criteria.

For the purpose of balancing the quality of sequence library and preserving read length, non critical FastQC warnings were disregarded in certain libraries. Criteria such as GC plots, base quality, removal of adapter presence and removal of kMer sequences were favoured for trimming. These factors have the largest effect on accurate and precise representation of an organism's genome and as such will have an impact on assembly and downstream annotation processes. Moreover, trimming was focused on removing SolexaPE adapters for the 100 bp libraries and NexteraPE adapters for the 250 bp libraries.

As the quality of the 100 bp libraries were of satisfactory quality, light trimming was undertaken to remove low quality spots. This was to maximise both base quality and coverage of the entire library. For the 250 bp libraries the loss of quality was mostly confined to distal portion of reads. Cropping removed these low quality distal regions. The reads were also subjected to further stringent distal quality trimming. A less stringent MAXINFO trimming was applied to discard and crop any low quality reads. This was to improve the quality of medially located base pairs.

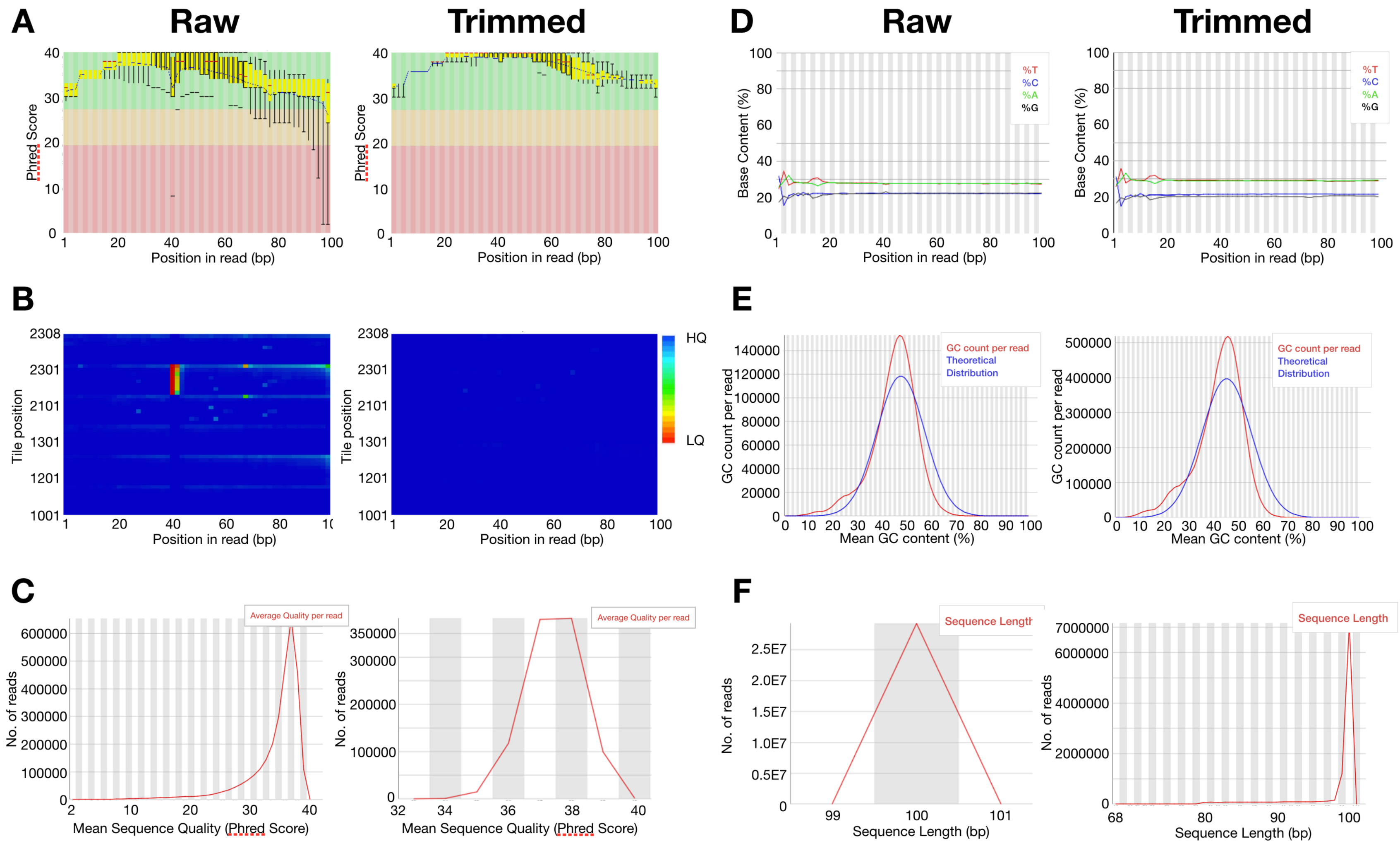


Figure 4.2. FastQC analysis of 170.56's 100 bp reverse Ilumina library. This library is typical in characteristic to the rest of the 100 bp libraries. Raw read library and its corresponding trimmed library counter part were analysed for quality. A) Average quality by our base position B) Average quality by tile positioning. HQ = high quality and LQ = Low quality C) Average quality per read. D) Base content distribution by per base location. E) Theoretical Gaussian GC content distribution and observed GC content of a library. F) Number of reads by sequence length.

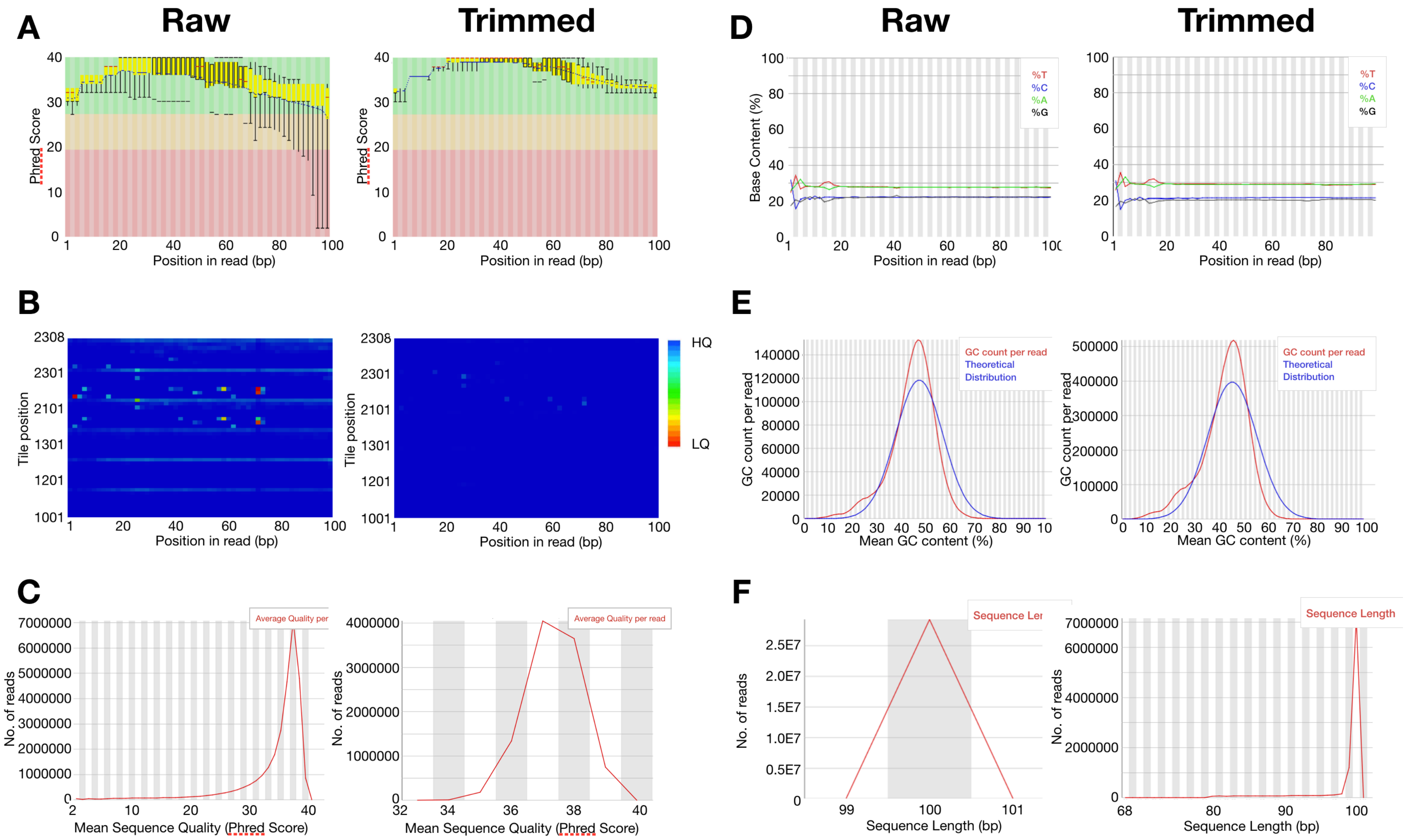


Figure 4.3. FastQC analysis of 170.56's 100 bp reverse paired library. This library is typical in characteristic to the rest of the 100 bp libraries. Raw read library and its corresponding trimmed library counter part were analysed for quality. A) Average quality by our base position B) Average quality by tile positioning. HQ = high quality and LQ = Low quality C) Average quality per read. D) Base content distribution by per base location. E) Theoretical Gaussian GC content distribution and observed GC content of a library. F) Number of reads by sequence length.

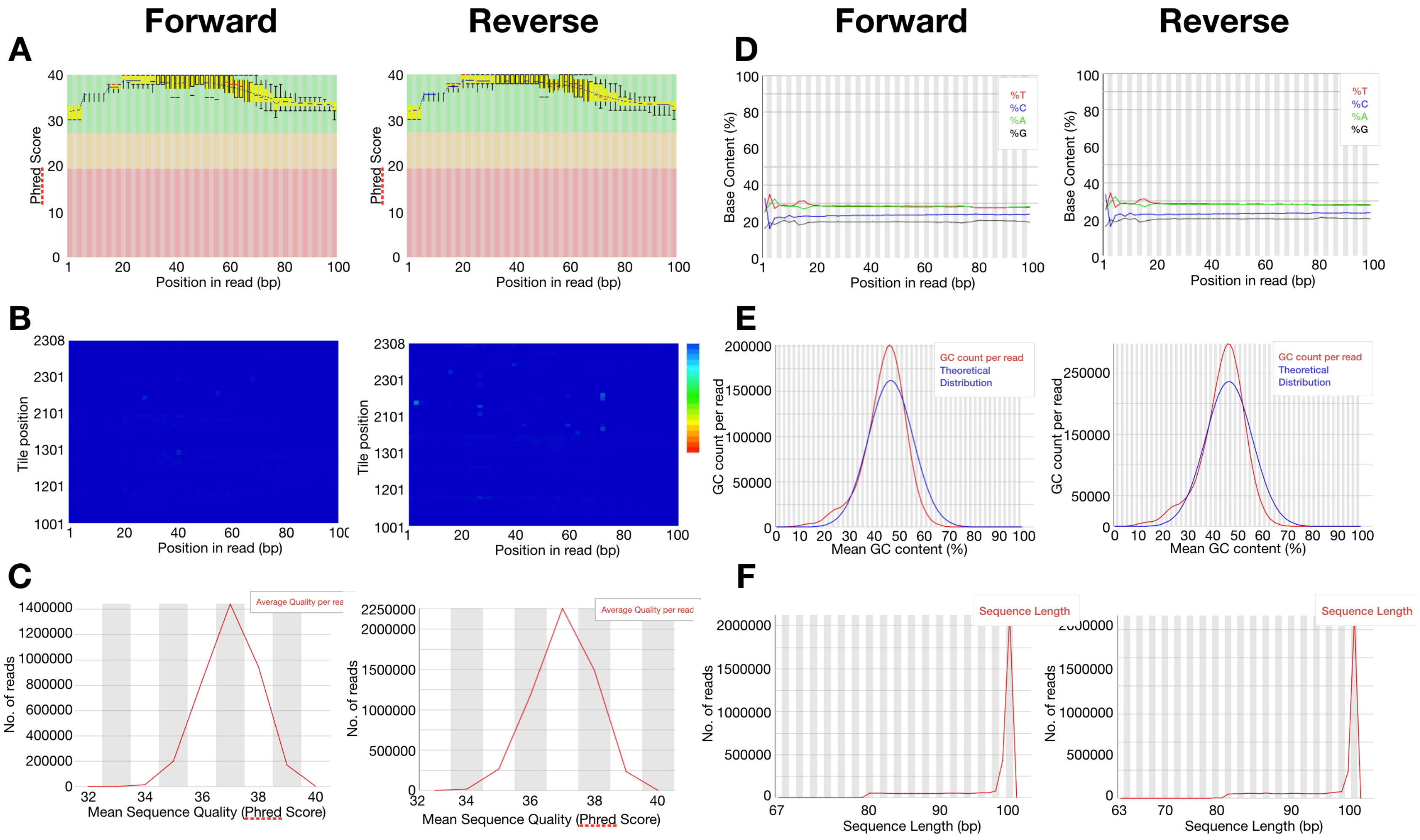


Figure 4.4. FastQC analysis of 170.56's 100 bp surviving post trimmed Ilumina library. This library is typical in characteristic to the rest of the 100 bp libraries. Raw read library and its corresponding trimmed library counter part were analysed for quality. A) Average quality by our base position B) Average quality by tile positioning. HQ = high quality and LQ = Low quality C) Average quality per read. D) Base content distribution by per base location. E) Theoretical Gaussian GC content distribution and observed GC content of a library. F) Number of reads by sequence length.

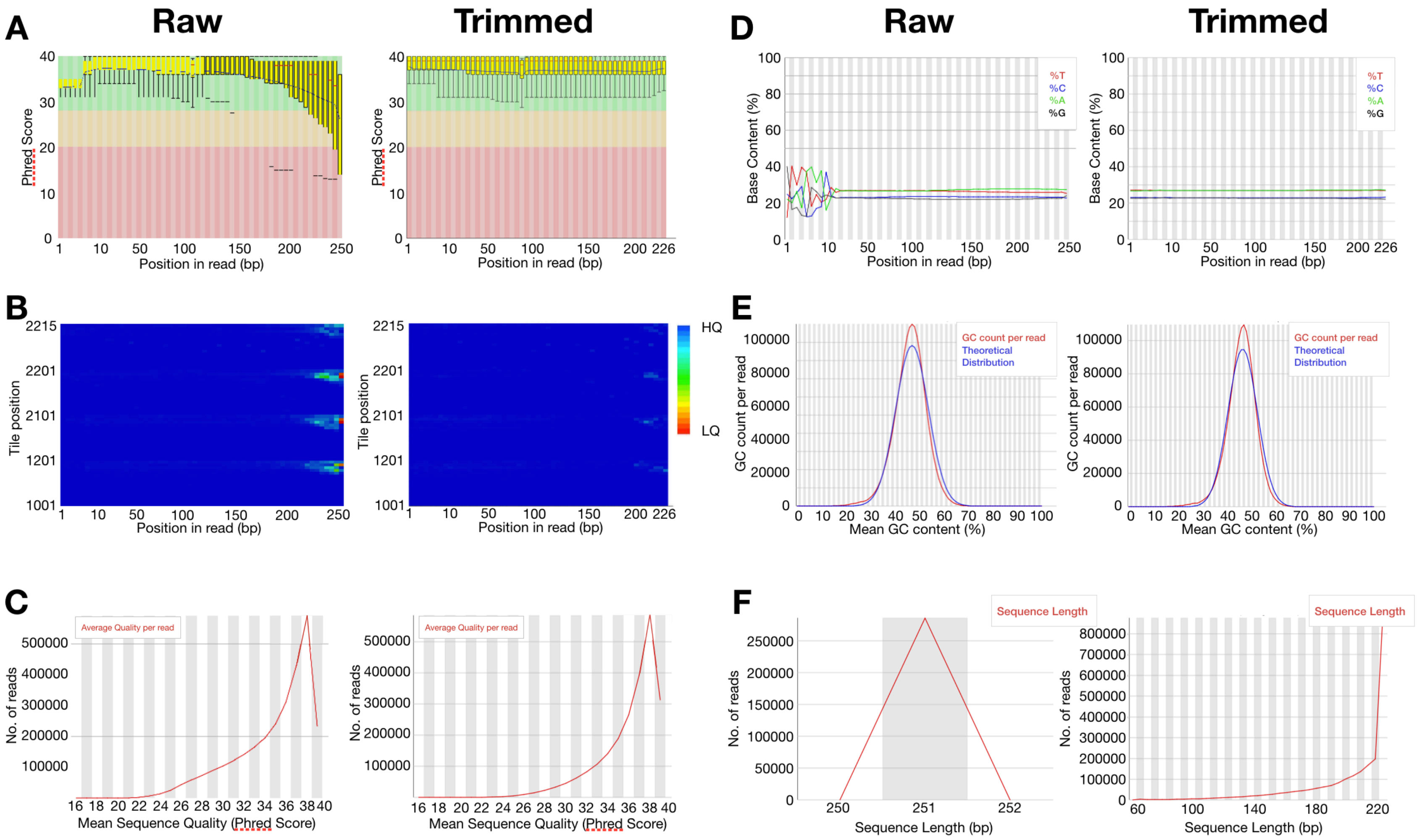


Figure 4.5. FastQC analysis of 309.76's 250 bp forward Ilumina library. This library is typical in characteristic to the rest of the 100 bp libraries. Raw read library and its corresponding trimmed library counter part were analysed for quality. A) Average quality by our base position B) Average quality by tile positioning. HQ = high quality and LQ = Low quality C) Average quality per read. D) Base content distribution by per base location. E) Theoretical Gaussian GC content distribution and observed GC content of a library. F) Number of reads by sequence length.

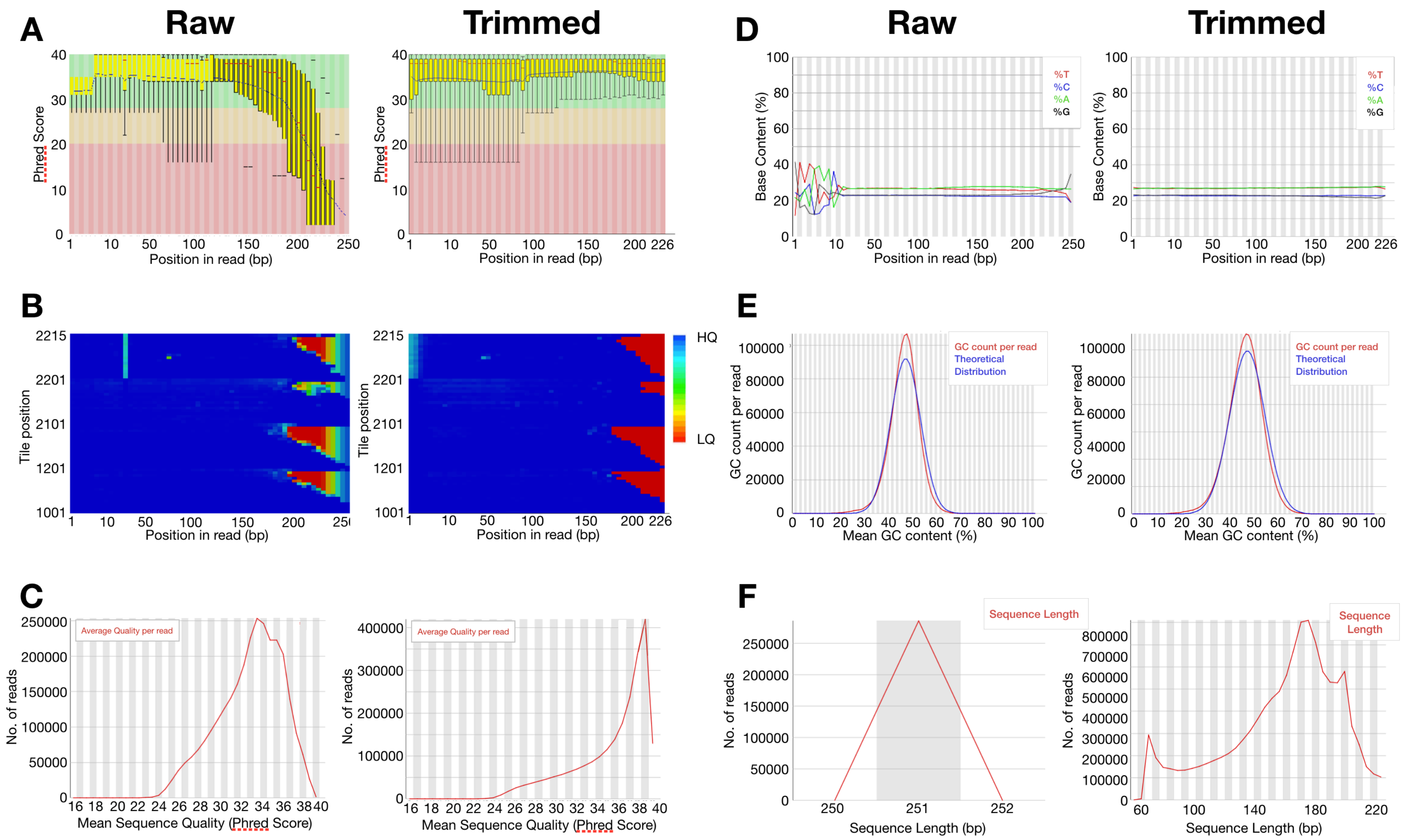


Figure 4.6. FastQC analysis of 309.76's 250 bp reverse Ilumina library. This library is typical in characteristic to the rest of the 100 bp libraries. Raw read library and its corresponding trimmed library counter part were analysed for quality. A) Average quality by our base position B) Average quality by tile positioning. HQ = high quality and LQ = Low quality C) Average quality per read. D) Base content distribution by per base location. E) Theoretical Gaussian GC content distribution and observed GC content of a library. F) Number of reads by sequence length.

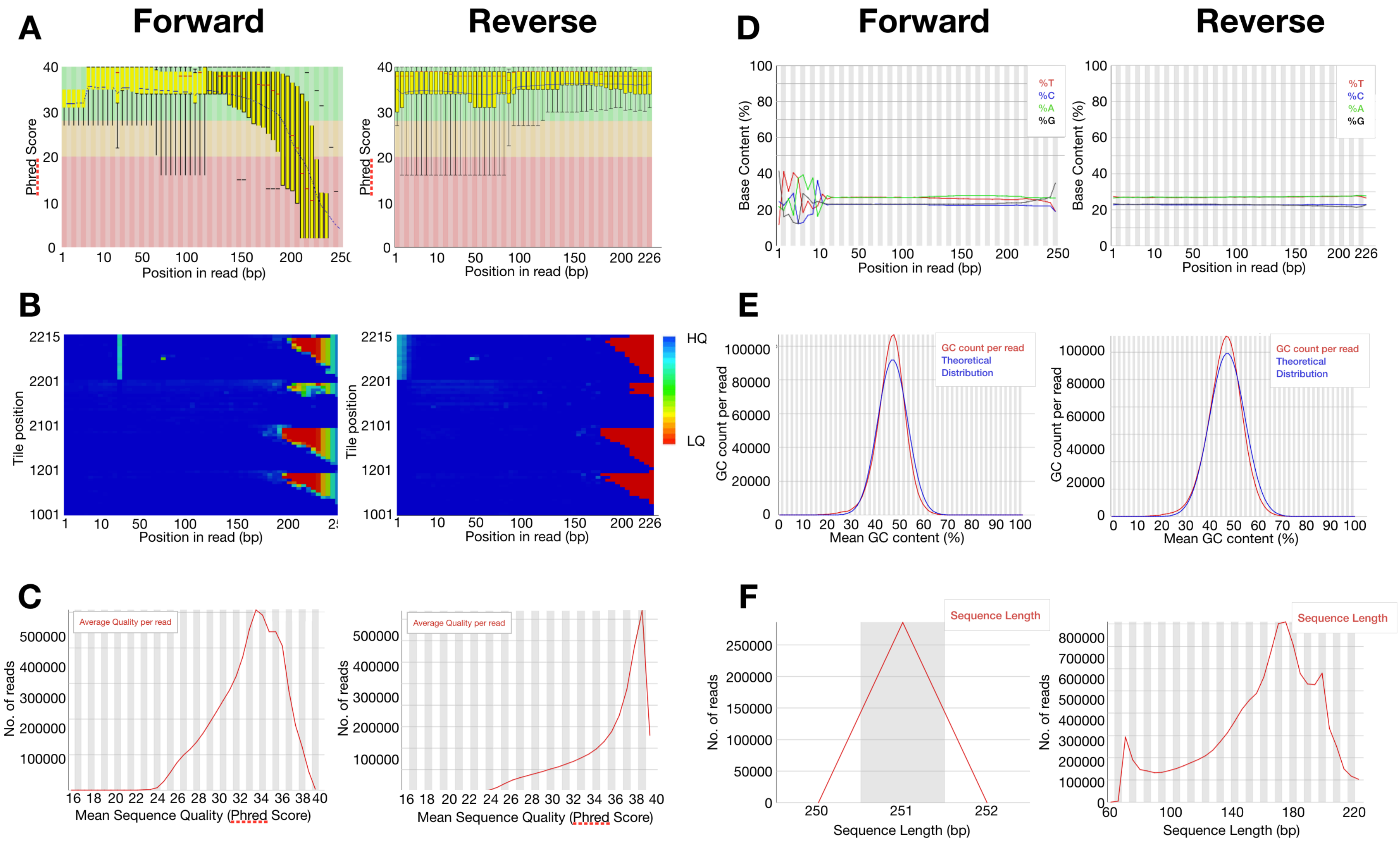


Figure 4.7. FastQC analysis of 309.71's 250 bp surviving post trimmed Ilumina library. This library is typical in characteristic to the rest of the 100 bp libraries. Raw read library and its corresponding trimmed library counter part were analysed for quality. A) Average quality by our base position B) Average quality by tile positioning. C) Average quality per read. D) Base content distribution by per base location. E) Theoretical Gaussian GC content distribution and observed GC content of a library. F) Number of reads by sequence length.

4.2.3 Post-trim Quality

With the 100 bp libraries, light trimming was enough to yield higher quality libraries. It was able to remove low quality reads which increase overall quality in every position of a read (Figure 4.1A and 4.2A) and resulted in the removal of low quality spots (Figure 4.1B and 4.2B). Figure 4.1C and 4.2C indicate that all reads have a quality of 32 or higher, with no surviving reads below that value.

Trimming the 250 bp libraries yielded post-trimmed libraries of 60-226 in read length (Table 4.1). Figure 4.4A and 4.5A indicate the trimming process to be successful in increasing the quality of the reads across all base position. This resulted in higher quality libraries as observed in figure 4.4C and 4.5C. Since the forward paired end libraries were of high quality, it did not benefit much from the quality trimming. The reverse libraries, however, benefited by cropping the distal bases from the reads. In Figure 4.4A, we observed the removal of low quality hotspots. Figure 4.4C also indicate the trimming process reduce the number of low quality reads. Importantly, the trimming process was able to preserved read length, with majority of these reads being >200 bp in length.

An improvement in quality is immediately apparent and beneficial in the reverse library, which suffers from low quality reads. Trimming has improved the average base call accuracy for some bases from 90% to ~99.99% in the reverse library (Figure 4.5A). This increase in quality has led to an increase in the average quality per read. This can be observed in Figure 4.5C where the mean score for the raw reads was 33 and the post-trim mean score was 39. A minor setback for the reverse libraries were the overall reduction in sequence length, as observed in Figure 4.5F. The trimming process has produced libraries of varied read length, with very few read length exceeding 200 bp. In Figure 4.5A, quality trimming did not remove the low quality hot spots (position 185-226 bp) from reads. While this is not favourable, further investigation of the N base plot indicates the removal of all N bases from the reverse paired-end libraries in the affected base position. We were also able to remove artefacts associated with Illumina-Seq from both forward and reverse paired end libraries. In Figure 4.4D and 4.5D, trimming was able to remove hexamers from the first 20 bp of all reads.

Trimming also produced surviving unpaired reads for both the forward and reverse reads. These surviving reads are unpaired and contained the discarded trimmed portion of the paired reads that passed the quality trimming process. When the unpaired libraries from the 100 bp set were analysed (Figure 4.4), we observed high quality surviving reads on both forward and reverse libraries. These were comparable in every FastQC metric to the corresponding paired library. However, in the 250 bp libraries, we observed that only the forward unpaired libraries contained a significant high number of quality reads (Figure 4.7). In these libraries, almost all of the reads were high quality reads, with a mean quality score of 38 in almost every base position (Figure 4.7A and 4.7C) and they were comparable to the high quality trimmed paired libraries. It should be noted, however, that these libraries contain reads that are highly varied in read length (Figure 4.7F). In contrast, the reverse unpaired library inherited reads that can be considered unusable. This library contained reads with a mean quality score of 17 (Figure 4.7F). Moreover, the majority of the libraries were composed of short 60 bp reads. Furthermore, in Figure 4.7D and 4.7E, we can see the artefacts and biases in these libraries. These characteristics found in these unpaired reverse reads render these libraries unusable for assembly.

Finally, in all libraries we were unable to observe any remaining adapters and artefacts. Furthermore, high quality surviving unpaired libraries present an opportunity to complement their counterpart trimmed paired end libraries during assembly.

Table 4.1. QUAST analysis based on draft genomes from the *de novo* assemblies of six *A. sarcoides* isolates. Assembly metrics for each draft genome. NNRL 50072, a publicly available draft genome, was not assemble here but is used as a point of reference.

	NRRL 50072	170.56	171.56	246.80	309.71	44013	64019
Contigs							
# contigs	219	694	393	443	6,086	2,088	2,464
>= 0 bp	219	1,799	1,434	1,575	15,011	4,570	4,194
>= 1000 bp	208	643	312	355	3,502	1,671	2,061
>= 5000 bp	179	531	165	204	1,858	1,110	1,376
>= 10000 bp	160	459	128	150	1,123	840	976
>= 25000 bp	137	341	114	118	333	471	461
>= 50000 bp	114	235	100	107	48	193	133
Largest contig	1,148,004	495,260	1,349,290	2,018,474	95,707	241,082	142,253
Length							
Total length	34,170,264	35,605,577	35,512,469	35,639,682	36,027,225	35,129,103	34,558,154
>= 0 bp	34,170,264	35,748,300	35,688,558	35,828,617	39,140,590	35,988,335	35,163,020
>= 1000 bp	34,161,224	35,568,786	35,455,190	35,576,066	34,299,627	34,851,363	34,281,800
>= 5000 bp	34,095,542	35,261,556	35,111,593	35,235,480	30,389,930	33,424,311	32,525,810
>= 10000 bp	33,957,701	34,747,741	34,860,823	34,858,197	25,080,181	31,477,105	29,635,007
>= 25000 bp	33,608,825	32,756,561	34,642,722	34,304,914	12,688,724	25,540,599	21,164,443
>= 50000 bp	32,734,677	29,009,437	34,142,554	33,922,358	3,058,660	15,771,321	9,822,451
Score							
N50	395,555	109,481	494,504	540,291	17,390	43,792	31,962
N75	216,138	64,045	255,622	246,158	8,155	22,771	16,592
L50	27	89	24	21	590	232	323
L75	55	195	48	47	1,338	505	690
GC (%)	46	46	46	46	46	46	47
Mismatches							
# N's	9	1,680	567	914	56,028	903	142
# N's per 100 kbp	0.03	4.72	1.6	2.56	155.52	2.57	0.41

4.2.4 Genome assembly and assembly quality

Once libraries trimming were of satisfactory quality, the forward paired, reverse paired and the high quality unpaired libraries were used for genome assembly by SPAde assembly tool (Bankevich et al., 2012). Using this approach, we were able to achieve *de novo* draft genome assembly for six different *A. sarcoides* isolate. Each assembly were benchmarked by Quast and BUSCO. For benchmarking, we included NRRL 50072, a previously studied *A. sarcoides* isolate sequenced by Roche 454 system and assembled by Roche's assembler (Gianoulis et al., 2012).

All *A. sarcoides* assemblies have similar characteristics to NRRL 50072. Each isolates' draft genomes were approximately 35 ± 1 Mb in length, with only minor deviations in size. We also observed minimal deviation in GC % content, 46.2 ± 0.2 %. When the GC % were plotted, we observed that the GC plot to be perfectly aligned to the theoretical bell-shaped curve with no shoulders. Both of these observations support the assumptions that these isolates are closely related and indicates the absence of contamination. Moreover, similarity in assembly length allows us to perform meaningful quality comparison. The quality of our assemblies were dependent on the sequencer used. The assemblies sequenced by 100 bp Illumina yielded higher quality assemblies in contrast to those sequenced by 250 bp Illumina. Furthermore, we also observed assemblies contained a lower frequency of contigs derived from the 100 bp Illumina set. This indicate that most reads were successfully aligned to a small number of contigs. Supporting this observation were the N50, which is defined as the minimum contig length required to cover half the genome. In practice, N50 can be seen as a measurement for sequence contiguity and the distribution of contigs within a given assembly. It is a metric that favours long contigs. Here we report the assemblies compiled from 100 bp Illumina libraries to have a similar N50 value to the previously studied NRRL 50072 model. In contrast, the assemblies from 250 bp Illumina set yielded poorer N50 scores reduced by approximately a factor of 10 (Figure 4.8). However, N50 metric does not account for errors such as misassemblies, read fidelities, or gaps.

The metric N's per 100 kbp estimates the number of uncalled bases within a given assembly. All assemblies have similar values, 0.41 - 4.72. An outlier is the draft genome of 309.71, with a value of 155.52 (Table 4.1). This is undesirable. Uncalled bases are functionally gaps within our assemblies thus indicating low coverage.

These gaps can be problematic in downstream annotation processes and can culminate in misalignment, fragmentation, or absent annotation.

Although correct assembly is important for fidelity, ultimately it is unable to measure the functionality. For this we benchmark with BUSCO to quantify the functionality within all draft genomes. It is capable of searching for well conserved single copy ortholog genes in distantly related datasets by leveraging HMM profiling. In practice, it is an important tool for estimating the “recovery” of genes from our assemblies. For our analysis, we ran BUSCO on BUSCO’s ascomycetes database. This contained 1315 curated single copy genes. All draft genomes returned more than 90% complete BUSCO genes (Table 4.2). This threshold indicates each assembly to be of a satisfactory quality, especially for distantly related dataset (Simão et al., 2015). Again, we observed the BUSCO scores for the100 bp Illumina sets to be comparable to NRRL 50072. While the 250 bp set scored adequately, it contained higher number of fragmented BUSCOs. However, the sum of completed and fragmented BUSCOs were constant throughout all assemblies. This indicated that, despite fragmentation and high amount of uncalled bases, BUSCO was able to return genes for the 250 bp library.

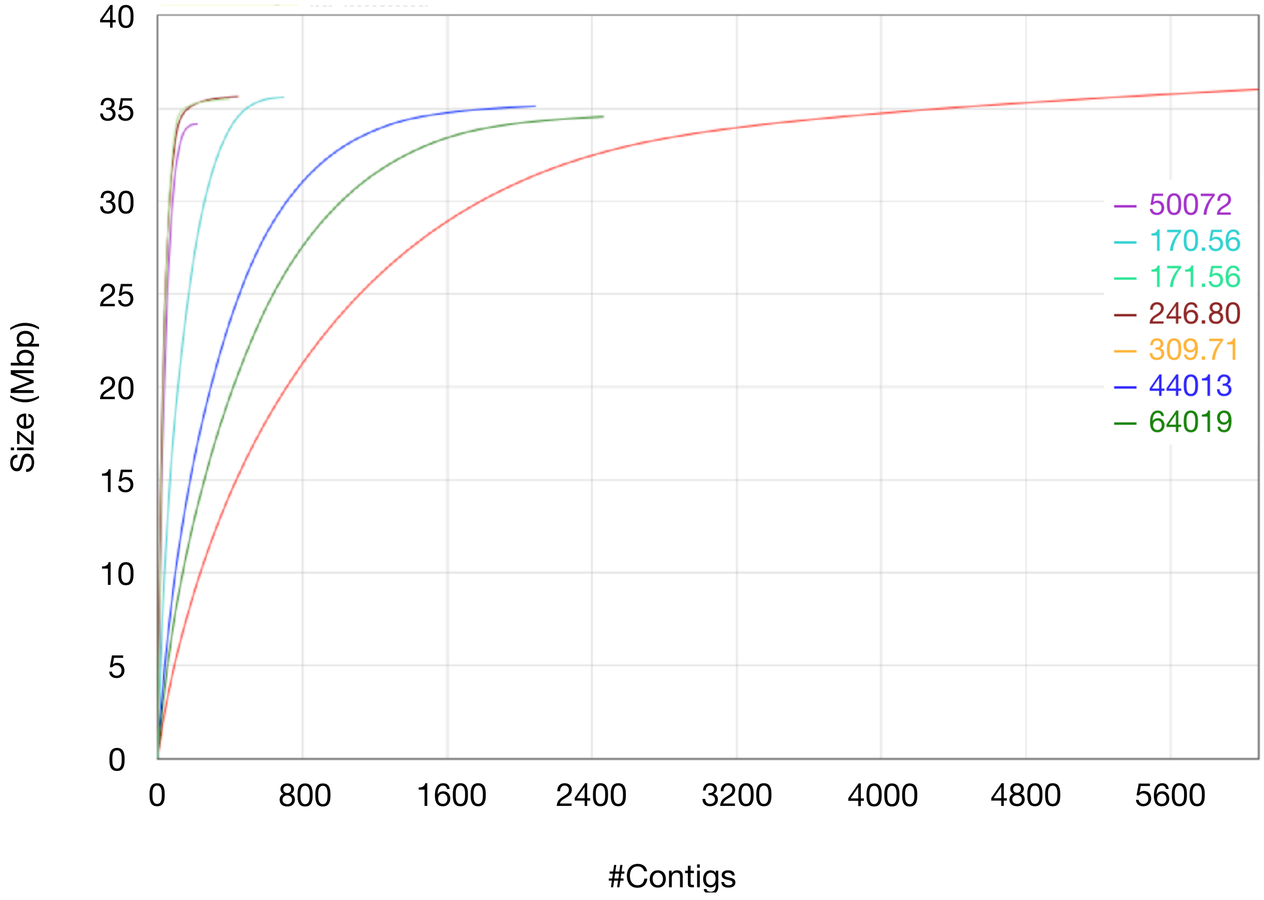


Figure 4.8. Cumulative plots of six draft genomes and a reference NRRL 50072 draft for comparisons. The cumulative plots indicate the distribution of a draft genome to the #contigs. The shorter the plot, the better the coverage for the draft genome.

Table 4.2. BUSCO summary report of each assembly. Complete indicates the sum of single and duplicated BUSCOs. Fragmented are partial alignment and missing are genes not detected from assembly. Ascsa1 is the assembled genome of NRRL 50072 and was analysed as a reference point for other isolates' assembly.

Isolate	Complete	Complete single	Complete duplicate	Fragmented	Missing	Total	Complete (%)
Ascsa1	1,297	1,292	5	7	11	1,315	98.6%
170.56	1,287	1,281	6	12	16	1,315	97.8%
171.56	1,292	1,286	6	8	15	1,315	98.2%
246.80	1,294	1,288	6	8	13	1,315	98.4%
309.71	1,218	1,208	10	58	39	1,315	92.6%
44013	1,268	1,261	7	25	22	1,315	96.4%
64019	1,254	1,257	7	28	33	1,315	95.3%

4.2.5 Predicting gene structure

The draft genomes were then analysed by MAKER2 for the purpose of gene prediction. For this we use the MAKER2 pipeline that houses several tools that enables gene prediction (Holt et al., 2011). This streamlines the usually complex process of using multiple tools to a relatively simple pipeline. We also enable repeats masking on all draft genomes to mask low complexity regions with RepeatMasker within MAKER2 (Tarailo-Graovac et al., 2009). Low complexity regions are amino acid sequences that can be found within protein sequences. These regions are characterised by repetitive single amino acid or motifs. When included in analysis, low complexity regions can bring about biases. Thus by filtering repeats we are able to; first, increase sensitivity than conducting BLAST search. Second, we reduce the query size thus is more computationally efficient.

Evidence based gene prediction was performed by using expressed sequence tag (EST) evidence and protein alignment from the published *A. sarcoides* NRRL 50072 genome (Gianoulis et al., 2012). Moreover, we used *ab initio* gene prediction approaches by leveraging HMM tools such as SNAP and Augustus (Stanke et al., 2004, Johnson et al., 2008). This double approach, while computationally expensive, is comprehensive in its prediction approach. We include published transcripts and annotated proteins from NRRL 50072 to align against our draft genomes to elucidate gene structure features. This is especially crucial for eukaryotic systems as current gene structure prediction pipelines are not viable due to the lack of sensitivity and specificity. Moreover, this also means that our pipeline strictly favour predictions based on biologically relevant data. Once gene structures were identified, we are able to train HMM pipelines to be genome specific. This approach is useful for tailoring protein prediction approach against distantly related organisms. MAKER2 compiled these different results generated by different pipelines to produce a final annotation. This final annotation is mathematically measured by a metric known as Annotation Edit Distance. This is a measurement for congruency between synthesised annotation and its supporting evidence. BUSCO analysis was again used to benchmark our annotation performances (Simão et al., 2015). The reason for this was to compare the recovery of well conserved orthologs of our annotation process to that of BUSCO and to compare recovery of our annotation dataset to our previous assemblies. On the first run we observed a slightly lower BUSCO score for all annotation set than genomes from the initial assembly.

During the first round of MAKER annotation, we used an evidence-based approach and *ab initio* approach using an Augustus' *Botrytis cinere*an HMM profile. This was to leverage the profile of a closely related ascomycetes which shares the same order (Helotiales) with *A. sarcoides*. On the first pass using MAKER, our approach predicted approximately 10,000 ± 100 protein encoded genes per draft genome. A second round of MAKER annotation was performed. This round of included isolate specific HMM profiles from SNAP, derived from iterative round of training, and Augustus, generated from BUSCO analysis. The rationale for including these different HMM profiles was to increase specificity for each isolate in the *ab initio* annotation procedure. On average, we see an increase of 91 protein encoded genes. In total we were able to predict around 10,100 - 10,200 protein encoded genes per isolate (Table 4.3). Finally, BUSCO protein analysis was performed on the protein set generated by each round of MAKER (Table 4.2). We observed three trends when we compared the results to the BUSCO analysis performed on the draft genomes. First, we see an increase in the detection of fragmented proteins. Second, we observed a reduction in complete protein annotation by BUSCO. Third, we were able to recover more missing proteins with MAKER.

4.2.6 Annotating protein encoded genes

Once prediction was complete, gene annotation was achieved by using BLASTX with databases such as NCBI's NR, NCBI's refseq and Uniprot's SwissProtKB. These databases were chosen for their comprehensiveness (NCBI's NR) and for their curation (refseq and SwissProtKB). Overall, our annotation strategy was to maximise sensitivity, specificity and accuracy for the annotation of well conserved and novel genes. With this process we were able to achieve: >90% successful annotation against the aforementioned databases.

For gene ontology annotations, BLASTX was used to query transcripts derived from MAKER2. Each transcript set was queried against the UniProt database for *A. nidulans* FGSC A4 and *S. cerevisiae S*288C. Each isolate is able to achieve > 99% match when e-value is set to 10 against both databases. When e-value is increased to 1e-10, this match is reduced to ~7200 hits for *A. nidulans* and ~5800 hits for *S. cerevisiae*. Based on higher significant hits, we used the annotation set from *A. nidulans* FGSC A4 for gene ontology functional annotations. For this, we employed

PANTHER version 14.0 databases and collection of tools. PANTHER was chosen because of its robust workflow, which incorporates BLAST and HMMer search strategies. This maximises sensitivity. When BLAST queries returned with no significant hits, it is re-queried with HMMer. With HMMer, an HMM profile of the query is generated for a tailored search. The analysis was carried out by PANTHER's gene list analysis and run on the *A. nidulans* database (designated as *Emericilla nidulans* on PANTHER) (Mi et al., 2013). The workflow was successful in mapping approximately ~5600 IDs to each isolate. From this functional annotation, PANTHER was able to assign PANTHER's GO-Slim terms for molecular function, biological process, and cellular component. We were able to annotate approximately 3200 molecular functions (Figure 4.9), 3600 biological processes (Figure 4.10), and 3200 cellular components (Figure 4.11).

To map our predictions to detailed pathways, we used the KEGG Automatic Annotation System (KAAS) to annotate our MAKER derived transcripts (Moriya et al., 2007). Our transcripts were set to BLAST against fungi from each of KEGG's fungal phyla databases (Filtered with e-value of 1e-60). Approximately 3300 transcripts were successfully annotated. These were mapped onto KEGG's ontology, modules, reaction modules, and pathways. Approximately, ~370 KEGG pathways were identified in each isolate, though it has to be noted that for some predicted pathways only a singular gene was identified

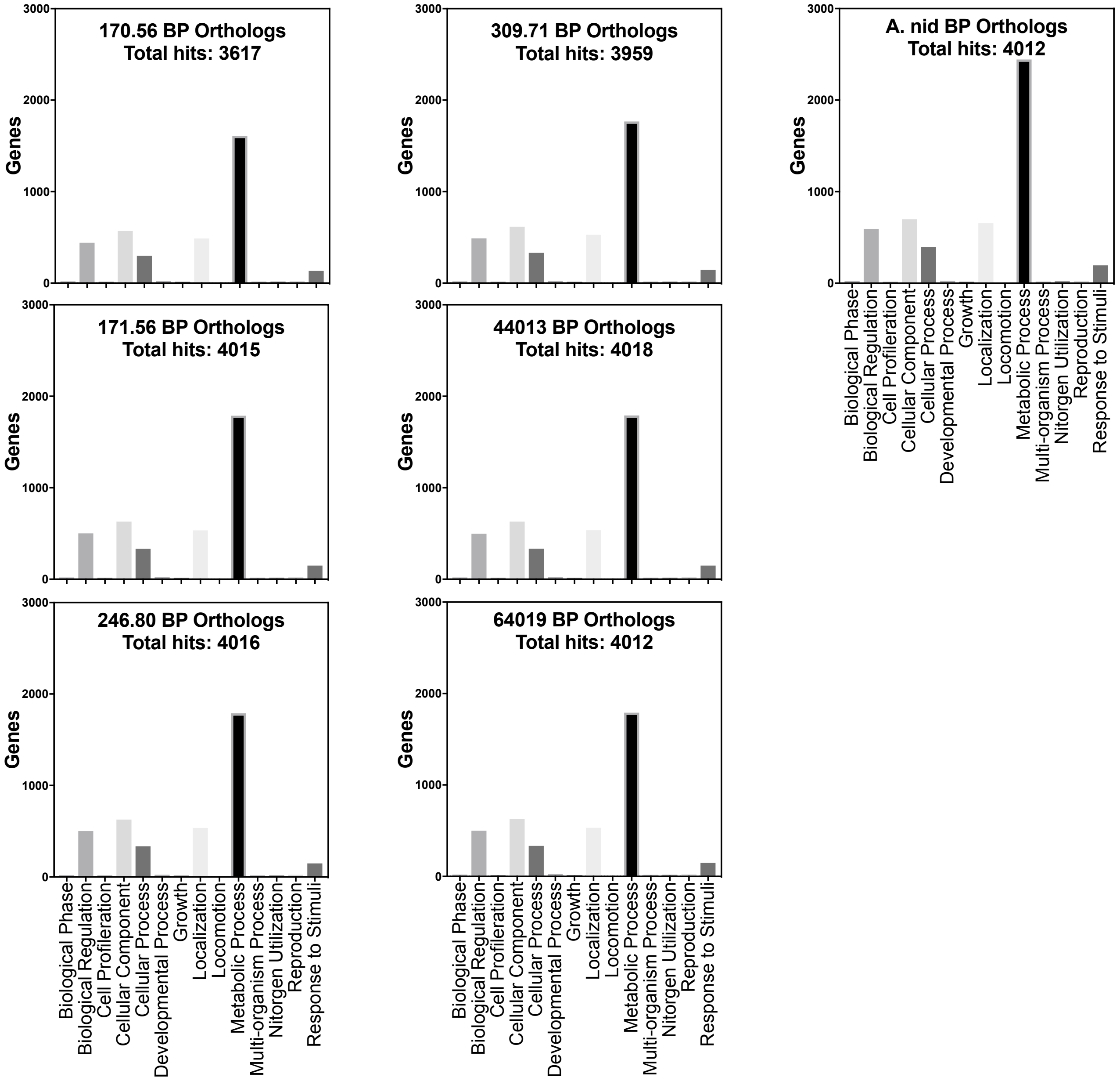


Figure 4.9. Biological processing (BP) ontology terms of six *A. sarcoides.* Annotation was achieved by PANTHER's workflow and annotated with PANTHER's *A. nidulans* database. These ontological terms to describe genes with a specific biological function.

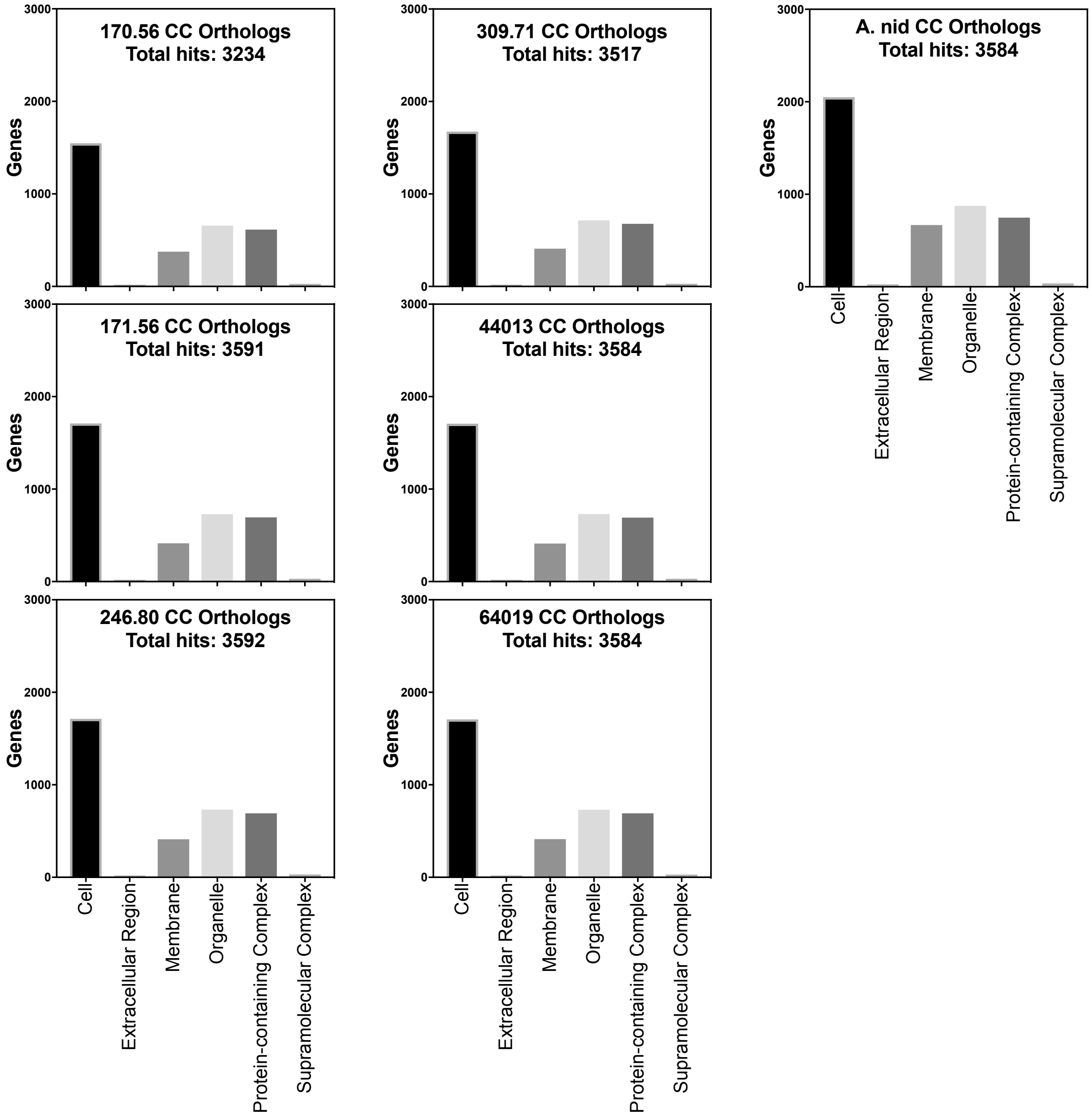


Figure 4.10. Cellular components (CC) ontology terms of six *A. sarcoides.* Annotation was achieved by PANTHER's workflow and annotated with PANTHER's *A. nidulans* database. These ontological terms to describe genes that are localised to a particular cellular partition.

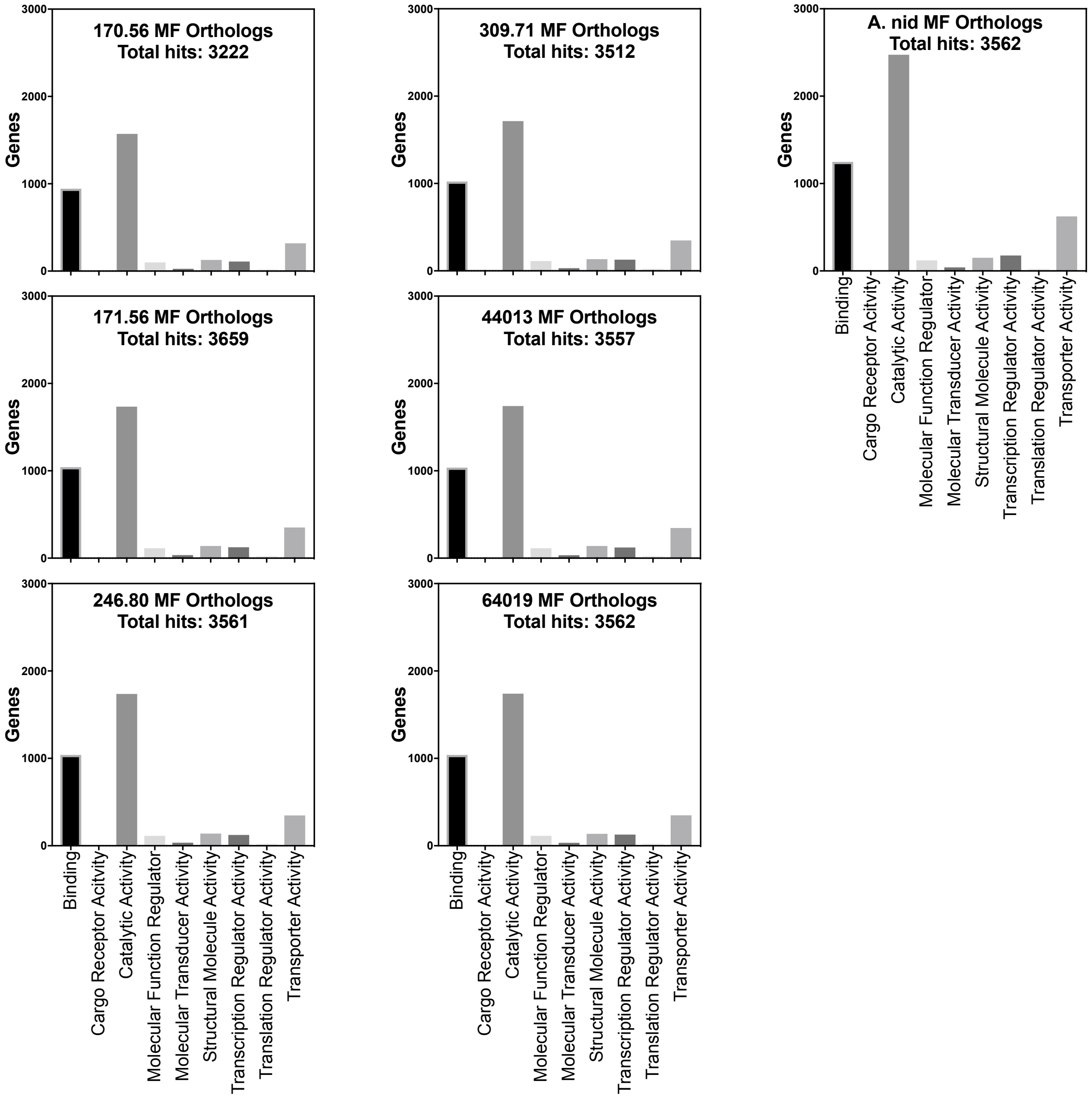


Figure 4.11. Molecular functions (MF) ontology terms of six *A. sarcoides.* Annotation was achieved by PANTHER's workflow and annotated with PANTHER's *A. nidulans* database. These ontological terms to describe genes with a specific molecular function. Usually, MF genes interact with molecules to achieve biological function.

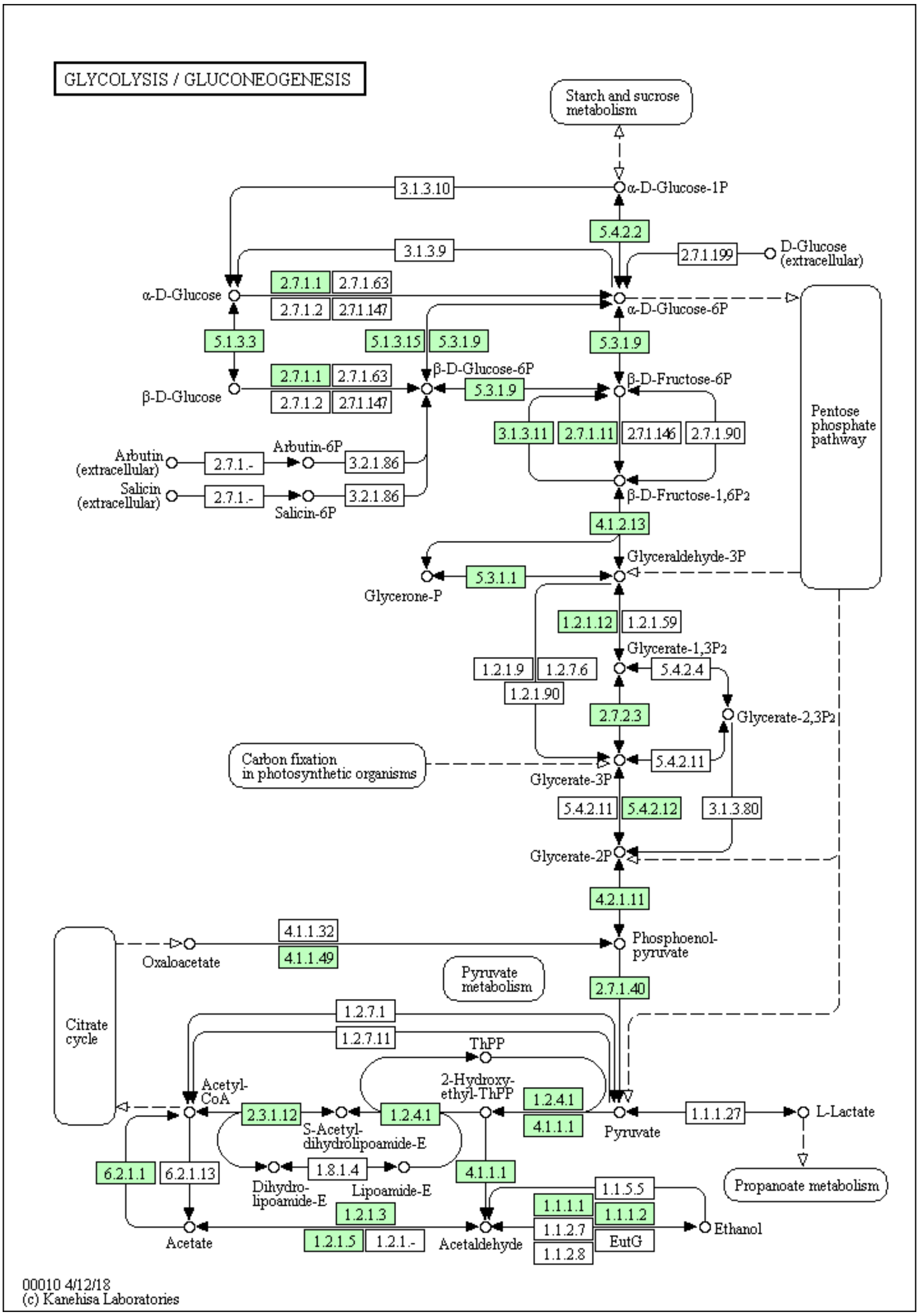


Figure 4.12. Pathway annotation of MAKER-identified genes. Example of KEGG annotation: Complete annotation for the glycolysis pathway present in *A. sarcoides* 170.56. Boxes highlighted in green indicate the presence of *A. sarcoides* genes that are good significant hits (E-value 1E-60) to genes involved in theglycolysis pathway. Annotation was based on MAKER2's predicted nucleotide sequence and was searched with BLAST in the KAAS pipeline.

4.2.7 Phylogeny

A phylogenetic tree was constructed by Multiple Alignment using a Fast Fourier Transform (MAFFT) program from amino acid sequences of 1092 highly conserved single orthologs identified by BUSCO. These BUSCO orthologs were derived from all isolates and NRRL 50072. In addition, for this analysis we also included BUSCO orthologs from *A. nidulans*, *S. cerevisiae* and *B. cinerea*. Only BUSCO orthologs that were common in all datasets were used. A custom script was used to parsed BUSCO output. This identified common genes across all datasets and concatenate amino acid sequences into a FASTA format. The average length of the concatenated amino acid sequence is 6,921,781. These concatenated sequence were organised under the organism they were from. These were then aligned and analysed by the MAFFT web server to yield a phylogenetic tree in Figure 4.12 (Katoh et al., 2013).

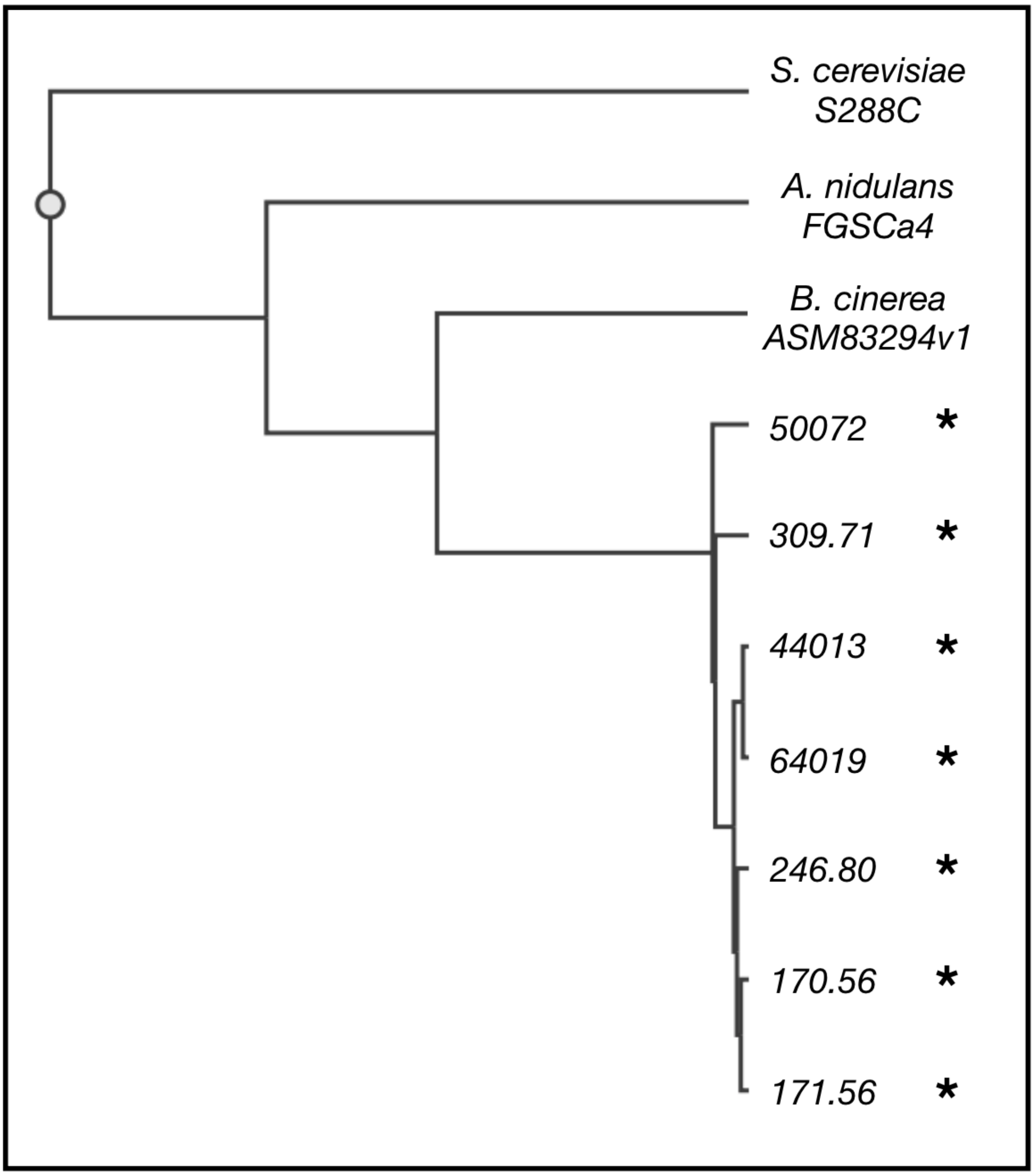


Figure 4.13. Phylogeny tree based on BUSCO genes was aligned and constructed by MAFFT. BUSCO genes were derived from seven *A. sarcoides* isolates (*), *B. cineria*, *S.* cerivisiae, and *A.* nidulans. This tree is made up of 1039 concatenated amino acid sequences that are common to all organisms in this analysis and these genes' are highly conserved single copy sequences recovered by BUSCO.

4.3 Discussion

This work represents the first in depth bioinformatics study into all publicly deposited and available *A. sarcoides* isolates. Here we present six isolates that have been *de novo* sequenced, trimmed (Figure 4.2 to 4.7), assembled (Figure 4.8), and functionally annotated into draft genomes (Figure 4.9 to 4.12). The data generated here provides an essential framework for further investigation of the genetic basis of alkane biosynthesis.

4.3.1 Assembling Draft Genomes

For phylogenetically distant organisms it is difficult to assemble high quality genomes. There are currently few biologically relevant resources for non-model organisms and distantly related organisms. Historically, producing a high quality functional genome requires vast resources and collaboration on a global scale. To produce genomes to such a standard is unachievable within this project. As such, our goal here is to focus on delivering highly functional and accurate draft genomes for each *A. sarcoides* isolates

For each isolate, we were able to trim low quality reads from the corresponding isolate's library. Our rationale here is that high quality data leads to a high quality assembly ensuring the integrity of downstream processing. The objective of the trimming is to remove low quality regions, large sections of uncalled bases and sequencing artefacts, while balancing read length. This is important in *de novo* assemblies where no reference data can be used for the assembly stage. Thus it is necessary to produce highly mappable reads and to remove artefacts that may confound assemblers. Note that some assemblers incorporate the removal of artefacts such as adapter sequences from the final assembly, nonetheless we apply appropriate trimming to remove such sequences

We were successful in assembling our isolates' read libraries into draft genomes, albeit with varying quality. We can see the lower quality of 250 bp libraries (Figure 4.5 to 4.6) reduces the quality of its assembly (Figure 4.8). The lower quality reads in 250 bp read libraries meant that a vast majority of reads are trimmed yielding high quality but truncated reads (60 to 225 bp). The longer the surviving reads, the easier it is to produce overlaps thus increasing the length of the contigs. It has been shown that quality assemblies are possible with library containing reads of 30 - 50 bp

(Worley et al., 2010). In fact, a 2.25 Gb giant panda genome was assembled with reads with an average length of 52 bp (Li et al., 2010).

From our assembly, we can observe the loss of quality in isolate 309.71 where the quality of the assembly suffered the most in relation to 44013 and 64019. BUSCO analysis of 309.71 indicates the loss of quality limited gene annotations. We observed a higher percentage of fragmented and missing genes to the rest of the 250 bp libraries. When comparing our 250 bp libraries to 100 bp libraries, we observed a lower quality in our 250 bp libraries in terms of assembly metrics and annotation metrics. Our 100 bp assemblies were comparable in assembly and annotation metrics to NRRL 50072, which was sequenced with a Roche 454. Currently, we are unable to explain the quality discrepancies between our two libraries despite using the same platform. It may be that read length over certain threshold do not affect downstream assemblies or annotations. The severity of gene fragmentation can be problematic for downstream annotation pipelines and recovering missing genes remains difficult without re-sequencing the entire genome. Still, fragmented genes can be useful as a starting point for downstream annotation by changing alignment scores.

We are unable to elucidate the chromosomal organisation of *A. sarcoides* with our current data. Ascomycetes are known to have highly variable numbers of haploid chromosomes and varied structure (Wieloch, 2006). Though a previous publication relating to NRRL 50072 (Gianoulis et al., 2012) claimed to have assigned their scaffolds to chromosomes, their publicly available data repository did not include chromosomal data. This mean that there are currently no reference points to assign our data at the chromosome level. Assigning our drafts to a chromosomal level would be useful for downstream heterologous expression, by identifying the loci of interesting genes. Furthermore, depending on the fragmentation of the genome, having a contiguous genome allows us to investigate genes on a structural level by performing synteny analysis. Many fungal secondary metabolite genes that participate in the biosynthesis of a specific secondary metabolite are loosely clustered in one locus. This also includes transcriptional regulators and associated accessory genes such as transporters (Gianoulis et al., 2012). By having such detailed information with regards to the gene structures we can elucidate the relationship of genes on at cluster level. Lastly, chromosomal structures can also facilitate a detailed study of the phylogenetic relationship of *A. sarcoides* with other fungi.

We have shown in the previous section the limitation of high number and short reads Illumina sequencing and its impact on our data. Ideally, to deliver a *A. sarcoides* genome of the highest quality, resolution and fidelity, we need to complement our current sequencing data with long read DNA sequencing data of >100,000 bp length (e.g. PacBio, Nanopore, 10x Chromium). Advances in sequencing technology in the last decade have meant that it is now viable to produce long sequence reads (1000 - 10000 bp). Current limitations of long reads are the costs associated and the inability to provide high sequence resolution relative to sequences derived from Illumina. In such scenarios, we would be able to map the repetitive and redundant Illumina sequences on to longer and more informative context of long read sequences. We would be able to resolve the disadvantages of low quality assembly associated with Illumina-seq and low resolution of long read sequencing technologies. Numerous studies have also shown the possibility of assembling an entire genome or chromosomes within a single contig. Ultimately, such a methodology will lead to a highly structured and high quality gapless genome. Without a doubt, such a genome will be invaluable for exhaustive gene mining projects like ours.

Across all of our isolates we note the similarity of GC content, 46%, for each genome. This observation is in agreement to NRLL 50072 46%. In relation to the Ascomycota fungi, we see diversity in GC content across all well studied fungi. Species such as, *A. nidulans* is 50% (NCBI: GCA_000011425.1), *A. fumigatus* is 49.8% (NCBI: GCA_000002655.1), *S. cerevisiae* is 38% (NCBI: GCA_000146045.2), and *F. oxysporum* is 48% (NCBI: GCA_000149955.2). While within the Helotilales order, which includes *A. sarcoides*, we observed similar GC values to *B. cinerea* is 42%(NCBI: GCA_000143535.4), *M. brunnea* of 43% (NCBI: GCA_000298775.1) and *G. lozoyensis* of 46% (NCBI: GCA_000409485.1).

4.3.2 Functional Draft Genomes

With our MAKER2 pipeline we were able to identify over >10,100 protein coding genes in our isolate. This is a similar hit rate to those presented by NRRL 50072 , which reported 10,672 protein coding genes (Gianoulis et al., 2012). It is also comparable to the well studied *A. nidulans*, which has 10,687 protein coding genes (Inglis et al., 2013). Isolate 309.71 contained the fewest identified genes. The cause of this is thought to be the low quality sequencing resulting in high fragmentation of 309.71 genomes, culminating in an increase in poor annotation alignments.

Our results when querying *A. nidulans* and *S. cerevisiae*, returned a similar number of matched queries at E-value of 10, of approximately 10,000 significant hits each. However, when E-value was set to a stricter settings, at 1e-10, we see a larger discrepancies in number of matched queries between the two databases. This yielded a higher number of matches for the *A. nidulans* database than for the *S. cerevisiae* database. This is unsurprising as both *A. sarcoides* and *A. nidulans* share the same Pezizomycotina subdivision, which encompasses all filamentous fungi. In contrast, *S. cerevisiae* is part of the Saccharomycotina subdivision. Thus we see a correlation with the number of matches to evolutionarily distances of databases used. Overall, we observed a very similar number of matched queries for both database and at both e-value scores. This is an indication of the biological similarity of our isolates.

Across all of our isolates we are able to demonstrate the functionality of our draft genomes by assigning gene ontology (GO) terms to each successful queried transcript. We were able to describe our ontology in three different major domains; molecular function, biological process, and cellular component. This allowed us to observe our data from a biological perspective and gave us the ability to interrogate our genomes. Molecular function represents the set of annotation at the molecular level involved in catalytic or binding activity. Each isolate contained approximately ~3,500 molecular function annotations. Biological processes represent annotation with biological roles within an organism and organise it into hierarchical levels. This culminated in approximately ~4,000 annotations involved in biological process. Finally, cellular component describes the locality of our annotation within a cell and each isolate contained approximately ~3600 cellular compartment annotations. Interestingly, isolate 170.56 contained less annotation consistently in every ontology domain.

Both the PANTHER and KEGG annotation sets indicate the presence of a complete primary metabolism genes and pathways in all isolates. Pathways relating to respiration such as glycolysis, citrate cycle, pyruvate metabolism, and fatty acid degradation were completely annotated. Observed were also a full complimentary set of oxidative phosphorylation pathways. Our pathway annotation also included a full set of eukaryotic DNA replication system, splicesomes, RNA transport systems, RNA degradation, and eukaryotic ribosome biogenesis.

4.3.4 Phylogeny of A. sarcoides isolate

In earlier studies, *A. sarcoides* NRRL 50072 was misidentified as *Gliocladium roseum* (Stinson et al., 2003, Strobel et al., 2010b, Strobel et al., 2010a)*.* The morphological observation and production of reddish pigment after 10 days of culturing were a characteristic phenotype relating to *Gliocladium spp* (Strobel et al., 2010b). Internal transcribed spacer (ITS) sequencing was also undertaken to aid in the identification of NRRL 50072. ITS are spacer DNA and is used as an identification method by using molecular fingerprinting. This is because ITS sequences undergo rapid evolution resulting in a high degree of sequence variation even among closely related species (Song et al., 2012). Phylogenetic analysis indicate the ITS sequence of NRRL 50072 is similar to that of *A. sarcoides* (Strobel et al., 2010b)*.* In this case morphological evidence was favoured over ITS evidence. The study concluded NRRL 50072 to be an atypical isolate of *G. roseum* and may be an anamorph of *Ascocoryne spp.* Phylogenetic analysis indicates that the ITS sequence of NRRL 50072 to be similar to that of *A. sarcoides* (Strobel et al., 2010b). Subsequent study performed phylogenetic analysis on the ITS sequence and included tricarboxylate transport protein (CTP) gene sequence (Griffin et al., 2010). This was able to provide enough phylogenetic resolution to reclassify NRRL 50072 as an *Ascocoryne sarcoides*, albeit distant from other isolates of *A. sarcoides* (Griffin et al., 2010)*.* Similarly, isolate 309.71 was initially deposited as *Ascocoryne cylichnium*, a closely related species to *A. sarcoides*. This isolate was also reassigned to *A. sarcoides* when phylogenetic analysis indicate 309.71's ITS sequence achieved better alignment with the *A. sarcoides* clade (Griffin et al., 2010).

Although the phylogeny of *A. sarcoides* has been described, we hope to complement published results and provide further clarity with our data by leveraging whole genome sequence data. However, our analysis lacks data for other *Ascocoryne spp* as there were no publicly available genomes for these organisms. In our phylogenetic analysis, we based our analysis on the genetic variation of 1092 conserved single copy orthologs instead of ITS sequences. Thus our analysis is based on a larger amount of biological evidence. This means that we are able to provides greater clarity within the *A. sarcoides*. With this approach, we were able to confirm previous reassignment of NRRL 50072 to the *A. sarcoides* clade and it being the most distant in relation to other *A. sarcoides* isolates (Figure 4.13). Moreover, our analysis agrees with published results that *A. sarcoides* is organised into its own clade even among other Helotiales, in this case *B. cineria*. In MycoBank database (MycoBank #353), the *Ascocoryne* genus included 9 species. Of these *A. sarcoides*, *A. cylichnium*, and *A. solitaria* were phylogenetically assessed. While six other *Ascocoryne* species *A. albida, A. javanica, A. microspora, A. striata, A. trichophora,* and *A. turficola* may require further studies to confirm their identity and phylogenetic relationship.

4.4 Conclusion

We have sequenced and assembled 6 draft genomes of *A. sarcoides*. With these 6 draft genomes we assessed the quality of the genome to improve confidence in them. For three draft genomes, we were able to achieve the same quality as the published NRRL 50072 draft genome., while the other three draft genomes were of useable quality. Of these, 309.71 was the most fragmented. All draft genomes achieved higher than 90% complete gene identification on BUSCO analysis. This indicate that we have assembled draft genomes that are functional. We were able to predict over 10,000 protein encoding genes for each draft genome. These were annotated for function to describe biological ontology and biological pathways of all *A. sarcoides* draft genomes. The identification of biological ontology and pathways complements metabolic data generated in Chapter 3 and 4 in understanding alkane metabolism in *A. sarcoides*. Phylogenetic analysis was also conducted. This was based on conserved single copy orthologs to run a Multi Loci Sequence analysis. With this analysis, we were able to describe with great confidence the *A. sarcoides* clade. However, without genomic data from other *Ascocoryne* species and the Helotiaceae family, it is difficult to draw significant conclusions about the phylogeny of our isolates.

CHAPTER 5 IN SILICO GENE DISCOVERY

5.1 Introduction

In Chapter 4, we successfully produced high quality draft genomes for the following *Ascocoryne sarcoides* isolates: 170.56, 171.56, 246.80, 309.71, 44013, and 64019. This included the successful identification of genes by the use of machine learning and transcriptomics of NRRL 50072. Furthermore, genes were annotated by three different annotation pipelines, of which two were pipeline annotated against *Aspergillus nidulans* FGSCa4 and *Sacchromyces cerevisiae* S288c and two general annotation pipelines against the Kyoto Encyclopedia of Genes and Genomes (KEGG) and PANTHER database (Mi et al., 2013).

Although we were successful in establishing bioinformatic resources for each isolate, these resources remained unexplored for their biological significance. This can be achieved by comparative genomics, thereby highlighting novel genes by subtraction against a reference resource. Usually this is achieved by subtracting homologous genes by basic local alignment search tools (BLAST). It can also be achieved by mapping annotated genes to known pathways to describe their biochemical potential. This is the approach used by KEGG's automated annotation server (KAAS) (Moriya et al., 2007). There are issues with both methods: subtraction may remove paralogous genes, genes that are have high homology to reference genes but may have deviated in function. Mapping can also be confounded and made more complex with multiple annotation evidences. Thus, their is a need for a custom pipeline to aid comparative analysis and collate different annotations evidence across six isolates and against two reference genomes. This will aid mapping of putative pathways and, in addition, exploration of the biochemical potential of these isolates to pinpoint the biosynthetic pathways for alkane.

5.1.1 Aims and Objectives

To aid the elucidation of an alkane pathway, it is critical to develop the genomic resources. This chapter will cover and describe:

- Production of a library of non-redundant orthologs
- Evalaution of the results of comparative genomics
- Identification of established alkane genes in the genomes of *A. sarcoides*
- Proposed pathways relating to alkane metabolism

5.2.0 Results

5.2.1 Identifying orthologs in A. sarcoides isolates

There are 60, 656 predicted proteins present across all six isolates of *A. sarcoides*. To use these annotations, it is important to conduct meaningful cross-isolate comparisons by identifying clusters of highly orthologous protein clusters. For this analysis, each protein in our six *A. sarcoides* datasets was given a unique accession code for traceability. Data from the publicly available genome was included to increase the accuracy of the analysis. Identification of orthologous proteins is achieved first by discerning reciprocal pairs. A reciprocal pairing in this analysis was defined by the following definition: when Gene A significantly aligns with Gene B and Gene B significantly aligns with Gene A, it is considered a reciprocal pair. To achieve this, we have built a computational program (Figure 5.1). The program begun with an indiscriminate proteome vs proteome BLASTP search of 7 isolate genomes and two reference genomes, in which no filter was applied and BLASTP produced a single highest scoring pair (Boratyn et al., 2012). This was chosen as it allows for downstream tuning for filter values (e.g. E-Value and Percentage ID) and to reduce computational complexity. This BLASTP process includes a recursive BLASTP search of every proteome to identify potential isoforms or gene duplicates.

Once reciprocal pairs are identified, it is possible to computationally assigns a specific gene to its reciprocal pair and to associate of the reciprocal pair to other pairs if the matches pass an e-value cut-off of 1e-3 and a percentage ID of 90%. Membership to a specific orthologous protein cluster requires every protein to be reciprocal to each other or to a specific protein within a cluster, thus this analysis assumes orthology from relative evidence. Using Figure 5.2 as an example, Gene 1, 2 and 4 are reciprocal to each other. If gene 4 and 5 are reciprocal, Gene 5 is included in the cluster. Moreover, any proteins that are unable to cluster with another protein are considered to be a singlet. This approach networks reciprocally homologous protein, independent of the gene's annotation.

By cross-comparing six different proteomes, 16, 488 orthologous clusters and, a total of 127 intersections were identified. Each cluster was given a unique accession for identification of the cluster and proteins within the cluster. Of this, 6, 756 clusters (50%) contain proteins that can be found in all seven isolate of *A. sarcoides* (Figure 5.3). Each intersection describes the isolate's proteome cover by a specific cluster and with the number within the intersection describing the number of clusters associated with a specific intersection. There are 10, 661 (64.66%) clusters which contain proteins that successfully network with one or more protein from another isolate and the remaining 5, 827 are not reciprocal with any other proteins from another isolate (Table 5.1). Interestingly, the second and third largest group of clusters are proteins from isolate *A. sarcoides 309.71* and the publicly available *A. sarcoides NRRL 50072* (Figure 5.3). Importantly, this analysis enables the identification of unique and shared genes in each and every isolate's proteome, the elucidation potential protein isoform within the dataset, and also produced a non-redundant list of orthologous protein clusters.

Proteins belonging to *A. sarcoides* NRRL 50072 were excluded from further analysis at this point and this reduced the number of clusters from 16, 488 to 13, 489 (Table 5.2). Insights from orthologous clusters were combined with annotation data (Chapter 4) to identify clusters of known and unknown identity in our *A. sarcoides* isolates (Table 5.2). Clustering and annotation evidence from OrthoVenn2, an orthology clustering program based on OrthoMCL, were also included (Xu et al., 2019). This has the advantage of adding further annotation and clustering evidence to our combined dataset. Following the previous assumption, any cluster can be considered annotated when one or more proteins within the cluster forms a reciprocal pair with an annotated protein. In our annotations, 8,493 clusters contained one or more annotation and, conversely, 4,996 clusters remain unannotated and represent unique genes found only in *A. sarcoides* (Table 5.2). Furthermore, we can elucidate unique genes by subtracting reciprocal hits from the well-studied model organisms *A. nidulans FGSC a4* and *S. cerevisiae S288c.* With *A. nidulans FGSC a4,* 6, 834 clusters were matched and were subtracted while 6, 655 clusters have no annotation (Table 5.2). Likewise, 3, 654 clusters reciprocally matched with proteins from *S. cerevisiae S288c* were subtracted, while 9, 835 clusters did not produce an annotation (Table 5.2). A total of 7, 090 clusters that reciprocate with *A. nidulans FGSC a4* and/or *S. cerevisiae S288c* were removed. This yielded 6, 399 clusters unique to *A. sarcoides* (Table 5.2).

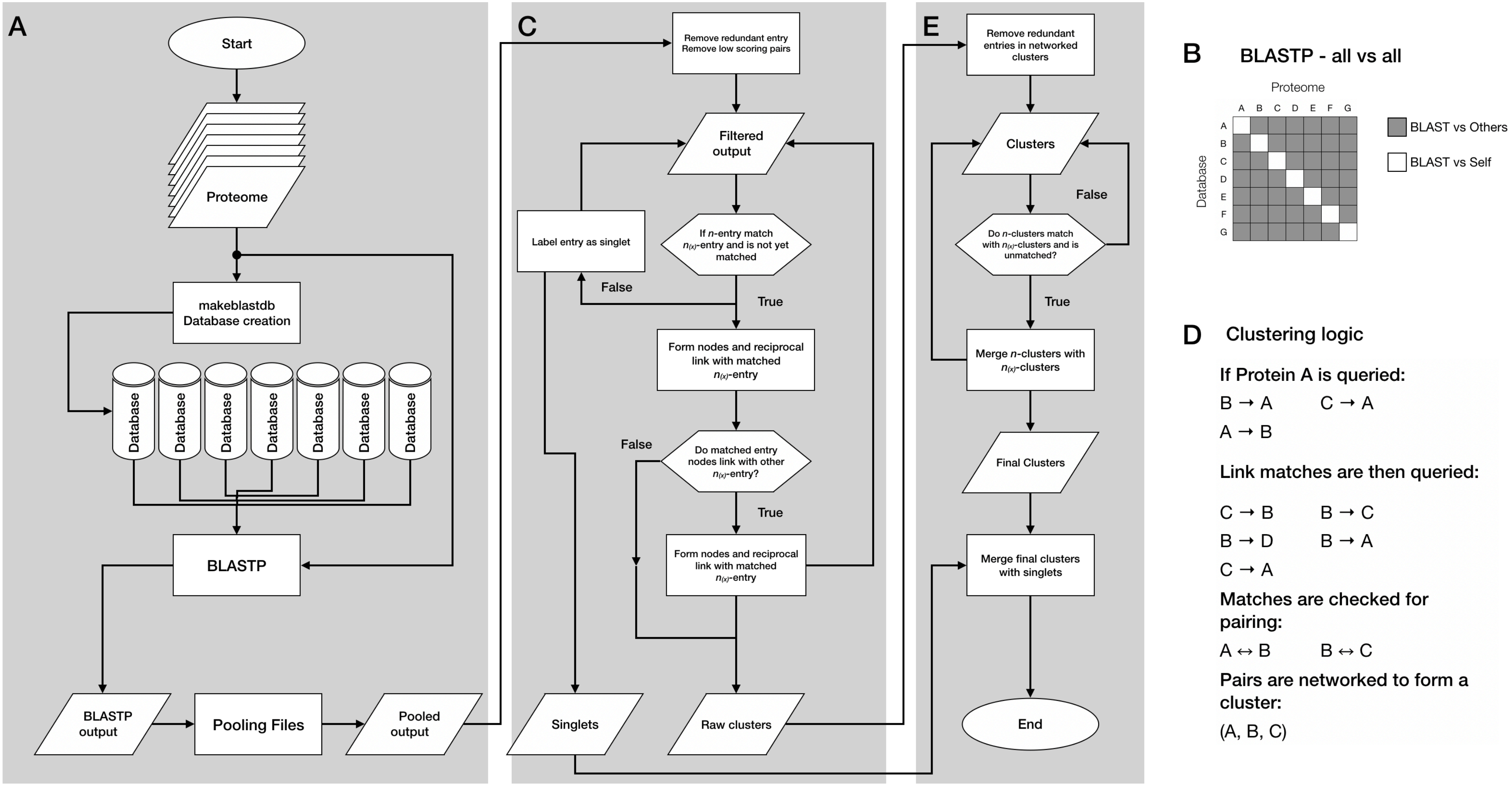

A
Start
Proteome
makeblastdb
Database creation
Database
Database
Database
Database
Database
Database
Database
BLASTP
BLASTP output
Pooling Files
Pooled output
C
Remove redundant entry
Remove low scoring pairs
Filtered output
Label entry as singlet
If n-entry match $n_{(x)}$-entry and is not yet matched
False
True
Form nodes and reciprocal link with matched $n_{(x)}$-entry
False
Do matched entry nodes link with other $n_{(x)}$-entry?
True
Form nodes and reciprocal link with matched $n_{(x)}$-entry
Singlets
Raw clusters
E
Remove redundant entries in networked clusters
Clusters
False
Do n-clusters match with $n_{(x)}$-clusters and is unmatched?
True
Merge n-clusters with $n_{(x)}$-clusters
Final Clusters
Merge final clusters with singlets
End
B
BLASTP - all vs all
Proteome
A B C D E F G
Database
A B C D E F G
BLAST vs Others
BLAST vs Self
D
Clustering logic
If Protein A is queried:
B → A C → A
A → B
Link matches are then queried:
C → B B → C
B → D B → A
C → A
Matches are checked for pairing:
A ↔ B B ↔ C
Pairs are networked to form a cluster:
(A, B, C)

Figure 5.1. Flow chart of reciprocal network analysis method used for clustering proteins. In 5.1A. a custom SLURM bash script creates a BLAST index database for each query proteome with NCBI's makeblastdb. The BLASTP output details the query, matched subject and associated score (5.1B). These outputs are then pooled for further analysis. To match reciprocal pairs, a custom python program identifies each protein entry as a query (e.g. Protein A in 5.1D) within the pooled output. The program will then attempt to identify the queried protein (A) against the pooled output and return successful hits (5.1D). Any queries that cannot form reciprocal pairs are considered to be singlets. With this evidence, reciprocal pairs can be networked together to find clusters. To achieve non-redundancy (5.1E), clusters that are identical are removed and clusters that are similar are pooled. The program will then output a list of clusters and singlets.

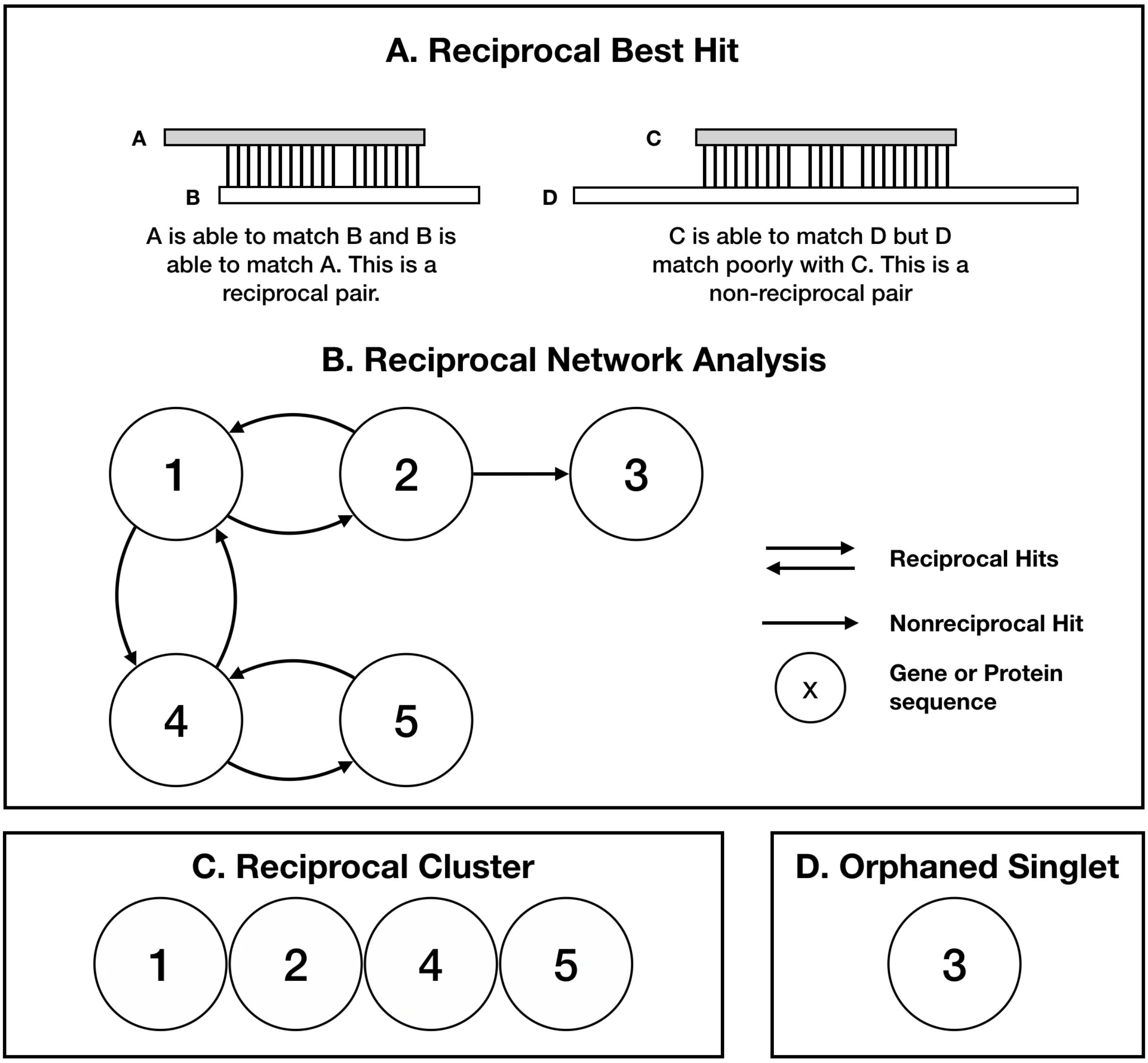

A. Reciprocal Best Hit
A
B
C
D
A is able to match B and B is able to match A. This is a reciprocal pair.
C is able to match D but D match poorly with C. This is a non-reciprocal pair
B. Reciprocal Network Analysis
1
2
3
4
5
Reciprocal Hits
Nonreciprocal Hit
X
Gene or Protein sequence
C. Reciprocal Cluster
1
2
4
5
D. Orphaned Singlet
3

Figure 5.2. Schematics of reciprocal network analysis, a scaleable method to cluster proteins by sequence homology. Reciprocal network analysis is based on reciprocal best hits. For example (5.2A), if Gene A is a match to Gene B and Gene B is a match to Gene A, both genes are considered to be a reciprocal pair. BLAST is used to identify reciprocal pairs against *n* dataset. Network analysis is then performed to cluster reciprocal pairs. For a pair to be a member of a cluster, it must be reciprocal to any nodes within a cluster and have the required parameters (e-value and percentage identity). As an example in 5.2B, Gene 1, 2 and 4 are reciprocal. Since Gene 4 and 5 are reciprocal, Gene 5 is included in the cluster (5.2C). While Gene 2 aligns significantly to Gene 3. However, it is not reciprocal as Gene 3 does not significantly align with Gene 2. Therefore Gene 3 is considered an orphaned singlet (5.2D).

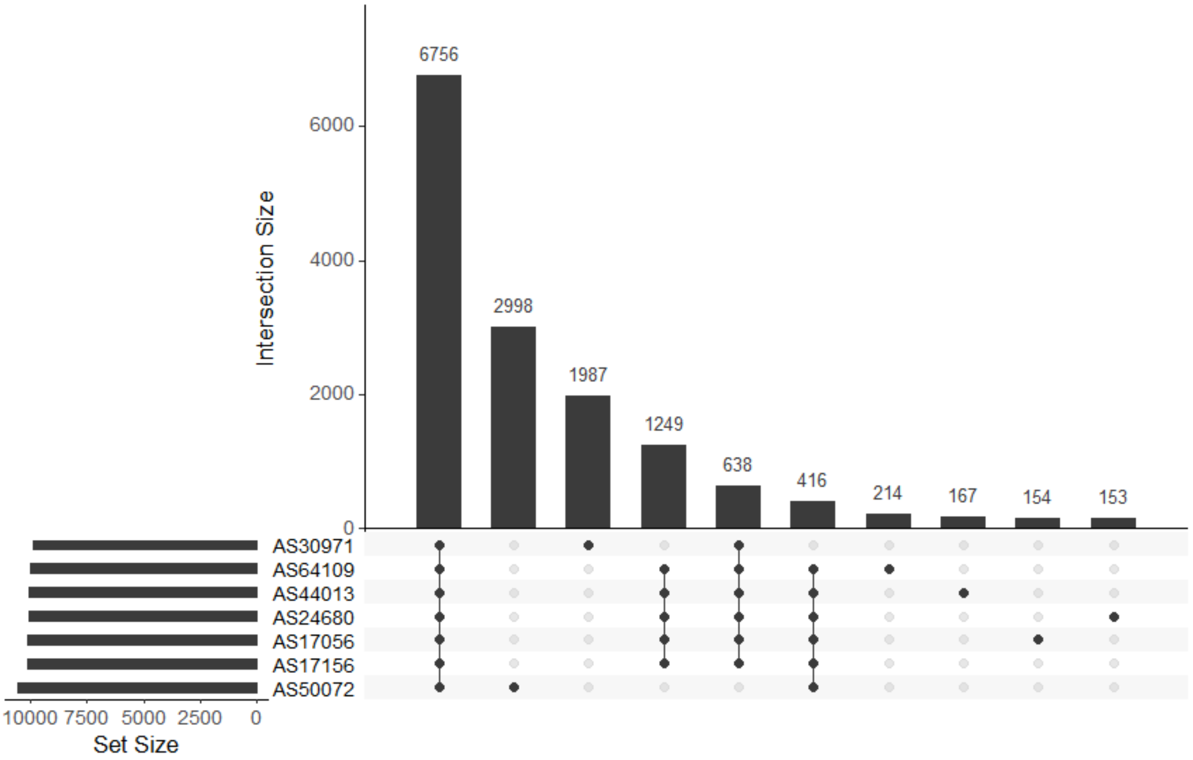


Figure 5.3. Number of orthologous clusters identified by reciprocal network analysis across seven isolates of *A. sarcoides*. The top 10 intersection by number of clusters (Y axis) are presented with clusters sorted by the size of intersection size. Intersection size is the number of clusters present within a specific overlap. Filled dots represent the overlapping proteome (categorical X axis).

Table 5.1. Number of orthologous protein clusters generated by reciprocal network analysis. Total clusters include clusters (2 or more proteins) and singlets (single protein clusters). This total number includes proteins from *A. sarcoides 50072* to improve the analysis result.

Category	Number of cluster
Total Clusters	16, 488
Clusters containing proteins from all isolates	6, 756
Clusters with 2 or more proteins	10, 661
Singlets	5, 827

Table 5.2. Number of orthologous protein clusters with numbers of annotations. Clusters with one or more proteins annotated by one or more annotation pipeline is considered to be annotated. Whereas clusters with no annotations represent proteins with no annotation evidence. Total clusters represent the number of clusters without proteins from *A. sarcoides 50072*.

Category	Number of annotated clusters	Number of unannotated clusters
Total Clusters	13, 489	
>1 annotations vs all pipeline	8, 493	4, 996
>1 annotations vs *A. nidulans FGSCa4*	6, 834	6, 655
>1 annotations vs *S. cerevisiae S288c*	3, 654	9, 835
>1 annotations vs *A. nidulans FGSCa4* and *S. cerevisiae S288c*	7, 090	6, 399

5.2.2 Evaluating the known alkane biosynthetic pathway

The pathway for alkane biosynthesis has been elucidated in cyanobacteria, *D. melanogaster*, and *A. thaliana*. These studies independently confirm a two-step mechanism for the conversion of fatty acid to n-1 alkane product. Initially, the fatty acid's carboxyl group is reduced to an equivalent length fatty aldehyde by an acyl reductase (AR). Fatty acids entering this pathway can either be in the form of free fatty acids or bound to either acyl carrier protein or coenzyme A. The fatty aldehyde undergoes carbon-carbon cleavage at the terminal carbonyl group by an aldehyde decarbonylase, producing an oxygenated carbon compound in the form of formate or CO_2. We examined the proteome of six isolates of *A. sarcoides* for genes that encode acyl reductase and aldehyde decarbonylase, which participates in known alkane biosynthesis pathway. A BLAST search analysis was undertaken to identify possible homologous analogs for acyl reductase and aldehyde decarbonylase. No previously known alkane biosynthesis gene was found with BLAST searches (Table 5.3). However, for CYP4G1, BLAST returned results associated with other genes encoded for P450 and were associated with specific secondary metabolite biosynthesis. These are thought not to be orthologous to CYP4G1. These findings support a novel genetic basis for alkane biosynthesis in *A. sarcoides*.

Table 5.3. TBLASTN annotation of protein sequence associated with alkane biosynthesis in other biological systems against the draft genome of six *A. sarcoides* isolates. BLAST search was conducted with a cutoff e-value 1e-5. Hits returned P450 genes associated with other well defined biosynthetic and are thought to be P450-related false hits.

TBLASTN Query	Isolate						
	170.56	171.56	246.8	309.71	44013	64019	NRRL 50072
CER1	0	0	0	0	0	0	0
CER3	0	0	0	0	0	0	0
AAR	0	0	0	0	0	0	0
ADO	0	0	0	0	0	0	0
UndA	0	0	0	0	0	0	0
OleT	2	2	2	2	2	2	1
CYP4G1	26	26	22	22	22	22	25

5.2.3 Elucidating the lipid hyperoxidase pathway

In 2012, a study hypothesised a possible lipid hyperoxidase pathway for alkane biosynthesis for *A. sarcoides* NRRL 50072 (Gianoulis et al., 2012). The hypothesised pathway was based on gene expression levels that correlate with metabolic observations. In this hypothesised pathway (Figure 5.4), free fatty acid in the form of a linoleic acid is oxygenated by a psi-producing oxygenase (PPO). The double bond present in linoleic acid coordinates PPO to introduce a mid chain hydroperoxy group at C10 by oxidising the mid chain acyl group and yields a 10-Hydroperoxyoctadecanoic acid. The next step requires hydroperoxide lyase to cleave the acyl chain into two fragments by coordinating the hydroperoxy group to catalyse carbon-carbon lysis. This generates two fragments, an unsaturated decanoic acid and either an 1-octen-3-one or an 1-octen-3-ol. Further reduction by a 3-oxo-acyl reductase, dehydratase and enoyl reductase yields an 1-octene.

Evaluating lipid hyperoxidase annotation

We evaluated the annotations of a previously proposed hypothetical alkane pathway. Reciprocal network analysis were able to cluster some of the proposed protein annotation with proteins from our *A. sarcoides* isolates. This was able to provide annotation for eight out of eleven proposed annotation. With reciprocal network analysis, three proposed annotations form singlets and were unable to cluster with other *A. sarcoides* proteins. Moreover, with Orthovenn2 pipeline two of the eleven proposed annotation could not form clusters. In summary, ten out of eleven were clustered by one of the two methods and from these ten cluster, it was able to provide coverage for the entire lipid hyperoxidase pathway. This approach was able to suggest the proposed annotation did not have the required activity for the lipid hyperoxidase pathway. For direct proof of sequence homology, each proposed gene sequence was extracted for BLASTP and HMMer search approaches were used to confirm the sequence identity of each proposed gene (Table 5.4). With these two approaches, it was revealed that the proposed annotation does not correlate with the suggested function at each step of the hypothesised pathway and largely agreed with annotation by clustering approaches. These findings suggest the hypothesised pathway, with the provided accession, is unlikely to have the biochemical potential to conduct the conversion of fatty acids to alkane or alkane.

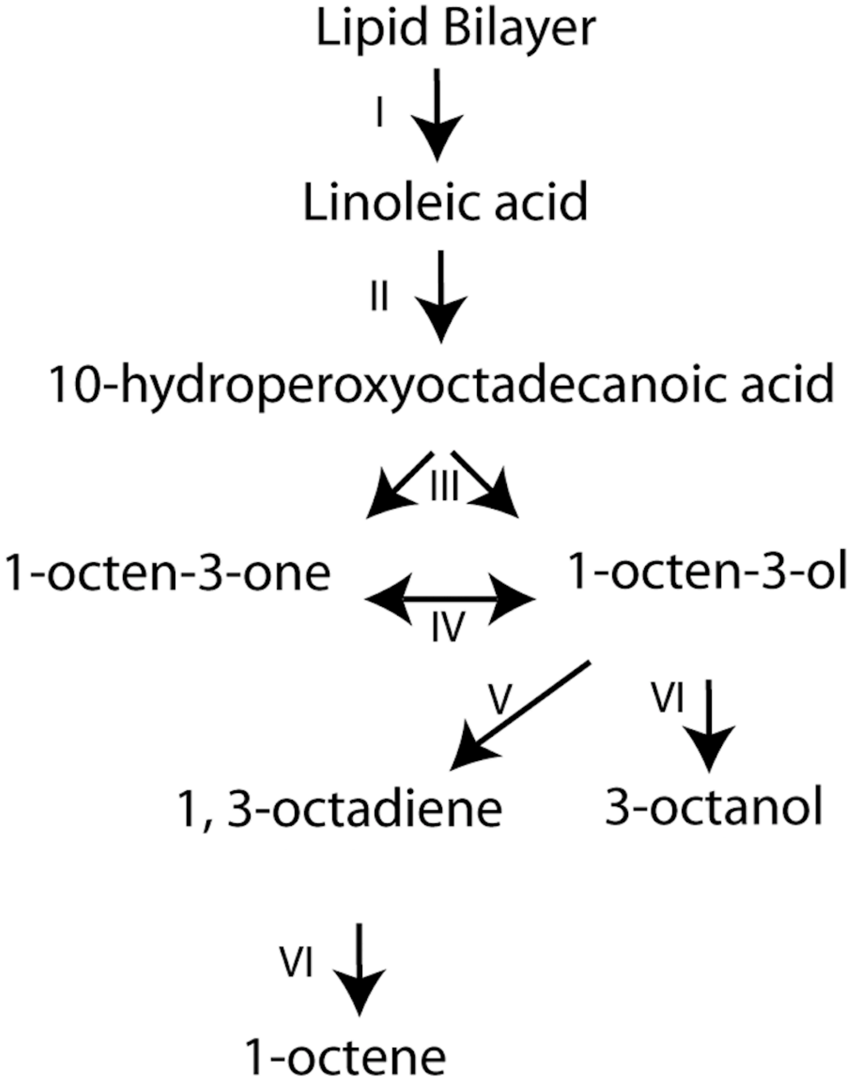


Figure 5.4. Proposed *A. sarcoides* 50072 1-octene pathway. The hypothesised pathway was based on gene expression levels that correlates with metabolic observation Gianoulis et al. (2012). Numbers I-VI refer to the enzymes listed in Table 5.4.

Table 5.4. BLAST and HHpred annotation of the proposed *A. sarcoides* 50072 1-octene pathway. Two method were used to identify the annotations proposed by Gianoulis et al. (2012). Both BLAST and HHpred suggests the proposed pathway do not have the biochemical potential required for 1-octene biosynthesis.

Hypothetical pathway position	Accession	BLASTP vs NR and Uniprot	HHpred vs *S. cerevisiae*
I	10218	PQ-loop motif hypothetical protein	Permease/transporter
II	2804	snRNP/formin like protein	Pre-mRNA-processing protein PRP40; FF domain, Prp40, *Saccharomyces cerevisiae*
II	3405	Hypothetical/ Uncharacterised protein	Histone chaperone RTT106
III	9537	ABC Type 2/P-loop NTPase	Iron-sulfur clusters transporter ATM1, mitochondrial; ABC transporter
IV	1593	Hypothetical/ Uncharacterised protein	No significant hits
IV	4820	Dephospho kinase/ CoaE-domain-containing protein	URIDYLATE KINASE (E.C.2.7.4.-) COMPLEXED WITH; TRANSFERASE; HET: ADP, AMP
IV	5565	ATP-dependent metallopeptidase Hfl	Proteasome subunit alpha type-1 (E.C.3.4.25.1); 26S Proteasome, ATPase, AAA+, Protease

Table 5.4 (continued). BLAST and HHpred annotation of the proposed *A. sarcoides* 50072 1-octene pathway. Two methods were used to identify the annotations proposed by Gianoulis et al. (2012). Both BLAST and HHpred suggest the proposed pathway do not have the biochemical potential required for 1-octene biosynthesis.

Hypothetical pathway position	Accession	BLASTP vs NR and Uniprot	HHpred vs *S. cerevisiae*
V	4033	Hypothetical/ Uncharacterised protein	No significant hits
V	9422	Serine/threonine-protein phosphatase 6 regulatory ankyrin repeat subunit B	Regulatory protein SWI6; transcriptional, Ankyrin repeats
VI	5457	putative mitochondrial import inner membrane translocase subunit	Mitochondrial import inner membrane translocase
VI	8716	No significant hits	Ubiquinol--cytochrome-c reductase subunit, Ubiquinol--cytochrome-c reductase; Respiratory chain

Reconstructing the hypothesised pathway

Assuming a potential mislabelling as the reason for discrepancies between our results and the reported activity, we endeavoured to reconstruct their hypothesised pathway *in silico* by elucidating similar proteins with the proposed activity in all *A. sarcoides* isolates including NRRL 50072. For this analysis, we exclude any genes that do not form orthologous clusters with proteins from other isolates and exclude any proteins that were not found in NRRL 50072, with either our reciprocal network analysis or Orthovenn. This is because for this hypothesised pathway to hold true, this alkane pathway must be present in all isolates of *A. sarcoides*.

Step II

To reconstruct the hypothetical pathway, we first assumed that there are already pathways for free fatty acids within *A. sarcoides*. For step II (Table 5.5), two clusters were found to annotate for PPO-type activity. Both RBHcluster_3513 and RBHcluster_6191 are annotated to be Psi-producing oxygenase A. Furthermore, RBHcluster_3513 is supported by three annotations, each annotation further supports the cluster to annotate for PPO. By using the ontology that described RBHcluster_3513 and RBHcluster_6191, three further clusters were found to have PPO-like activity (RBHcluster_6513, RBHcluster_6008, and RBHcluster_1861.

Table 5.5. Orthologous protein clusters that are annotated for the hyper oxidation of linoleic acid. This pathway is able to oxidise linoleic acid, producing a hyperoxide lipid. Results indicate the annotation evidence and *A. sarcoides* annotation attached to the cluster.

Reciprocal network analysis					
Cluster	Highest E-value	Average percentage ID (%)	Putative annotation	Annotation evidence	*A. sarcoides* protein accession
RBHcluster_ 3513	0	98.3	PPOA	Psi-producing oxygenase A (K17863) *A. nidulans* Psi-producing oxygenase A (G5EB19) *N. fumigata* Psi-producing oxygenase A (B0Y6R2)	AS17056_P7269 AS17156_P5110 AS24680_P7769 AS30971_P1173 AS44013_P7454 AS64019_P2843
RBHcluster_ 6191	0	98.8	PPOA	*A. nidulans* Psi-producing oxygenase A (Q6RET3)	AS17056_P4014 AS17156_P5532 AS24680_P10039 AS30971_P4931 AS44013_P7243 AS64019_P6377

Table 5.5 (continued). Orthologous protein clusters that are annotated for the hyper oxidation of linoleic acid. This pathway is able to oxidise linoleic acid, producing a hyperoxide lipid. Results indicate the annotation evidence and *A. sarcoides* annotation attached to the cluster.

Reciprocal network analysis			Putative annotation	Annotation evidence	*A. sarcoides* protein accession
Cluster	Highest E-value	Average percentage			
RBHcluster_ 6513	0	99.1	Manganese Lipoxygenase	*F. oxysporum* Manganese lipoxygenase	AS17056_P9966 AS17156_P8869 AS24680_P5888 AS30971_P5632 AS44013_P7120 AS64019_P193
RBHcluster_ 6008	0	96.2	Linoleate 9S-lipoxygenase 1	*Oryza sativa* Linoleate 9S-lipoxygenase 1	AS17056_P158 AS17156_P6853 AS24680_P5590 AS30971_P559 AS44013_P5608 AS64019_P6660
RBHcluster_ 1861	0	98.2	Linoleate 9S-lipoxygenase 1	*Oryza sativa* Linoleate 9S-lipoxygenase 1	AS17056_P3044 AS17156_P3954 AS24680_P1436 AS30971_P6893 AS44013_P8337 AS64019_P3207

Step III

For lysing the hydroperoxide lipid in Step III (Table 5.6), two clusters (RBHcluster_8002 and RBHcluster_1013) and a singlet (RBHcluster_12720) were annotated for this function. RBHcluster_8002 and RBHcluster_12720 are able to form orthologous clusters with the OrthoVenn2 pipeline. Furthermore, OrthoVenn2 also annotated these clusters linolenate hydroperoxidase. However, the BLAST pipeline also annotated these clusters to have function in hyphae growth with a monooxygenase activity. For RBHcluster_1013, annotations from different pipelines did not come to an agreement. FGSCa4 BLAST annotation suggested this cluster to be a Hydroperoxide monoxygenase, whereas both OrthoVenn2 and S288c BLAST annotation pipeline annotate this cluster to have a monothiol forming activity.

Table 5.6. Orthologous protein clusters that are annotated for lysing hyperoxidise lipds. This pathway is able to cleave linoleic acid, producing oxygenated acyl compounds such as aldehydes and dicarboxylic acid. Results indicate the annotation evidence and *A. sarcoides* annotation attached to the cluster.

Reciprocal network analysis			Putative annotation	Annotation evidence	*A. sarcoides* protein accession
Cluster	Highest E-value	Average percentage			
RBHcluster_ 8002	0	95.8	Hyperoxidase lyase	AhbB, monooxygenase activity and hyphae growth function (C8VF00) *A. thaliana* Probable inactive linolenate hydroperoxide lyase (B3LF83)	AS17056_P6748 AS17156_P4044 AS24680_P3996 AS44013_P8034 AS64019_P6684
RBHcluster_ 12720	-	-	Hyperoxidase lyase	AhbB, monooxygenase activity and hyphen growth function (C8VF00) *A. Thaliana* Probable inactive linolenate	AS30971_P8263
RBHcluster_ 1861	0	99.8	Hyperoxide/ Monothiol glutaredoxin	*A. nidulans* Hydroperoxide and superoxide-radical responsive, glutathione-dependent oxidoreductase (Q5AVW3) *S. cerevisiae* Monothiol glutaredoxin-4 (P32642)	AS17056_P3044 AS17156_P3954 AS24680_P1436 AS30971_P6893 AS44013_P8337 AS64019_P3207

Step IV

For 3-oxoacyl reductases in Step IV (Table 5.7), annotations that were localised in the mitochondria or were annotated to utilise acyl-ACP were subtracted. This revealed four clusters that are annotated for peroxisomal and microsomal 3-oxoacyl reductases. Of this, two clusters (RBHcluster_6464 and RBHcluster_7557) were annotated for long chain fatty acid elongation. Furthermore, the one remaining cluster (RBHcluster_2268) was annotated to be 3-oxoacyl-CoA reductase localised to the microsomes, which have the required activity to catalyse step IV of the pathway.

Table 5.7. Orthologous protein clusters that are annotated for 3-oxoacyl reductase (FabG). This pathway is able to reduce 3-hydroxyacyl-ACP to 3-oxoacyl-ACP. Results indicate the annotation evidence and *A. sarcoides* annotation attached to the cluster.

Reciprocal network analysis			Putative annotation	Annotation evidence	*A. sarcoides* protein accession
Cluster	Highest E-value	Average percentage			
RBHcluster_ 6464	0	99.1	3-oxoacyl reductase	*B. taurus* ELOVL7, Elongation of very long chain fatty acids protein (A0JNC4)	AS17056_P425 AS17156_P793 AS24680_P7977 AS30971_P9261 AS44013_P8256 AS64019_P4868
RBHcluster_ 7557	0	94.3	3-oxoacyl reductase	*D. discoideum* Putative elongation of fatty acids protein (Q86JM5)	AS17056_P8845 AS17156_P9507 AS24680_P4402 AS30971_P5798 AS44013_P232 AS64019_P6756
RBHcluster_ 2268	0	99.6	very-long-chain 3-oxoacyl-CoA reductase	very-long-chain 3-oxoacyl-CoA reductase (K10251) *B. fuckeliana* Very-long-chain 3-oxoacyl-CoA reductase (A6SG70)	AS17056_P1991 AS17156_P6841 AS24680_P4011 AS30971_P1496 AS44013_P1212 AS64019_P2289

Step V

Step V of the hypothesised pathway requires a dehydratase. Only one cluster (RBHcluster_3076) was annotated for this step (Table 5.8). RBHcluster_3076 is annotated as a 3-hydroxyacyl-CoA dehydratase, which is capable of dehydrating hydroxyl on the C3 octane backbone to an unsaturated bond.

Table 5.8. Orthologous protein clusters that are annotated for long chain 3-hydroxyacyl-CoA dehydratase. This pathway is able to reduce 3-hydroxyacyl-ACP to 3-oxoacyl-ACP. Results indicate the annotation evidence and *A. sarcoides* annotation attached to the cluster.

Reciprocal network analysis					
Cluster	Highest E-value	Average percentage ID (%)	Putative annotation	Annotation evidence	*A. sarcoides* protein accession
RBHcluster_ 3076	0	99.5	Long chain 3-hydroxyacyl-CoA dehydratase	very-long-chain (3R)-3-hydroxyacyl-CoA dehydratase (K10703) *S. pombe* very-long-chain (3R)-3-hydroxyacyl-CoA dehydratase (O14346)	AS17056_P425 AS17156_P793 AS24680_P7977 AS30971_P9261 AS44013_P8256 AS64019_P4868

Step VI

Table 5.9. Orthologous protein clusters that are annotated for Trans-enoyl reductase. This pathway is able to reduce unsaturated bonds. Results indicate the annotation evidence and *A. sarcoides* annotation attached to the cluster.

Reciprocal network analysis			Putative annotation	Annotation evidence	*A. sarcoides* protein accession
Cluster	Highest E-value	Average percentage ID (%)			
RBHcluster_ 5589	0	99.2	trans-enoyl reductase	*G. zeae* Trans-enoyl reductase FSL5 (A0A0E0RXA7) *A. nidulans* Trans-enoyl reductase apdC (Q5ATH1)	AS17056_P484 AS17056_P7334 AS17156_P211 AS17156_P8079 AS24680_P1148 AS24680_P2164 AS30971_P7017 AS30971_P7550 AS44013_P7864 AS44013_P887 AS64019_P409 AS64019_P4618
RBHcluster_ 373	0	99.4	trans-enoyl reductase	*A. nidulans* Dehydrogenase orsE (Q5AUW6) *A. japonica* Trans-enoyl reductase himH (A0A2Z5TIQ0)	AS17056_P2805 AS17156_P7000 AS24680_P2556 AS30971_P3516 AS44013_P7537 AS64019_P907
RBHcluster_ 45	0	98.5	trans-enoyl reductase	*M. ulcerans* Trans-acting enoyl reductase (A0PQ21)	AS17056_P2892 AS17156_P7776 AS24680_P8643 AS44013_P2926 AS64019_P3452
RBHcluster_ 11927	-	-	trans-enoyl reductase	*M. ulcerans* Trans-acting enoyl reductase (A0PQ21)	AS30971_P4908
RBHcluster_ 9887	0	93.4	trans-enoyl reductase	*A. flavus* Trans-enoyl reductase lepG (B8NJH1)	AS17056_P869 AS24680_P9273 AS44013_P4828
RBHcluster_ 9722	0	92.3	trans-enoyl reductase	*A. flavus* Trans-enoyl reductase lepG (B8NJH1)	AS17156_P6189 AS30971_P7569 AS64019_P4572

Step VI requires an enoyl reductase to covert the unsaturated bond to a a saturated C3/C4 bond. Nine clusters and one singlet were detected to have an annotation for enoyl reduction (Table 5.9). RBHcluster_5589 is an orthologous cluster that contains gene duplication event. This cluster represents two highly similar copies of enoyl reductase in each isolate. RBHcluster_9887 and RBHcluster_9722 are orthologous clusters that each contain proteins from isolate 170.56, 246.80, and 44013 and 171.56, 309.71, and 64019 respectively. OrthoVenn2 was able to group these into its own cluster. Of the clusters mentioned, all clusters were involved in either secondary metabolite or steroid biosynthesis. No satisfactory consensus can be formed for an annotation that is relevant to Step VI.

Table 5.9 (continue). Orthologous protein clusters that are annotated for Trans-enoyl reductase. This pathway is able to reduce unsaturated bonds. Results indicate the annotation evidence and *A. sarcoides* annotation attached to the cluster.

Reciprocal network analysis					
Cluster	Highest E-value	Average percentage ID (%)	Putative annotation	Annotation evidence	*A. sarcoides* protein accession
RBHcluster_ 4707	0	99.3	trans-enoyl reductase	*M. tuberculosis* Putative trans-acting enoyl reductase (O53176) *A. nidulans* Putative NADP domain protein (Q5AWN7)	AS17056_P505 AS17156_P245 AS24680_P2143 AS30971_P7024 AS44013_P7862 AS64019_P4612
RBHcluster_ 1729	0	99.4	di-enoyl reductase	Oxidoreductase, short-chain dehydrogenase/ reductase family (Q5AVB0) peroxisomal 2,4-dienoyl-CoA reductase (K13237) *S. cerevisiae* Peroxisomal 2,4-dienoyl-CoA reductase (P32573)	AS17056_P10106 AS17156_P8473 AS24680_P1196 AS30971_P3545 AS44013_P1946 AS64019_P4374
RBHcluster_ 5982	0	99.7	trans-enoyl reductase	*G. zeae* Trans-enoyl reductase FSL5 (A0A0E0RXA7) *A. nidulans* Trans-enoyl reductase apdC (Q5ATH1)	AS17056_P7334 AS17156_P8079 AS24680_P1148 AS30971_P7550 AS44013_P887 AS64019_P409

Table 5.9 (continue). Orthologous protein clusters that are annotated for Trans-enoyl reductase. This pathway is able to reduce unsaturated bonds. Results indicate the annotation evidence and *A. sarcoides* annotation attached to the cluster.

Reciprocal network analysis					
Cluster	Highest E-value	Average percentage ID (%)	Putative annotation	Annotation evidence	*A. sarcoides* protein accession
RBHcluster_ 1592	0	98.8	trans-enoyl reductase	very-long-chain enoyl-CoA reductase (K10258) *S. pombe* Putative enoyl reductase (O94511) *S. cerevisiae* very-long-chain enoyl-CoA reductase (Q99190) *A. nidulans* Steroid alpha reductase family protein (Q5BBW2)	AS17056_P7496 AS17156_P8315 AS24680_P8800 AS30971_P1757 AS44013_P1406 AS64019_P731

5.2.4 Terpene and sesquiterpene biosynthesis

In all six *A. sarcoides* isolates, with a text-mining approach of KEGG annotations, we observed the presence of 17 genes that participate in terpenoid backbone biosynthesis. Out of 17 genes, six of these genes cover the entirety of the mevalonate pathway (MVA). These are crucial in the generation of isomeric isoprene monomers, such as dimethylallyl pyrophosphate (DMAPP) and isopentenyl pyrophosphate (IPP). For MVA, acetyl-CoA is the primary substrate. From our annotation we can infer the following MVA pathway: two units of acetyl-CoA are condensed by acetyl-CoA acetyltransferase (KEGG: K00626) to form a single unit of acetoacetyl-CoA. Reciprocal network analysis has revealed two clusters (RBHcluster_479 and RBHcluster_2584) that are annotated for this function (Table 5.10). Furthermore, all annotation pipelines suggest that both clusters are acetyl-CoA acetyltransferase. This reveals two isoforms of acetyl-CoA acetyltransferase for each isolate. The acetoacetyl-CoA is condensed with a single unit of acetyl-CoA by a hydroxymethylglutaryl-CoA synthase (KEGG: K01641) to form a single unit of 3-hydroxy-3-methyl-glutaryl-CoA. For this step, only RBHcluster_1226 is annotated for this function indicating that it is present in all six isolates and suggesting that there is a single isoform of this enzyme (Table 5.10). The single unit of 3-Hydroxy-3-methyl-glutaryl-CoA undergoes reduction, consuming two units of NADP+, leading to CoA acylation by an oxidoreductase, hydroxymethylglutaryl-CoA reductase (KEGG: K00021), generating a single unit of mevalonate. RBHcluster_340, indicates the presence of this protein sequence in all six isolates (Table 5.10). Mevalonate undergoes two independent cycles of phosphorylation. The first phosphorylation is achieved by mevalonate kinase (KEGG: K00869) to generate a phosphomelavonate and the second phosphorylation is conducted by phosphomevalonate kinase (KEGG: K00938) to generate a diphosphatemevalonate, Both steps are clustered respectively into RBHcluster_2932 and RBHcluster_3302 (Table 5.10). The presence of these proteins was observed and confirmed in each isolate. Finally, diphosphomevalonate decarboxylase (KEGG: K01597) catalyses a carboxy-lyase reaction to remove alpha carbon from a single unit of diphosphatemevalonate yielding a single unit of isopentenyl pyrophosphate. This step was annotated in RBHcluster_2089 Table 5.10, also indicating its presence in all isolates. This five-step pathway is a typical route for eukaryotic and archea lifeforms to synthesise IPP and DMAPP. Typically, the archea MVA pathway is slightly modified in its generation of melavonate and phosphorylation of melavonate. In contrast, plants contain both mevalonate independent pathway (also known as MEP) and MVA pathway for the generation of isoprenes. For isoprenoid synthesis, there are two distinct pathways. In bacteria, terpenoid synthesis is dependent on MEP to achieve isoprenoid synthesis. Despite producing DMAPP and IPP, each pathway requires different initial substrates. The MEP pathway is dependent on condensing glyceraldehyde 3-phosphate with pyruvate, with glyceraldehyde 3-phosphate a product of the glycolysis pathway. For both pathways, the end product is the formation of IPP. The generation of DMAPP is dependent on IPP delta-isomerase (KEGG: K01823) to catalyse a reversible isomerisation of IPP (Table 5.10). Both DMAPP and IPP act as substrates for downstream condensation reactions to form longer chains terpene precursors.

Table 5.10. Orthologous protein clusters that are annotated for mevalonate. This pathway is able to annotate for the complete pathway producing isoprenes by mevalonate. Results indicate the annotation evidence and *A. sarcoides* annotation attached to the cluster.

Reciprocal network analysis					
Cluster	Highest E-value	Average percentage ID (%)	Putative annotation	Annotation evidence	*A. sarcoides* protein accession
RBHcluster_ 479	0	92.3	Acetyl-CoA acetyltransferase 1	Acetyl-CoA acetyltransferase, putative (C8VM21) Acetyl-CoA acetyltransferase (P41338) acetyl-CoA C-acetyltransferase (K00626) *S. pombe* Acetyl-CoA acetyltransferase (Q9UQW6)	AS17056_P5467 AS17156_P6697 AS24680_P562 AS30971_P9894 AS44013_P1333 AS64019_P4834
RBHcluster_ 2584	0	98.7	Acetyl-CoA acetyltransferase 2	Acetyl-CoA-acetyltransferase, putative (C8V4N2) acetyl-CoA C-acetyltransferase (K00626) *B. taurus* Acetyl-CoA acetyltransferase, mitochondrial (Q29RZ0)	AS17056_P1633 AS17156_P3765 AS24680_P2892 AS30971_P4813 AS44013_P6752 AS64019_P4031
RBHcluster_ 1226	0	96.9	3-hydroxy-3-methylglutaryl coenzyme A synthase	3-hydroxy-3-methylglutaryl coenzyme A synthase (Q5B3F7) 3-hydroxy-3-methylglutaryl coenzyme A synthase (P54839) hydroxymethylglutaryl-CoA synthase (K01641) *S. pombe* Hydroxymethylglutaryl-CoA synthase, HMG-CoA synthase (P54874)	AS17056_P3690 AS17156_P4213 AS24680_P685 AS30971_P4094 AS44013_P7837 AS64019_P2785

Table 5.10 (continue). Orthologous protein cluster that are annotated for mevalonate**.** This pathway is able to annotate for the complete pathway producing isoprenes by mvelonate. Results indicate the annotation evidence and *A. sarcoides* annotation attached to the cluster.

Reciprocal network analysis					
Cluster	Highest E-value	Average percentage ID (%)	Putative annotation	Annotation evidence	*A. sarcoides* protein accession
RBHcluster_ 340	0	96.1	3-hydroxy-3-methylglutaryl coenzyme A reductase	3-hydroxy-3-methylglutaryl coenzyme A reductase (Q5B6L3) 3-hydroxy-3-methylglutaryl-coenzyme A reductase 2 (P12684) hydroxymethylglutaryl-CoA reductase (K00021) *G. fujikuroi* 3-hydroxy-3-methylglutaryl-coenzyme A reductase 2 (S0DQM8)	AS17056_P2095 AS17156_P1501 AS24680_P5405 AS30971_P4934 AS44013_P2357 AS64019_P4428
RBHcluster_ 2932	0	93.2	Mevalonate kinase	Mevalonate kinase (Q5B6G1) Mevalonate kinase (P07277) mevalonate kinase (K00869) Mevalonate kinase (Q09780)	AS17056_P2290 AS17156_P8732 AS24680_P1408 AS30971_P3444 AS44013_P9778 AS64019_P7422
RBHcluster_ 3302	0	95.5	Phosphomevalonate kinase	Phosphomevalonate kinase (Q5BAW9) Phosphomevalonate kinase (P24521) mevalonate kinase (K00938) *S. cerevisiae* Phosphomevalonate kinase *(*P24521)	AS17056_P9679 AS17156_P5915 AS24680_P8009 AS30971_P3430 AS44013_P5415 AS64019_P5826

Table 5.10 (continue). Orthologous protein cluster that are annotated for mevalonate**.** This pathway is able to annotate for the complete pathway producing isoprenes by mvelonate. Results indicate the annotation evidence and *A. sarcoides* annotation attached to the cluster.

Reciprocal network analysis					
Cluster	Highest E-value	Average percentage ID (%)	Putative annotation	Annotation evidence	*A. sarcoides* protein accession
RBHcluster_ 2089	0	99.6	Diphosphomevalonate decarboxylase	Diphosphomevalonate decarboxylase, (Q5B4W60 Diphosphomevalonate decarboxylase, (P32377) diphosphomevalonate decarboxylase (K01597) *S. pombe* Diphosphomevalonate decarboxylase (O13963)	AS17056_P6339 AS17156_P7887 AS24680_P9121 AS30971_P3326 AS44013_P9075 AS64019_P3373
RBHcluster_ 6105	0	99.1	Isopentenyl-diphosphate Delta-isomerase	Uncharacterized protein (G5EB72) Isopentenyl-diphosphate Delta-isomerase (P15496) Isopentenyl-diphosphate Delta-isomerase (K01823) *S. pombe* Isopentenyl-diphosphate Delta-isomerase (Q10132)	AS17056_P549 AS17156_P7704 AS24680_P7620 AS30971_P4557 AS44013_P3077 AS64019_P5975

Prenyltransferases (PT) are a family of enzymes that are capable of transferring prenyl groups to an acceptor molecule. Prenyl molecules can be isoprenes of any length. Acceptor molecules can be proteins, riboses, indoles, aromatics, and, crucially, other prenyls. A subset of prenyltransferases facilitate the condensation of prenyl pyrophosphates. In all *A. sarcoides* isolates we are able to annotate four sets of PT to five independent orthologous cluster (RBHcluster_3122, RBHcluster_1532, RBHcluster_779, RBHcluster_5652, and RBHcluster_4255) (Table 5.11), which annotate for prenyl elongation. PT condensation enzymes are divided into two categories that are dependent on *cis* or *trans* configurations during bond formation. The initial condensation reaction is facilitated by geranyl pyrophosphate synthase (GPPS, RBHcluster_3122). It condenses one unit of IPP with one unit of DMAPP to form geranyl pyrophosphate (GPP), an eight-carbon geranyl backbone. GPP can be further extended by condensation with a single unit of IPP. This reaction is facilitated by farnesyl pyrophosphate synthase (FPPS, RBHcluster_3122) to produce farnesyl pyrophosphate (FPP), a molecule with a 12-carbon backbone terpene. Repeated head to tail condensation reaction of terpenes continues this pattern of four carbon addition by condensing with IPP and DMAPP or with other PTs, such as GPP or FPP. The KEGG annotations suggest that *A. sarcoides* is capable of generating terpene backbones of around C24 length, based on the presence of hexaprenyl synthase.

Within our annotations, we note the presence two single copy PT genes (RBHcluster_3122 and RBHcluster_779) that are capable of catalysing analogous reactions similar to FPPS and GPPS (Table 5.11). This includes a farnesyl diphosphate synthase (RBHcluster_3122) that is capable of utilising IPP and DMAPP to generate GPP or FPP. The other is a putative hexaprenyl synthase (RBHcluster_779) that is able to accept either FPP or GPP as initial substrate to biosynthesise a hexaprenyl diphosphate. This suggests that PTs are capable of promiscuous activity with regards to substrate activity. In *A. thaliana*, radio-labelling assays indicate that, with a range of prenyl substrates ranging from C5 to C20, a GPPS encoded gene product is capable of generating GPP and other prenyl-pyrophosphate ranging from C20 to C40 (Hsieh et al., 2011). This study concluded that if sufficient IPP is provided, atGPP is able to continue the enzymatic reaction producing short/medium chain prenyl pyrophosphate (Hsieh et al., 2011). This enzymatic promiscuity is thought to be facilitated by a long hydrophobic tunnel observed in the crystal structure of atGPP (Hsieh et al., 2011). Moreover, the mechanism for chain length determination in *cis*-type isopentyl pyrophosphate synthase has been resolved. As mentioned previously, the hydrophobic tunnel is key to the length of product. The hydrophobic tunnel is formed by two α-helices and four β-strands. In *E. coli* undecaprenyl pyrophosphate synthase (UPPS), the bottom of this tunnel is lined with hydrophobic residues such as isoleucine, leucine, valine and histidine. When the leucine 137 residue in this part of the protein is replaced with an alanine, a smaller hydrophobic residue, the enzyme is capable of synthesising prenyl products that are longer than wild type variant. Substrate specificity can also be manipulated. *M. tuberculosis* possess a relatively short chain FPP enzyme. This is important as most *cis*-PT have products over C55. It is thought the formation of shorter prenyl product is due to the presence of leucine 84 and valine 156. These residues correspond to Alanine 69 and Alanine 143 in *E. coli* UPPS. When alanine 69 in *E. coli* UPPS was substituted with a leucine residue, it leads to greater formation of C30 intermediates. This suggest that substitution of the size of leucine residues interferes with chain elongation by blocking the hydrophobic channel. Despite conducting the same catalytic activity, reports suggest that prenyltransferase shares only 30% conserved amino acids, though structural studies indicate that PT are conserved on a structural level and conserved DD(*X*)*n*D motifs for substrate and cofactor binding. Based on these findings, it is hypothesised that *A. sarcoides* possesses PTs that are catalytically promiscuous. This is because prenyl condensations are ubiquitous throughout all forms of life and share common catalytic condensation reactions. KEGG also annotates a gene as a putative NUS1 (KEGG: K19177, Cluster: RBHcluster_4255) in all isolates of *A. sarcoides*. NUS1 is thought to encode a long chain PT, the dehydrodolichyl diphosphate synthase complex. This implies that all isolates are capable of producing long chain terpene pyrophosphates that are typically over C50 in length. In yeast, NUS1 gene affects survival and the function of N-linked glycosylation.

Table 5.11. Orthologous protein clusters that are annotated for Prenyl Transferases. Prenyl transferases are enzymes that are capable of head to head condensation of prenyls. Results indicate the annotation evidence and *A. sarcoides* annotation attached to the cluster.

Reciprocal network analysis					
Cluster	Highest E-value	Average percentage ID (%)	Putative annotation	Annotation evidence	*A. sarcoides* protein accession
RBHcluster_ 3122	0	97.9	farnesyl diphosphate synthase	*A. nidulans* Polyprenyl synthetase (Q5AUL8) *S. cervisiae* Farnesyl pyrophosphate synthase (P08524) farnesyl diphosphate synthase (K00787) *G. kujikuroi* Farnesyl pyrophosphate synthase *(*Q92235)	AS17056_P8206 AS17156_P2461 AS24680_P7068 AS30971_P6429 AS44013_P7266 AS64019_P2087
RBHcluster_ 1532	0	99.2	Geranylgeranyl diphosphate synthase	*A. nidulans* Geranylgeranyl diphosphate synthase (Q5BFM6) *S. cerevisiae* Geranylgeranyl diphosphate synthase (Q12051) geranylgeranyl diphosphate synthase, type III (K00804) *G. fujikeroi* Geranylgeranyl diphosphate synthase (Q92236)	AS17056_P8672 AS17156_P7808 AS24680_P7953 AS30971_P6274 AS44013_P3230 AS64019_P3071

Table 5.11 (continue). Orthologous protein clusters that are annotated for Prenyl Transferases. Prenyl transferases are enzymes that are capable of head to head condensation of prenyls. Results indicate the annotation evidence and *A. sarcoides* annotation attached to the cluster.

Reciprocal network analysis			Putative annotation	Annotation evidence	*A. sarcoides* protein accession
Cluster	Highest E-value	Average percentage ID (%)			
RBHcluster_779	0	98.9	Hexaprenyl pyrophosphate synthetase	Hexaprenyl pyrophosphate synthetase (C8VJN5) *S. cerevisiae* Hexaprenyl pyrophosphate synthetase, mitochondrial (P18900) hexaprenyl-diphosphate synthase (K05355) Probable hexaprenyl pyrophosphate synthase, mitochondrial (Q7S565)	AS17056_P419 AS17156_P3696 AS24680_P882 AS30971_P8164 AS44013_P5986 AS64019_P9276

Table 5.11 (continue). Orthologous protein clusters that are annotated for Prenyl Transferases. Prenyl transferases are enzymes that are capable of head to head condensation of prenyls. Results indicate the annotation evidence and *A. sarcoides* annotation attached to the cluster.

Reciprocal network analysis					
Cluster	Highest E-value	Average percentage ID (%)	Putative annotation	Annotation evidence	*A. sarcoides* protein accession
RBHcluster_ 4255	0	99.6	Dehydrodolichyl diphosphate synthase	*A. nidulans* Uncharacterized protein (Q5AS52) *S. cerevisiae* Dehydrodolichyl diphosphate synthase complex subunit NUS1 (Q12063) dehydrodolichyl diphosphate syntase complex subunit NUS1 (K19177) *S. pombe* Dehydrodolichyl diphosphate synthase complex subunit nus1 Q9Y7K8)	AS17056_P1352 AS17156_P9310 AS24680_P7561 AS30971_P6358 AS44013_P8150 AS64019_P8017
RBHcluster_ 5652	0	98.3	Dehydrodolichyl diphosphate synthase	*A. nidulans* Prenyltransferase, putative (C8VG05) *S. cerevisiae* Dehydrodolichyl diphosphate synthase complex(P35196) ditrans,polycis-polyprenyl diphosphate synthase (K11778) *S. pombe* Dehydrodolichyl diphosphate synthase complex subunit SPAC4D7.04c (O14171)	AS17056_P5597 AS17156_P8864 AS24680_P2194 AS30971_P8356 AS44013_P6886 AS64019_P2419

5.2.5 Alkane Degradation

In chapter 3, we observed that all six *A. sarcoides* isolate were capable of utilising tetradecane when it is the sole carbon source. Here, we elucidate the genes responsible for an alkane degradation pathway. Alkane assimilation requires the alkane molecule to be oxidised. The first step hydroxylates the terminal methyl group of an alkane and converts it to an equivalent length alcohol. In *A. sarcoides,* one protein cluster containing six proteins (RBHcluster_4080) and a singlet (RBHcluster_10974) was annotated to be a cytochrome P450 ALK1 (Table 5.12). Moreover, a cluster of five proteins (RBHcluster_7817) and a singlet (RBHcluster_12726) were also annotated as a Cytochrome p450 ALK2-A (Table 5.11). Phylogenetic analysis was conducted to assess the annotations to other well characterised terminal hydroxylase and omega-monooxygenase (Figure 5.5).

Two further oxidation steps are required to convert the alkanol to an equivalent length aldehyde and fatty acids respectively. Oxidation of alkanol to an aldehyde is catalysed by an alcohol dehydrogenase. Combining text-mining and KEGG's pathway annotation revealed the conversion of alcohol to aldehyde is encoded by 57 different clusters or singlets and is found across all isolates of *A. sarcoides*. This led us to propose a pathway from these observations (Figure 5.6).

Table 5.12. Orthologous protein clusters that are annotated for alkane monooxygenase. Alkane monooxygenases are P450 enzymes that catalyse the oxidation of alkane molecules. Ultimately, the alkane molecule is oxidised to a fatty acid of the same length to be assimilated by beta-oxidation pathway. Text mining has reported two major forms of alkane monooxygenase in *A. sarcoides*.

Reciprocal network analysis			Putative annotation	Annotation evidence	*A. sarcoides* protein accession
Cluster	Highest E-value	Average percentage			
RBHcluster_ 4080	0	92.3	P450 ALK2A	Cytochrome P450, putative (Q5B6B3) *C. maltosa* Alkane-inducible P450-ALK2-A (Q12581)	AS17056_P642 AS17156_P6273 AS24680_P8919 AS30971_P1383 AS44013_P3452 AS64019_P4509
RBHcluster_ 10974	-	-	P450 ALK2A	Cytochrome P450, putative (Q5B6B3) *C. maltosa* Alkane-inducible P450-ALK2-A (Q12581)	AS24680_P10122
RBHcluster_ 7817	0	98.6	P450 ALK	Cytochrome P450, putative (Q5B6B3) *C. tropicalis* Cytochrome P450 52A1 (P10615)	AS17056_P2976 AS17156_P514 AS24680_P4503 AS44013_P180 AS64019_P9072
RBHcluster_ 12726	-	-	P450 ALK	Cytochrome P450, putative (Q5B6B3) *C. tropicalis* Cytochrome P450 52A1 (P10615)	AS30971_P8296

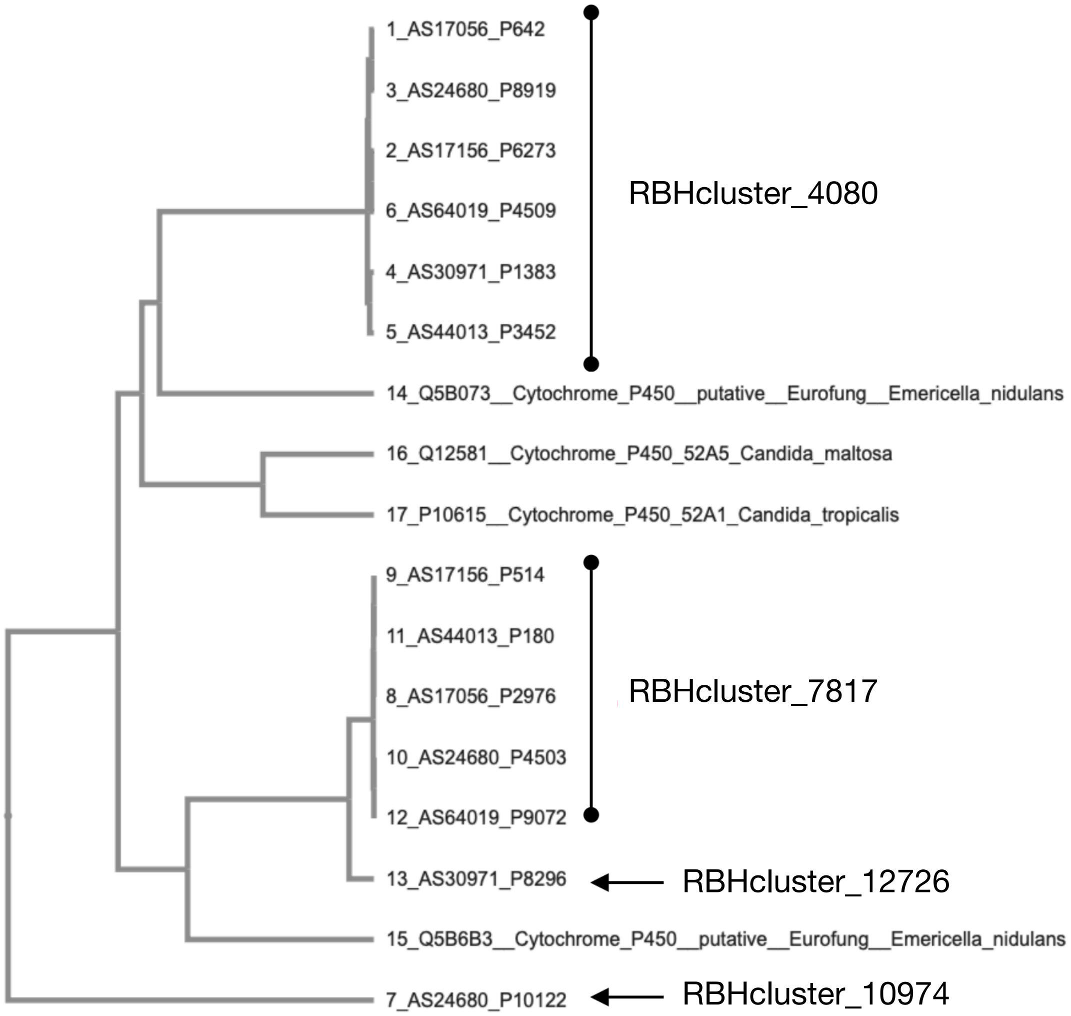


Figure 5.5. Phylogenetic tree of putative clusters encoding for ALK. Phylogeny indicated that there are three orthologs that encodes alkane degradation. MAFFT was used to analysed the amino acid sequence sagainst characterised P450 ALK.

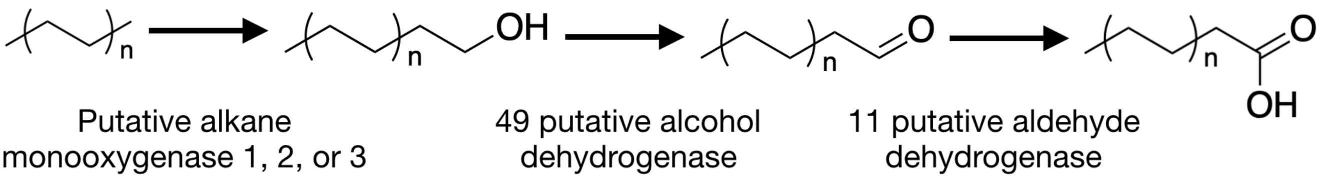

n
OH
n
O
n
O
n
OH
Putative alkane monooxygenase 1, 2, or 3
49 putative alcohol dehydrogenase
11 putative aldehyde dehydrogenase

Figure 5.6. Proposed pathway for alkane degradation in *A. sarcoides*.

5.2.6 Fatty acid biosynthesis

Fatty acid synthesis is facilitated by two different types of fatty acid synthases (FAS). Type I is responsible for fatty acid biosynthesis in non-plant eukaroytes and certain prokaryotes and is facilitated by multi-enzymes. For plants and prokaryotes, biosynthesis occurs via FAS II. Discrete enzymes catalyse each catalytic step in FAS Type II. In fungi, the biosynthesis of fatty acids is facilitated by Type I system although the mitochondria of eukaryotes are also capable of synthesising fatty acids by Type II system.

The Type I system is composed of two subunits: the alpha subunit (FAS2) and beta subunit (FAS1). In *S. cerevisiae*, the six subunits of FAS1 and six subunits of FAS2 complex together to form a heterododecameric complex with 8 active sites. Fungal type I FAS are capable of four major cyclical reactions to initiate fatty acid biosynthesis and its continual condensation: ketoacyl synthase (KS), ketoacyl reductase (KR), hydroxyacyl dehydratase (HD) and enoyl reductase (ER). This FAS system self elongates chain up to a chains length of C16. In the dataset of all *A. sarcoides* isolates, the annotation pipeline elucidated predicted proteins within RBHcluster_6305 to be FAS1 and RBHcluster_4001 to be FAS2 (Table 5.13). Further evidence by three independent annotation pipelines support this annotation. No other significant reciprocal hits were detected for either cluster. This indicates that the annotations agree with the well characterised type I FAS in *S. cerevisiae*.

Like other ascomycetes, *A. sarcoides*' Type I system is complemented by a fatty acid elongation (FAE) system. In *S. cerevisiae* ELO1, ELO2 and ELO3 are responsible for ketoacyl synthase reaction in FAE. These ketoacyl synthase enzymes initiates elongation for medium-chain fatty acids (C16-18) to long chain fatty acids (C20-30) fatty acids by catalysing a single unit of maloynyl-CoA with a single unit of a growing chain acyl-CoA which yield a (n+2) 3-ketoacyl-CoA and a CO_2. Across six isolates of *A. sarcoides*, RBHcluster_649 was annotated to be an ELO-like protein (Table 5.14). Initial BLAST analysis revealed proteins within RBHcluster_649 to significantly align with *S. cerevisiae* ELO1, ELO2 and ELO3. Further reciprocal BLAST analysis was able to confirm all proteins within RBHcluster_649 aligned significant to only ELO2. No other *S. cerevisiae* ELO protein returned a significant reciprocal hit against proteins from *A. sarcoides*. Moreover, further annotation evidence supports the conclusion that RBHcluster_649 is an ELO2-like protein. When the proteins from

RBHcluster_649 were analysed by BLAST against proteome sets that contain these proteins, it revealed reciprocal matches with proteins from clusters RBHcluster_6464 and RBHcluster_7557 (Table 5.15). Additionally, RBHcluster_6464 and RBHcluster_7557 proteins are reciprocal to each other and no other significant reciprocal matches were reported. RBHcluster_6464 and RBHcluster_7557 were annotated to be putative homologs of *A. nidulans'* FAE protein. Evidence from OrthoVenn2 annotation pipeline described the clusters respectively to be a putative 3-ketoacyl acyl-CoA synthase FAE from *Dictyostelium discoideum* and a FAE protein (ELOVL7) from *Bos taurus*. Further BLAST analysis and text mining revealed no other FAE-like proteins. Phylogenetic trees of the *A. sarcoides* FAE-like proteins with the known annotation sequence (Figure 5.7) revealed RBHcluster_649 to be evolutionarily closest to ELO-type FAE and other ELO2-like FAE proteins. While RBHcluster_6464 and RBHcluster_7557 form clades that are closer related to a putative *A. nidulans* FAE protein (Uniprot: Q5BEP9). This evidence supports the hypothesis that RBHcluster_6464 and RBHcluster_7557 are orthologous and paralogous towards RBHcluster_649. It is therefore possible for RBHcluster_6464 and RBHcluster_7557 to have functions beyond the elongation of mid-chain fatty acids to very long chain fatty acids.

The annotation pipelines and reciprocal network analysis were able to annotate Type II like enzymes that are thought to contribute to fatty acid biosynthesis. FAS Type II proteins function as discrete mitochondrial enzymes, catalysing the individual steps for the biosynthesis of fatty acids, unlike the FAS Type I proteins. In the first step of FAS Type II, a ketoacyl synthase is required to condense a single unit of acetyl-CoA with a malonyl-CoA. This step is similar, in terms of catalytic activity, to the aforementioned ELO and ELO-like proteins. Initial searches for ketoacyl synthases beyond ELO-like activities did not return any significant candidates. Reviewing previous annotation for RBHcluster_6464 and RBHcluster_7557 leads to the hypothesis that both clusters may encode a Type II ketoacyl synthase activity in *A. sarcoides* (Table 5.14). The second step of fatty acid biosynthesis is the removal of the beta-keto group (the product of KS activity) by a KR to reduce it to a beta-hydroxy group. Text-mining the annotation list revealed five reciprocal orthologous clusters which encode KR activity and KEGG annotation revealed a further two reciprocal clusters of FabG-like enzymes with KR activity (Table 5.14). Of this subset, RBHcluster_6955 reciprocally aligns with a putative unreviewed *A. nidulans*

FabG-like protein. Moreover, reciprocal network analysis revealed that none of the seven putative FabG-like proteins in *A. sarcoides* are reciprocal to other clusters. When phylogeny analysis was undertaken (Figure 5.8), each cluster is able to form individual clades. Furthermore, RBHcluster_6955, RBHcluster_879, and RBHcluster_6375 were observed to be homologous to other well characterised FabG enzymes. The rest of the clusters, in contrast, are evolutionarily distant to each other, indicating that these clusters may have different functional roles. The third step in FAS Type II is the dehydration of the beta-keto group to an unsaturated bond by HD enzyme. Text-mining the annotation revealed an OrthoVenn2 annotation for RBHcluster_3076, which was previously elucidated in Table 5.8. Annotation suggests that RBHcluster_3076 is homologous to a *S. pombe* very-long-chain (3R)-3-hydroxyacyl-CoA dehydratase (O14346). Additionally, KEGG also provide annotation evidence to support RBHcluster_3076 as being a VLC 3-hydroxyacyl-CoA dehydratase (KEGG: K10703). No other annotations revealed any significant hits to a HD-type enzyme. The last step of FAS Type II is the reduction of the unsaturated bond to a saturated C-C bond, which is usually catalysed by an enoyl reductase type enzyme. Text mining was able to find 11 different putative enoyl reductases within the annotation list, again this was elucidated previously in Table 5.9. Of this, six are related to reviewed genes with secondary metabolitesbiosynthesis functions. The remaining five annotations are also reviewed, and description of the annotation strongly suggests enoyl reductase activity. A proposed pathway in Figure 5.9 details fatty acid biosynthesis in *A. sarcoides.*

Table 5.13. Orthologous protein clusters that are annotated for Type I FAS. All isolates of *A. sarcoides* contain both FAS1 and FAS2 protein which complex into a heterododecameric multi enzyme for the biosynthesis of fatty acids in eukaryotes.

Reciprocal network analysis					
Cluster	Highest E-value	Average percentage ID (%)	Putative annotation	Annotation evidence	*A. sarcoides* protein accession
RBHcluster_ 4001	0	89.7	FAS1 alpha subunit	*A. nidulans* Fatty acid synthase, alpha subunit	AS17056_P6228 AS17156_P3748
RBHcluster_ 6305	0	98.8	FAS2 beta subunit	Fatty acid synthase, beta subunit (Q5AQM2)	AS17056_P6250 AS17156_P3755

Table 5.14. Orthologous protein clusters that are annotated for long chain fatty acids elongation. In eukaryotes, FAE proteins are responsible for the formation of long chain and very long chain fatty acid elongation. From our annotation, all isolates of *A. sarcoides* contain two FAE-like proteins.

Reciprocal network analysis					
Cluster	Highest E-value	Average percentage ID (%)	Putative annotation	Annotation evidence	*A. sarcoides* protein accession
RBHcluster_7557	0	94.3	FAE	*D. discoideum* Putative elongation of fatty acids protein (Q86JM5)	AS17056_P8845 AS17156_P9507 AS24680_P4402 AS30971_P5798 AS44013_P232 AS64019_P6756
RBHcluster_6464	0	99.1	FAE	*A. nidulans* Elongation of fatty acids protein *B. taurus* ELOVL7, Elongation of very long chain fatty acids protein (A0JNC4)	AS17056_P425 AS17156_P793 AS24680_P7977 AS30971_P9261 AS44013_P8256 AS64019_P4868
RBHcluster_649	0	99.4	FAE	*A. nidulans* Elongation of fatty acids protein (Q5AUB3) fatty acid elongase 2 (K10245) *S. cerevisiae* ELO2 Elongation of fatty acids protein 2 (P25358)	AS17056_P12 AS17156_P2473 AS24680_P3064 AS30971_P4940 AS44013_P2023 AS64019_P3127

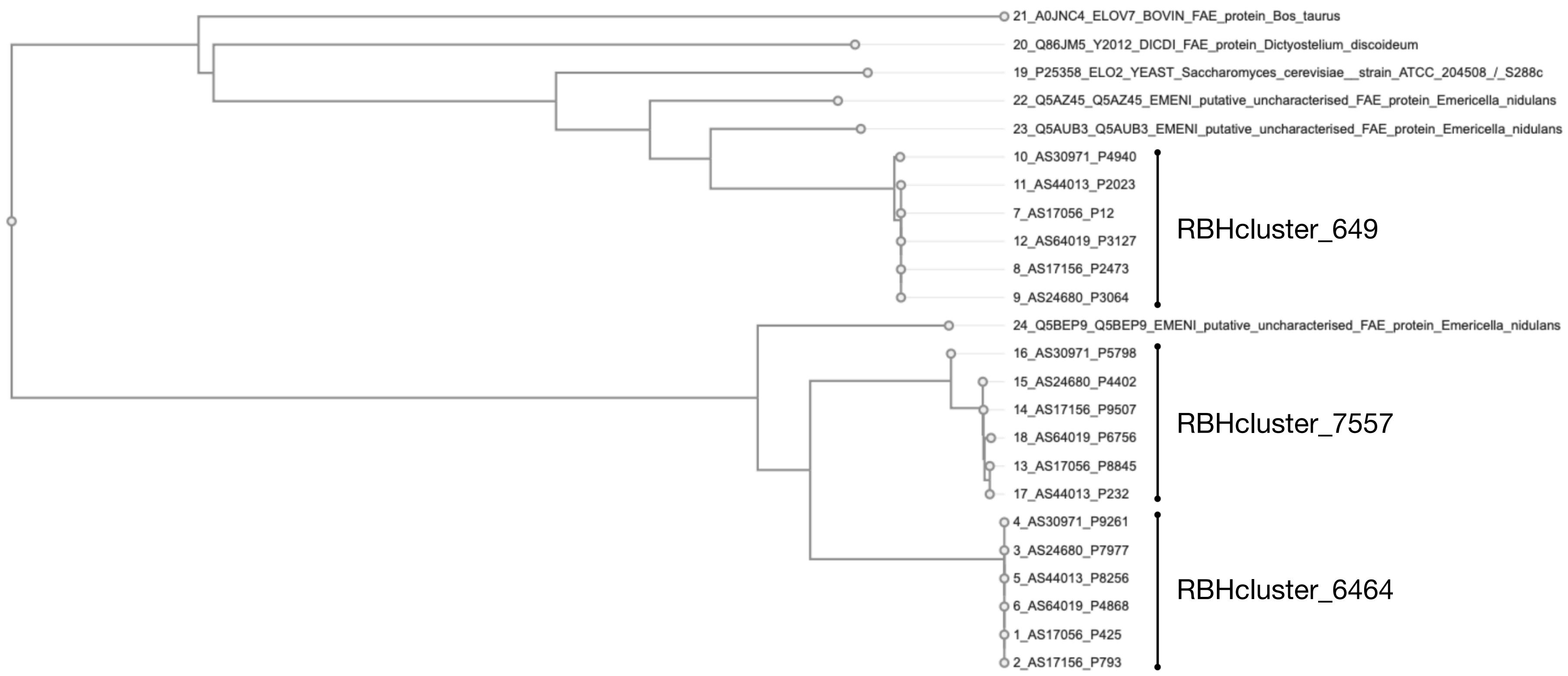

21_A0JNC4_ELOV7_BOVIN_FAE_protein_Bos_taurus
20_Q86JM5_Y2012_DICDI_FAE_protein_Dictyostelium_discoideum
19_P25358_ELO2_YEAST_Saccharomyces_cerevisiae__strain_ATCC_204508_/_S288c
22_Q5AZ45_Q5AZ45_EMENI_putative_uncharacterised_FAE_protein_Emericella_nidulans
23_Q5AUB3_Q5AUB3_EMENI_putative_uncharacterised_FAE_protein_Emericella_nidulans
10_AS30971_P4940
11_AS44013_P2023
7_AS17056_P12
12_AS64019_P3127
8_AS17156_P2473
9_AS24680_P3064
RBHcluster_649
24_Q5BEP9_Q5BEP9_EMENI_putative_uncharacterised_FAE_protein_Emericella_nidulans
16_AS30971_P5798
15_AS24680_P4402
14_AS17156_P9507
18_AS64019_P6756
13_AS17056_P8845
17_AS44013_P232
RBHcluster_7557
4_AS30971_P9261
3_AS24680_P7977
5_AS44013_P8256
6_AS64019_P4868
1_AS17056_P425
2_AS17156_P793
RBHcluster_6464

Figure 5.7. Phylogeny of FAE-like orthologous clusters with associated annotation sequence. RBHcluster_649 is the closest evolutionarily to all the known ELO-like FAE proteins, while the other two clusters more closely related to an uncharacterised putative FAE. This indicates the function may not be elongation of medium chain fatty acids for RBHcluster_7557 and RBHcluster_6464.

]

Table 5.15. Orthologous protein clusters that are annotated for FabG-like proteins. In eukaryotes, the mitochondria contains machinery for fatty acid biosynthesis. FabG encodes the 3-ketoacyl reductase, the second step in fatty acid synthesis pathway. In total seven clusters of 3-ketoacyl reductase are annotated.

Reciprocal network analysis			Putative annotation	Annotation evidence	*A. sarcoides* protein accession
Cluster	Highest E-value	Average percentage ID (%)			
RBHcluster_ 2070	0	98.5	fabG 3-oxoacyl-ACP	*A. nidulans* Uncharacterized protein (C8VG63) 3-oxoacyl-[acyl-carrier protein] reductase (K00059) *Synechocystis sp.* fabG 3-oxoacyl-[acyl-carrier-protein] reductase (P73574)	AS17056_P4060 AS17156_P501 AS24680_P4696 AS30971_P8539 AS44013_P8962 AS64019_P7407
RBHcluster_ 6955	0	95.1	fabG 3-oxoacyl-ACP	*A. nidulans* 3-oxoacyl-(Acyl-carrier-protein) reductase, putative (Q5B3E6) 3-oxoacyl-[acyl-carrier protein] reductase (K00059) *C. lanceolata* CLKR27 3-oxoacyl-[acyl-carrier-protein] reductase, chloroplastic (P28643)	AS17056_P7489 AS17156_P8295 AS24680_P8778 AS30971_P3465 AS44013_P9101
RBHcluster_ 7649	0	97.1	fabG 3-oxoacyl-ACP	*C. trachomatis* FabG 3-oxoacyl-[acyl-carrier-protein] reductase (P38004)	AS17056_P5061 AS17156_P2586 AS24680_P4839 AS30971_P8709 AS44013_P3524 AS64019_P1376

Table 5.15 (continued). Orthologous protein clusters that are annotated for FabG-like proteins. In eukaryotes, the mitochondria contains machinery for fatty acid biosynthesis. FabG encodes the 3-ketoacyl reductase, the second step in fatty acid synthesis pathway. In total seven clusters of 3-ketoacyl reductase are annotated.

Reciprocal network analysis					
Cluster	Highest E-value	Average percentage ID (%)	Putative annotation	Annotation evidence	*A. sarcoides* protein accession
RBHcluster_879	0	97.2	fabG 3-oxoacyl-ACP	*A. nidulans* Oxidoreductase, short-chain dehydrogenase/ reductase family, putative (Q5AT19) *T. maritima* FabG 3-oxoacyl-[acyl-carrier-protein] reductase (Q9X248)	AS17056_P5311 AS17156_P118 AS24680_P8124 AS30971_P1992 AS44013_P4760 AS64019_P6748
RBHcluster_7441	0	97.6	fabG 3-oxoacyl-ACP	*A. nidulans* Oxidoreductase, short-chain dehydrogenase/ reductase family (C8V6I5) *V. harveyi* FabG 3-oxoacyl-[acyl-carrier-protein] reductase (P55336)	AS17156_P9942 AS24680_P5826 AS44013_P2532 AS64019_P6440
RBHcluster_11687	-	-	fabG 3-oxoacyl-ACP	*A. nidulans* Oxidoreductase, short-chain dehydrogenase/ reductase family (C8V6I5)	AS30971_P3926

Table 5.15 (continued). Orthologous protein cluster that are annotated for FabG-like proteins. In eukaryotes, the mitochondria contains machinery for fatty acid biosynthesis. FabG encodes the 3-ketoacyl reductase, the second step in fatty acid synthesis pathway. In total seven clusters of 3-ketoacyl reductase are annotated.

Reciprocal network analysis			Putative annotation	Annotation evidence	*A. sarcoides* protein accession
Cluster	Highest E-value	Average percentage ID (%)			
RBHcluster_254	3.89E-10	96.8	fabG 3-oxoacyl-ACP	*A. nidulans* Short chain dehydrogenase/reductase family oxidoreductase, putative (C8VJX0) *C. pneumoniae* FabG 3-oxoacyl-[acyl-carrier-protein] reductase (Q9Z8P2)	AS17056_P2045 AS17156_P3853 AS24680_P3395 AS30971_P981 AS44013_P5502 AS64019_P10050
RBHcluster_6375	0	99.3	fabG 3-oxoacyl-ACP	*A. nidulans* Short chain dehydrogenase/reductase family oxidoreductase, putative (Q5BCA8) *T. maritima* FabG 3-oxoacyl-[acyl-carrier-protein] reductase (Q9X248)	AS17056_P1519 AS17156_P9216 AS24680_P4244 AS30971_P6773 AS44013_P6129 AS64019_P8312

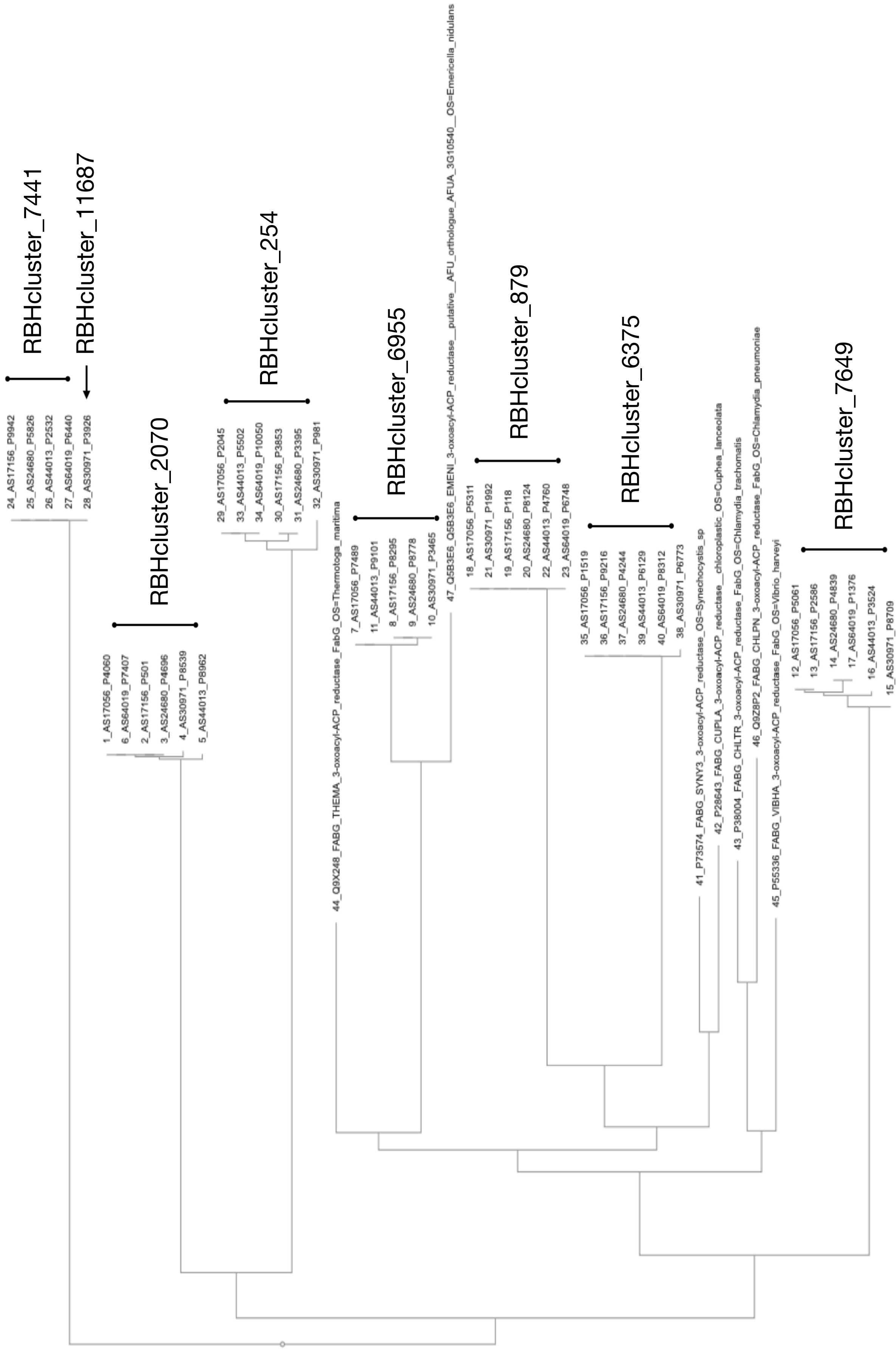

RBHcluster_7441
RBHcluster_11687
24_AS17156_P9942
25_AS24680_P5826
26_AS44013_P2532
27_AS64019_P6440
28_AS30971_P3926
RBHcluster_2070
1_AS17056_P4060
6_AS64019_P7407
2_AS17156_P501
3_AS24680_P4696
4_AS30971_P8539
5_AS44013_P8962
RBHcluster_254
29_AS17056_P2045
33_AS44013_P5502
34_AS64019_P10050
30_AS17156_P3853
31_AS24680_P3395
32_AS30971_P981
44_Q9X248_FABG_THEMA_3-oxoacyl-ACP_reductase_FabG_OS=Thermotoga_maritima
RBHcluster_6955
7_AS17056_P7489
11_AS44013_P9101
8_AS17156_P8295
9_AS24680_P8778
10_AS30971_P3465
47_Q5B3E6_Q5B3E6_EMENI_3-oxoacyl-ACP_reductase__putative__AFU_orthologue_AFUA_3G10540__OS=Emericella_nidulans
RBHcluster_879
18_AS17056_P5311
21_AS30971_P1992
19_AS17156_P118
20_AS24680_P8124
22_AS44013_P4760
23_AS64019_P6748
RBHcluster_6375
35_AS17056_P1519
36_AS17156_P9216
37_AS24680_P4244
39_AS44013_P6129
40_AS64019_P8312
38_AS30971_P6773
41_P73574_FABG_SYNY3_3-oxoacyl-ACP_reductase_OS=Synechocystis_sp
42_P28643_FABG_CUPLA_3-oxoacyl-ACP_reductase__chloroplastic_OS=Cuphea_lanceolata
43_P38004_FABG_CHLTR_3-oxoacyl-ACP_reductase_FabG_OS=Chlamydia_trachomatis
46_Q9Z8P2_FABG_CHLPN_3-oxoacyl-ACP_reductase_FabG_OS=Chlamydia_pneumoniae
45_P55336_FABG_VIBHA_3-oxoacyl-ACP_reductase_FabG_OS=Vibrio_harveyi
RBHcluster_7649
12_AS17056_P5061
13_AS17156_P2586
14_AS24680_P4839
17_AS64019_P1376
16_AS44013_P3524
15_AS30971_P8709

Figure 5.8. Phylogeny of FabG orthologous clusters with associated annotation sequence. FabG encodes 3-ketoacyl reductase for the elongation of fatty acids in Type II FAS. RBHcluster_6955, RBHcluster_879 and RBHcluster_6375 is the closest evolutionarily to known annotated FabG proteins.

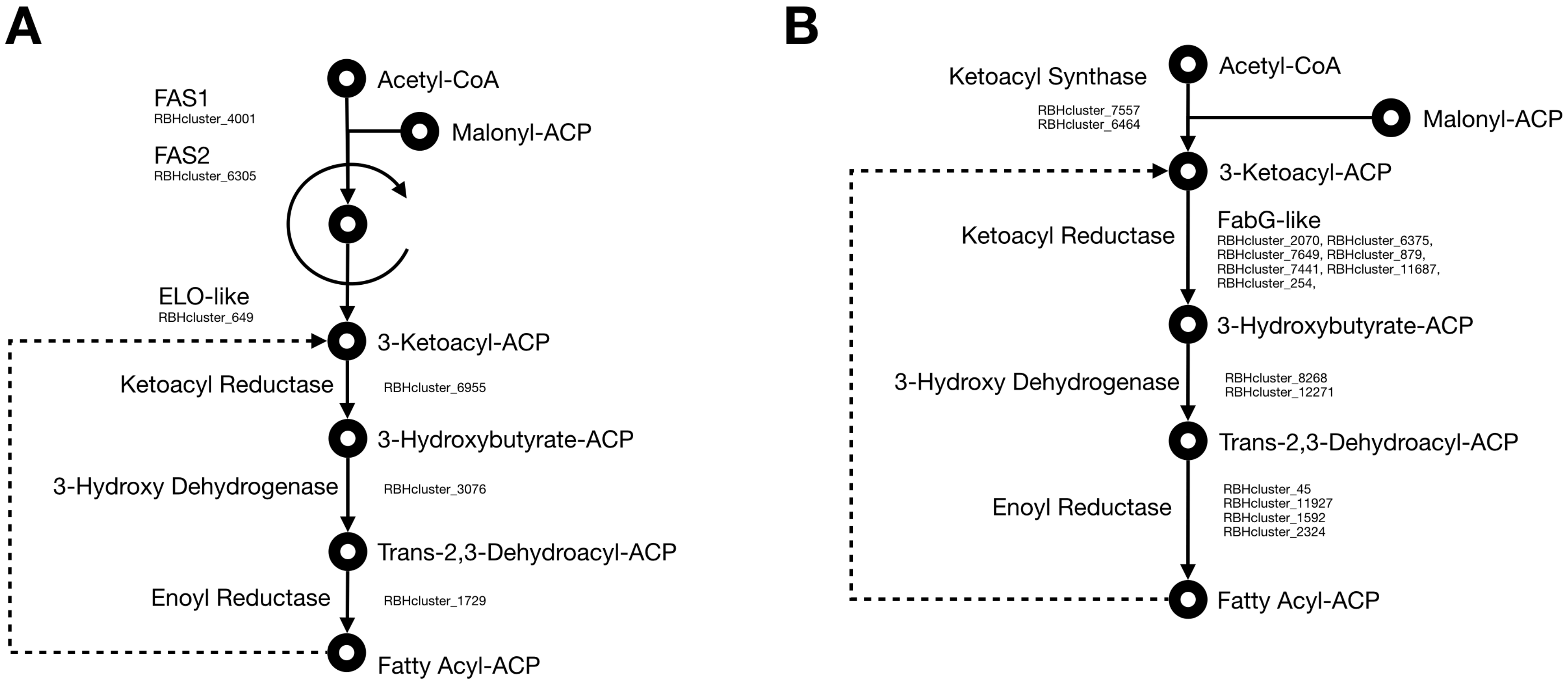


Figure 5.9. Annotated pathways of Type I FAS and Type II FAS. 5.9A. The Type I FAS and medium to long chain elongation systems annotated with orthologous clusters from *A. sarcoides. 5.9*B. The Type II FAS pathway, usually found in fungal mitochondria annotated with orthologous clusters from *A. sarcoides*.

5.2.7 Fatty acids degradation

The beta-oxidation pathway facilitates the degradation of fatty acids in fungi. The degradation pathway is the process of breaking down fatty acids to key precursors for energy generation and for the generation of metabolites. This yields acetyl-CoA, which is used for energy production in the citric acid cycle. In some respects, degradation can be seen as the reversal of the fatty acid biosynthesis pathway. Instead of elongating the fatty acyl chain by two per cycle, in the degradation pathway the acyl chain length is shortened by two carbons per cycle. Four discrete enzymes catalyse each step of the beta-oxidation pathway. First, the dehydrogenation of fatty acyl-CoA generates an unsaturated bond between C2 and C3 by an acyl-CoA dehydrogenase. For acyl-CoA dehydrogenase, there are a total of six (Table 5.15) orthologous clusters annotated (RBHcluster_1806, RBHcluster_2619, RBHcluster_6154, RBHcluster_6530 and RBHcluster_869). Although one of the cluster formed a six protein group by OrthoVenn2, reciprocal network analysis indicated that this orthologous cluster was incongruent and so it was separated into a cluster of two orthologous protein clusters (RBHcluster_10532) and four singlets (RBHcluster_10809, RBHcluster_10958, RBHcluster_11053 and RBHcluster_11412). This indicates a high degree of separation in the sequence of that cluster.

Second, the hydration of the unsaturated bond leads to the formation of L-3-Hydroxyacyl-CoA by an enoyl-CoA hydratase. Four orthologous clusters (RBHcluster_1218, RBHcluster_6369, RBHcluster_5531 and RBHcluster_5979) were annotated to have enoyl hydratase activity (Table 5.17). Upon reviewing the annotation further, RBHcluster_5531 was discarded from this set as KAAS indicates that it is most likely involved in valine, leucine and isoleucine degradation. Furthermore, RBHcluster_5979 was annotated to be an enoyl-CoA hydratase/ isomerase localised in the peroxisomal compartment, which was also discarded from this set. The RBHcluster_6369 cluster was annotated to be localised in the mitochondria. Of interest is the RBHcluster_1218; proteins from this cluster were thought to be a multifunctional beta-oxidation protein that is thought to have enoyl hydratase activity and 3-hydroxyacyl-CoA dehydrogenase activity localised in the peroxisome.

Third, the hydroxy group is further oxidised to a 3-ketoacyl-CoA by a 3-hydroxyacyl-CoA dehydrogenase. As mentioned previously, RBHcluster_1218 was previously annotated to contain a dual function enzyme that is thought to also catalyse 3-hydroxyacyl-CoA dehydrogenase activity (Table 5.17). Furthermore, a cluster of five orthologous protein (RBHcluster_7604) and a singlet (RBHcluster_12955) was also annotated to be 3-hydroxyacyl-CoA dehydrogenase. Although there are conflicting annotation evidences for this cluster and singlet, because Orthovenn's annotation indicate that this maybe part of a secondary metabolite biosynthesis gene.

The final step catalyses the cleavage of 3-ketoacyl-CoA by a beta-ketothiolase, which releases an acetyl group and forms an (n-2) fatty acyl-CoA (Table 5.18). Due to the thiolysis of the C2 and C3, an additional cofactor-A is required to stabilise the cleaved acyl chain. For this step of the pathway, RBHcluster_5728 was annotated to be a ketothiolase

Table 5.16. Orthologous protein clusters that are annotated for acyl-CoA dehydrogenase. This is the first step for fatty acid degradation in the beta oxidation process.

Reciprocal network analysis			Putative annotation	Annotation evidence	*A. sarcoides* protein accession
Cluster	Highest E-value	Average percentage ID (%)			
RBHcluster_869	0	99.8	Acyl-CoA dehydrogenase	*A. nidulans* Acyl-CoA dehydrogenase, putative (C8VQF1) Short/branched chain specific acyl-CoA dehydrogenase, mitochondrial (Q9DBL1)	AS17056_P2389 AS17156_P7163 AS24680_P5202 AS30971_P1474 AS44013_P2059 AS64019_P2240
RBHcluster_9046	3.10E-77	88.6	Acyl-CoA dehydrogenase	acyl-CoA dehydrogenase (K00249) *M. tuberculosis* fadE26 Acyl-CoA dehydrogenase (I6YCA3)	AS17056_P1918 AS17156_P7474 AS24680_P8220 AS44013_P1136 AS64019_P4701
RBHcluster_12888	-	-	Acyl-CoA dehydrogenase	acyl-CoA dehydrogenase (K00249) *M. tuberculosis* fadE26 Acyl-CoA dehydrogenase (I6YCA3)	AS30971_P8979

Table 5.15 (continued). Orthologous protein clusters that are annotated for acyl-CoA dehydrogenase. This is the first step for fatty acid degradation in the beta oxidation process.

Reciprocal network analysis			Putative annotation	Annotation evidence	*A. sarcoides* protein accession
Cluster	Highest E-value	Average percentage ID (%)			
RBHcluster_ 10532	-	-	Acyl-CoA dehydrogenase	*A. alternata* Acyl-CoA dehydrogenase (Q50LG2)	AS44013_P9956 AS64019_P2081
RBHcluster_ 10809	-	-	Acyl-CoA dehydrogenase	*A. alternata* Acyl-CoA dehydrogenase (Q50LG2)	AS17056_P9612
RBHcluster_ 10958	-	-	Acyl-CoA dehydrogenase	*A. alternata* Acyl-CoA dehydrogenase (Q50LG2)	AS17156_P9030
RBHcluster_ 11053	-	-	Acyl-CoA dehydrogenase	*A. alternata* Acyl-CoA dehydrogenase (Q50LG2)	AS24680_P5462
RBHcluster_ 11412	-	-	Acyl-CoA dehydrogenase	*A. alternata* Acyl-CoA dehydrogenase (Q50LG2)	AS30971_P2473
RBHcluster_ 1806	0	97.2	Acyl-CoA dehydrogenase	Acyl-CoA dehydrogenase (Q5ATG5) *C. elegans* Acyl-CoA dehydrogenase (Q9XWZ2)	AS17056_P2053 AS17156_P3857 AS24680_P3400 AS30971_P983 AS44013_P5507 AS64019_P10053

Table 5.15 (continued). Orthologous protein clusters that are annotated for acyl-CoA dehydrogenase. This is the first step for fatty acid degradation in the beta oxidation process.

Reciprocal network analysis			Putative annotation	Annotation evidence	*A. sarcoides* protein accession
Cluster	Highest E-value	Average percentage ID (%)			
RBHcluster_3910	0	97.3	Acyl-CoA dehydrogenase	*A. nidulans* Sad1/ UNC domain protein (Q5BB16) *S.cerevisiae* Protein BNS1 (P50084) acyl-CoA dehydrogenase (K00249) *G. gallus* Acyl-CoA dehydrogenase (Q5ZHT1)	AS17056_P2062 AS17156_P3880 AS24680_P7112 AS30971_P4665 AS44013_P6838 AS64019_P3367
RBHcluster_6154	0	92.7	Acyl-CoA dehydrogenase	*A. nidulans* Acyl-CoA dehydrogenase, putative (Q5BCN1) *A. alternata* Acyl-CoA dehydrogenase (Q50LG2)	AS17056_P6027 AS17156_P177 AS24680_P8457 AS30971_P4493 AS44013_P4138 AS64019_P8830
RBHcluster_6530	0	97.7	Acyl-CoA dehydrogenase	*A. nidulans* Acyl-CoA dehydrogenase family protein (Q5AZ86) *S.cerevisiae* Nucleoporin NUP2 (P32499) *A. nidulans* Acyl-CoA dehydrogenase (Q5ATG5)	AS17056_P8671 AS17156_P7810 AS24680_P7952 AS30971_P8922 AS44013_P3248 AS64019_P3070

Table 5.16. Orthologous protein clusters that are annotated for enoyl-CoA hydratase. This is the second step for fatty acid degradation in the beta oxidation process.

Reciprocal network analysis			Putative annotation	Annotation evidence	*A. sarcoides* protein accession
Cluster	Highest E-value	Average percentage ID (%)			
RBHcluster_1218	0	99.3	enoyl-CoA hydratase	*A. nidulans* Peroxisomal multifunctional beta-oxidation protein (MFP) (C8VDC2) *S. cerevisiae* Peroxisomal hydratase-dehydrogenase-epimerase, HDE (Multifunctional beta-oxidation protein, MFP) (Q02207) multifunctional beta-oxidation protein (K14729) *N. crassa* Peroxisomal hydratase-dehydrogenase-epimerase, HDE (Multifunctional beta-oxidation protein, MFP) (Q02207)	AS17056_P5229 AS17156_P6161 AS24680_P2226 AS30971_P2993 AS44013_P940 AS64019_P8506
RBHcluster_6369	0	93.8	enoyl-CoA hydratase	*A. nidulans* Uncharacterized protein (Q5B9R4) Probable enoyl-CoA hydratase, mitochondrial (Q1ZXF1)	AS17056_P7345 AS17156_P5876 AS24680_P2460 AS30971_P3159 AS44013_P7871 AS64019_P5662

Table 5.16 (continued). Orthologous protein clusters that are annotated for hydratase dehydrogenase. This is the second step for fatty acid degradation in the beta oxidation process.

Reciprocal network analysis					
Cluster	Highest E-value	Average percentage ID (%)	Putative annotation	Annotation evidence	*A. sarcoides* protein accession
RBHcluster_5531	0	89.1	enoyl-CoA hydratase	*A. nidulans* Mitochondrial methylglutaconyl-CoA hydratase (Q5B984) methylglutaconyl-CoA hydratase (K05607) Enoyl-CoA hydratase domain-containing protein 2 (Q3TLP5)	AS17056_P3447 AS17156_P1324 AS24680_P3578 AS30971_P9575 AS44013_P2203 AS64019_P3736
RBHcluster_5979	0	90.6	enoyl-CoA hydratase	Enoyl-CoA hydratase/ isomerase family protein (Q5AY63)	AS17056_P7775 AS17156_P6192 AS24680_P6760 AS30971_P7303 AS44013_P1257 AS64019_P4983

Table 5.17. Orthologous protein clusters that are annotated for 3-hydroxyacyl-CoA. This is the third step for fatty acid degradation in the beta oxidation process.

Reciprocal network analysis					
Cluster	Highest E-value	Average percentage ID (%)	Putative annotation	Annotation evidence	*A. sarcoides* protein accession
RBHcluster_7604	0	91.3	3-hydroxyacyl-CoA	*A. nidulans* 3-hydroxyacyl-CoA dehydrogenase, putative (C8V6H1) *C. heterostrophus* 3-hydroxyacyl-CoA dehydrogenase-like protein LAM1 (N4WEA4)	AS17056_P5567 AS17156_P1669 AS24680_P8943 AS44013_P5268 AS64019_P5324
RBHcluster_12955	-	-	3-hydroxyacyl-CoA	*A. nidulans* 3-hydroxyacyl-CoA dehydrogenase, putative (C8V6H1) *C. heterostrophus* 3-hydroxyacyl-CoA dehydrogenase-like protein LAM1 (N4WEA4)	AS30971_P9236

Table 5.18. Orthologous protein clusters that are annotated for 3-ketoacyl-CoA thiolase. This is the last step for fatty acid degradation in the beta oxidation process.

Reciprocal network analysis			Putative annotation	Annotation evidence	*A. sarcoides* protein accession
Cluster	Highest E-value	Average percentage ID (%)			
RBHcluster_ 5728	0	95.7	3-ketoacyl-CoA thiolase	*A. nidulans* 3-ketoacyl-CoA ketothiolase, putative (Q5BEI0) *M. musculus 3-ketoacyl-CoA thiolase A, peroxisomal*(Q921 H8)	AS17056_P9005 AS17156_P8272 AS24680_P4778 AS30971_P9819 AS44013_P8641 AS64019_P8044

5.2.8 Fatty acyl reductases

From the annotations we were able to elucidate a putative *A. sarcoides* fatty acyl-CoA reductase (FAR). As mentioned previously, reduction of fatty acid is crucial in forming fatty aldehyde in the two-step model for alkane biosynthesis. Our BLAST pipeline was able to annotate proteins within RBHcluster_515 (Table 5.19) to be a putative *Aspergillus nidulans* FAR (Uniprot: Q5B282). Further evidence supported by KAAS, annotates the cluster to be a FAR enzyme (KEGG: K13356) and the OrthoVenn2 annotation pipeline revealed it to be a *Simmondsia chinensis* alcohol-forming FAR (Uniprot: Q9XGY7). With reciprocal network analysis we were able to observe over 99% sequence identity, which suggests the putative FAR is highly conserved among *A. sarcoides* isolates. Furthermore, each isolates' proteome contains just a single copy of the putative FAR. When analysed with HHpred for structural homology (Table 5.20), the structure of asFAR was similar to that of well characterised aldehyde producing carboxylic acid reductases (CAR) such as *Mycobacterium marinum's* CAR (Probability: >99.86%), *Nocardia iowensis'* CAR (Probability: >99.81%), and *Segniliparus rugosus'* CAR (Probability: >99.77%). Other significant HHpred hits included short-chain dehydrogenases/reductases, enoyl-reductase, saccharide dehydratases, and saccharide epimerases. This further suggests asFAR has the capacity to conduct a redox reaction on carbon-oxygen bonds. Using PSI-BLAST (Figure 5.10), asFAR was revealed to contain a short-chain dehydrogenase/reductase (SDR) superdomain. Within the SDR domain, it also contained a NAD binding domain with a motif of [ST]GXXGXXG and a thioesterase reductase active site with a YXXXK motif and upstream [TS]. This implies that asFAR is able to reduce acyl-CoA thioester bonds by using NAD as a source of reducing power. Overall, this annotation suggests *A. sarcoides* is capable of reducing fatty acids to aldehyde.

Table 5.19. Orthologous protein clusters that are annotated for fatty acyl reductase proteins. In the currently understood model for alkane biosynthesis fatty acyl reductases are one part of the two step reaction to synthesis an alkane.

Reciprocal network analysis					
Cluster	Highest E-value	Average percentage ID (%)	Putative annotation	Annotation evidence	*A. sarcoides* protein accession
RBHcluster_ 5728	0	98.6	Fatty acyl reductase	Fatty acyl-CoA reductase (Q5B282) Fatty acyl-CoA reductase (K13356) *S. chinensis* Alcohol-forming fatty acyl-CoA reductase *(*Q9XGY7)	AS17056_P174 AS17156_P6867 AS24680_P5614 AS30971_P4229 AS44013_P5622 AS64019_P4865

Table 5.20. HHpred results of RBHcluster_515 as query. Structural homology indicate that proteins within RBHcluster_515 are highly similar to carboxylic reductase. Results were sorted by probability and only top 5 results were included.

PDB Accession	Name	Probability value (%)	E-value
5MSO_A	Carboxylic acid reductase (E.C.1.2.1.-,1.2.1.30); adenylation domain, carboxylic acid reductase; HET: NAP; 1.2A {*Mycobacterium marinum* M}	99.86	1.3E-23
4F6C_B	AusA reductase domain protein; Thioester reductase, OXIDOREDUCTASE; HET: MSE; 2.812A {*Staphylococcus aureus*}	99.86	2.7E-23
4U5Q_A	Peptide synthetase Nrp; nonribosomal peptide synthetase, oxidoreductase, Short-chain; HET: XPE; 1.811A {*Mycobacterium tuberculosis*}	99.87	4E-23
6GCS_E	NUAM protein (E.C.1.6.99.3), NADH dehydrogenase; Complex I, NADH dehydrogenase, Mitochondrion; HET: NDP, SF4, ZMP, 3PE, CDL, FMN; 4.32A {*Yarrowia lipolytica*}	99.86	4.3E-23
5XTB_J	NADH dehydrogenase [ubiquinone] flavoprotein 1; Respiratory, OXIDOREDUCTASE-ELECTRON TRANSPORT complex; HET: SF4, 8Q1, NDP, FMN; 3.4A {*Homo sapiens*}	99.86	5.9E-23

```
1     MPFEDEACARSMAGKPATIFLTGATGFLGKVVLEELFRRRHEINFEKVIVLTRPKRGKDP
                            ^ ^^^^                       ^^^
61    RSRFLNEIVSSPCFSNLPNGWTDSVQAVEGDLSLPSCGISEQTLEKICGEITHIIHCAGC
                                                              ^^^
121   VSFEAPILEAVGANITTALNVCNLAKDCPNLQRLTTTSTAYVQPHINGGIYEKLVDLPCP
                     ^                    ^^^

                                            *
                      °                     °
181   VGELLEDILEQRSTEAEILQLTAHQNTYTLTKCMAEHLLLERRGSIPITIVRPSIISASR
                                 ^   ^                    ^^^^

                                 *                          *
                                 °   °
241   AYPEPGWIDSKAAFAGFVTAFGAGILHVVDGNPNARGDIVPVDDVSRRLIDETLFSTAAP
            *               *    *
301   EKPKIVYAVAGLENCLQWSTGLRTLVDFFEGKPPGPGGKASLSYFGPRNLRFRFHEMIQQ

361   RLRFSLAELWYELRGDEKMRTRMVKVSNLVIMLNRIFPYFIHNTFDFRGDSDLLGDGFDS

421   EKYVGLNAVGVLADTQFEIYLINCLADGKSVDADVKGQEEGEVAVKGRS
```

Figure 5.10. Annotated SDR domain and key residues in the sequence of *A. sarcoides*' putative Fatty acyl-CoA Reductase. A putative asFAR was annotated by NCBI’s PSI-BLAST. The highlighted sequence indicate the short-chain dehydrogenase/reductase domain (SDR). Within the SDR domain it was able to highlight key residues that participate in active site (^), NAD binding (*), and putative binding substrate (°).

5.2.9 Putative Decarboxylase/Decarbonylation

In searching for a protein that is capable of decarbonylating aldehydes, we examined the annotations for evidence of carbon-carbon lyase (EC 4.1.-.-). This revealed 77 annotations that fit this criteria. From this set, one cluster (RBHcluster_2048) stood out to be interesting and is present in all isolates (Table 5.21). This cluster contains an annotation that encodes Ferulic acid decarboxylase 1 (FDC1), an enzyme that is capable of decarboxylating derivatives of phenolic acids, such as cinnamic acids to styrene (Nagy et al., 2019). This reaction is similar to that of $C_{(n-1)}$ alkane biosynthesis, in which both pathways generate a $C_{(n-1)}$ product and C1 by-product. Network analysis revealed two further clusters (RBHcluster_4872 and RBHcluster_2048) that were identified to also encode FDC1 (Table 5.21), making a total of three genes that encode FDC1 in each isolate of *A. sarcoides*. Studies of FDC1 showed that there is only one copy of *fdc1* in each of the *fdc1*-containing organisms. Phylogenetic analysis of the three clusters and of characterised FDC1 genes indicate that RBHcluster_6878 is closely homologous to FDC1 derived from *A. niger* (A2QHE5) (Figure 5.11A). This suggests that RBHcluster_6878 is capable of similar activity to that of FDC1 in *A. niger* (Figure 5.11B). The remaining two clusters form distinct individual clades. This suggests that both proteins within both clusters might be capable of alternate decarboxylase activity that is unique to FDC1 (Figure 5.11C).

Table 5.21. Orthologous protein clusters that are annotated for FDC1. In the currently understood model for alkane biosynthesis a carbon-carbon lyase is required. FDC1 is a decarboxylase that was detected to be carbon-carbon lyase in all isolates of *A. sarcoides*, with a total of three copies.

Reciprocal network analysis					
Cluster	Highest E-value	Average percentage ID (%)	Putative annotation	Annotation evidence	*A. sarcoides* protein accession
RBHcluster_ 2048	0	94.4	FDC1	*A. niger* Ferulic acid decarboxylase 1 (A2QHE5)	AS17056_P6894 AS17156_P8944
RBHcluster_ 4572		98.1	FDC1	*S. cerevisiae* Ferulic acid decarboxylase 1	AS17056_P7906 AS17156_P9661
RBHcluster_ 6878		90.9	FDC1	*A. nidulans* Uncharacterized protein (Q5AX16)	AS17056_P4152 AS24680_P287

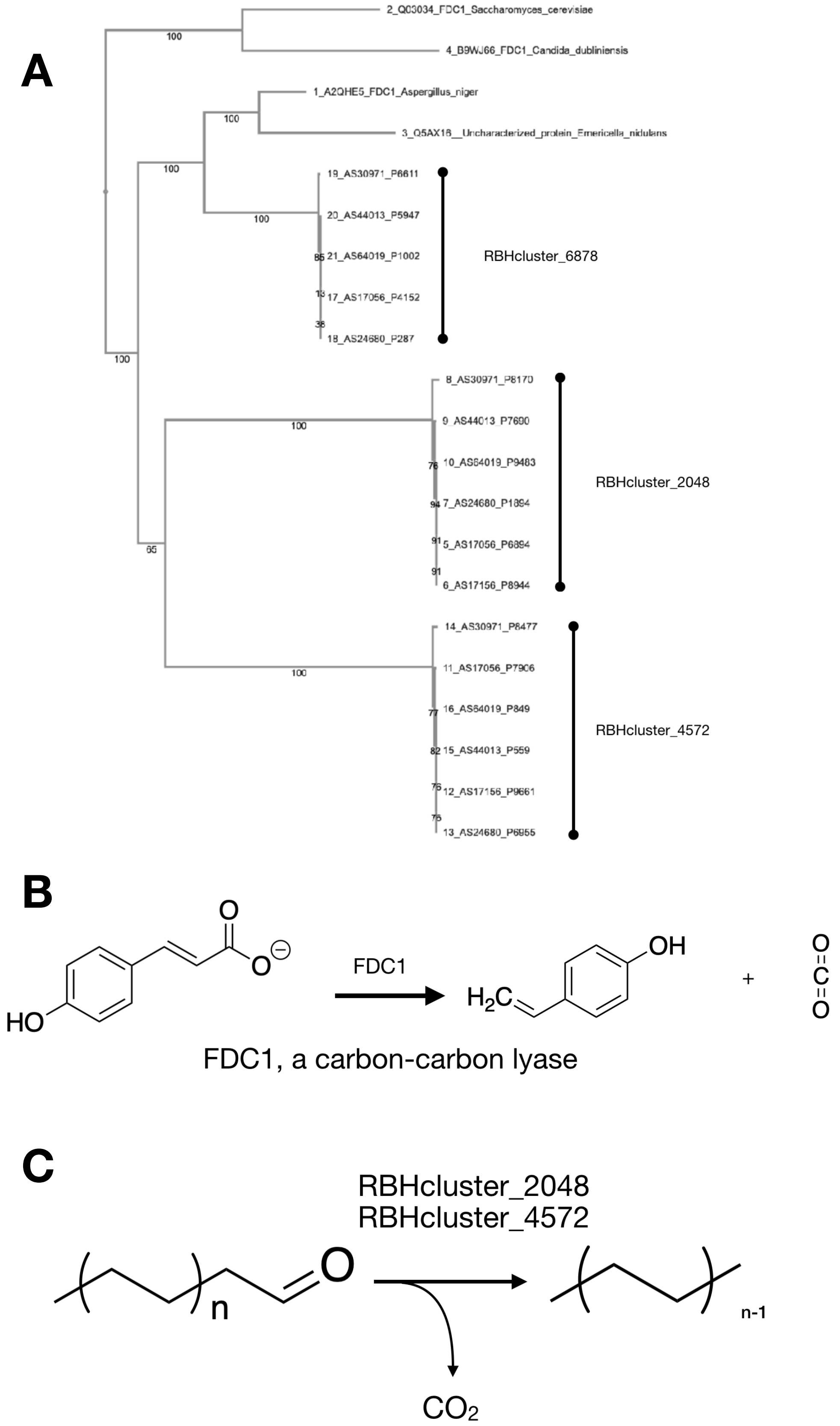
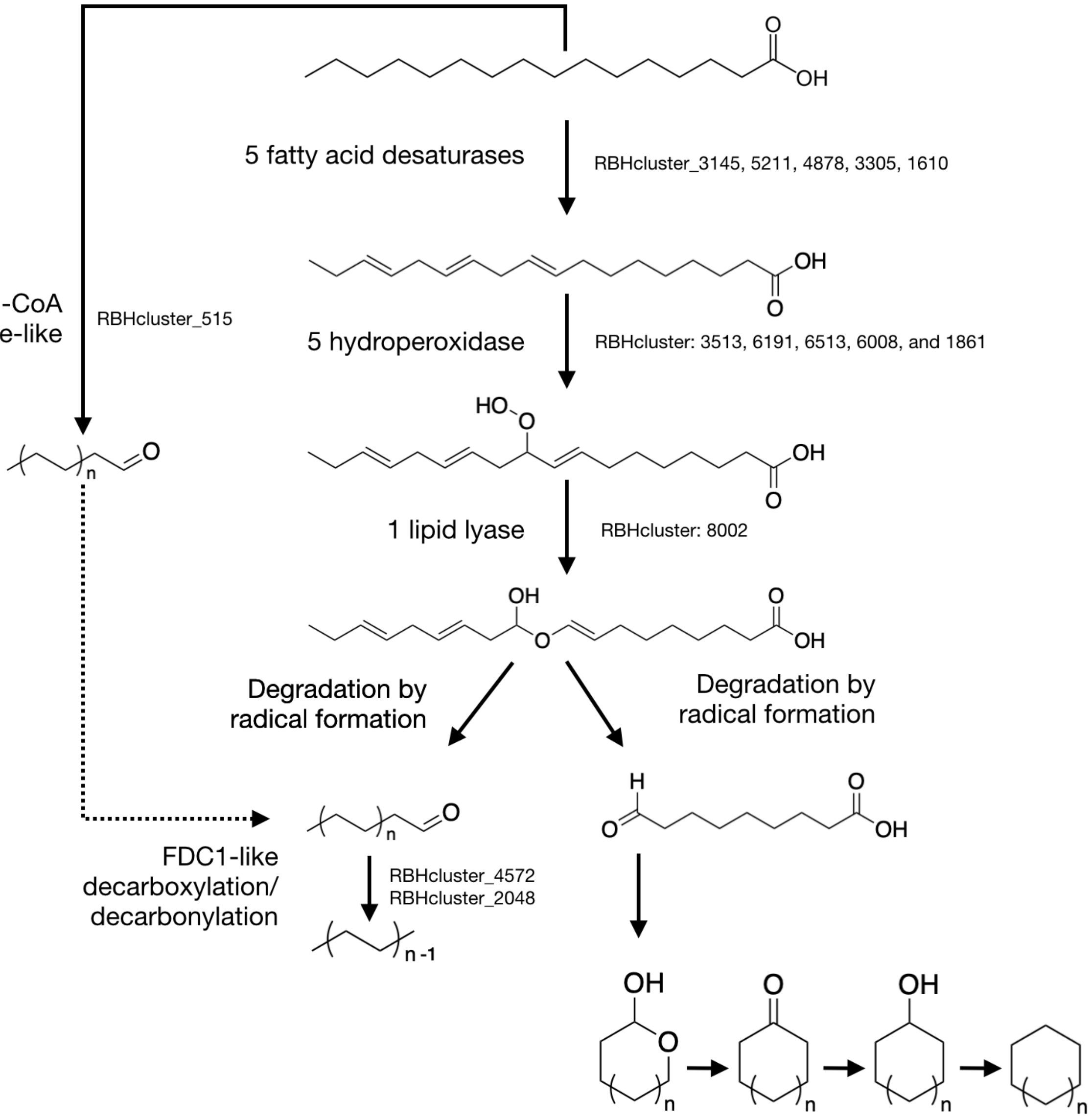


Figure 5.11. Summary of FDC1. A, phylogenetic tree of putative FDC1 clusters against characterised FDC1. Results indicated two unique orthologs that are paralogous to known FDC1. **B,** Reported activity of FDC1. **C,** proposed encoded function of the two paralogous clusters within *A. sarcoides.*

5.2.10 Cyclic alkane biosynthesis

There was no evidence for a cyclic alkane pathway with the reciprocal network analysis approach. This is because there were no previously characterised gene for cyclic alkanes. Because of this, we took the approach of reviewing synthetic pathways that contain cyclic alkane motifs. Civetone and exaltone are macrocyclic ketones. Synthetic routes for these macrocyclic ketone compounds rely on terminal oxygenated fatty acids as precursors. In this case, terminal oxygenated fatty acids can include terminal aldehyde fatty acids or dicarboxylic fatty acids. In fact, chemical synthesis of civetone is possible from palmitic acid (Choo et al., 1994). A diester derivative of palmitic acid (diethyl 9-octadecenedioate) condense with itself via Dieckmann condensation in the presence of WCl_6 and $SnMe_4$. This reaction produces 2-ethoxycarbonyl-9-cycloheptadecenone. A further decarboxylation step is able to yield civetone, a 9-cycloheptadecenone (Choo et al., 1994). For exaltone, a similar terminal oxygenated fatty acids precursor in the form of aleuritic acid was used (Mathur et al., 1963). Aleuritic acid, a major component of shellac, was chemically degraded to form pentadecanoic acid and is cyclises into exaltone, a C15 macrocyclic ketone. In contrast, chemical and biological degradation of cyclic ketones and cyclic alkanes also leads to the generation of terminal oxygenated fatty acids (Ruzicka, 1926, Stirling et al., 1977). Similarly, the biological degradation of cyclohexane revealed an oxygenated fatty acid as the degradation product of cyclic hexane. Cyclohexane is oxidised to form a cyclohexanol and further successive oxidation leads to the formation of an ε-caprolactam ring and hydrolysis of this leads to the formation of adipic acid, a dicarboxylic acid (Stirling et al., 1977). In this proposed biosynthetic pathway, the lipid hyperoxidase pathway is able to generate terminal oxygenated fatty acids of various length either by degradation by lyase or oxygenation by LOX (Figure 5.12). An unknown cyclase will initiate head to tail auto-condensation via Dieckmann condensation resulting in a cycloalcohol or a cycloketone. In the event of a mid-chain oxidation, the cyclisation can still take place and ring closure will occur at the oxygenated carbon via Claisen condensation. An unknown dehydratase will then reduce it to a cycloalkane. This model would be able to account for the wide ranging lengths of cyclic alkanes and branched cycloalkanes observed in this study and other studies. It is thought that a Dieckmann condensation is responsible for the beta-lactam motifs found in the antibiotic equisetin and trichostatin (Campbell et al., 2010).



5 fatty acid desaturases
RBHcluster_3145, 5211, 4878, 3305, 1610
Fatty acyl-CoA reductase-like
RBHcluster_515
5 hydroperoxidase
RBHcluster: 3513, 6191, 6513, 6008, and 1861
1 lipid lyase
RBHcluster: 8002
Degradation by radical formation
Degradation by radical formation
FDC1-like decarboxylation/ decarbonylation
RBHcluster_4572
RBHcluster_2048
OH
O
HO
n
n-1

Figure 5.12. Proposed pathway for linear and cyclic alkanes in *A. sarcoides.* The proposed linear alkane biosynthesis is dependent on FDC1 paralogs for decarboxylation/decarbonylation. Even chain alkanes are proposed to be derived from hyperoxidase lipid pathway. The proposed cyclic alkane biosynthesis is derived from examining synthetic pathways.

5.3.0 Discussion

5.3.1 Previously proposed pathway for A. sarcoides

A proposed biosynthetic pathway for linear alkane biosynthesis in *A. sarcoides* must be able to account for the diverse alkane profiles observed. It must take into account of even and odd chain alkanes. In 2012, a proposed lipid hyperoxidase pathway was published that is based on retrosynthetic analysis of two observations in *A. sarcoides* 50072. Firstly, gene expression levels that correlate with hydrocarbon producing conditions. Secondly, the observed metabolic pathway derived from the hydrocarbon producing conditions. A retrosynthesis approach is applied to detected metabolites to construct a pathway and this is correlated with gene expression to annotate genes within the hypothesised pathway. In this work, their model's algorithm assumes that observed volatile metabolites could be part of an intermediate of a metabolic pathway rather than the end product of a metabolic pathway.

By clustering annotations of NRRL 50072 to the annotations found in this study, we were able to conduct a comparable analysis by reducing both sets of annotations to a non-redundant list. Initial assessment of the annotations that participate in hypothesised pathways lead to annotations that do not have the required activity. Three independent lines of evidence revealed that these annotations did not have the biochemical potential to participate in the hypothetical hydrocarbon pathway. We attribute this discrepancy to clerical error in the process of forming unique identifiers in the annotations of NRRL 50072 (Gianoulis et al., 2012).

Thus, we endeavoured to reconstruct the hypothesised pathway proposed by the previous study with the combined annotations generated in Section 5.2.1. We are able to reconstruct a pathway with the activity required for the proposed pathway. In this reconstruction, we assumed that only orthlogous clusters that are found in all isolates are possible for the lipid hyperoxidase pathway. This was attributed to observations of alkane biosynthesis in all isolates of *A. sarcoides* (Griffin et al., 2010).

Once reconstructed, it allows for the pathway to be investigated in our six isolates. In this pathway, linear hydrocarbon biosynthesis deviates from the two-step aldehyde decarbonylase model. Instead, the generation of linear hydrocarbons is dependent on hydroperoxidation of unsaturated fatty acids, in this case linoleic acid. For this step, two annotations within NRRL 50072 suggests it to be an ortholog of *ppoC*, a psi-producing oxygenase C is required. Our annotation largely agrees with this possibility and was able to annotate two *ppoA* orthologs. No *ppoC* orthologs were detected. Moreover, three lipoxygenase orthologous clusters were detected: two linoleate 9S-lipoxygenase orthologs and a manganese lipooxygenase ortholog. Lipoxygenase (LOX) are a class of enzyme that catalyse a specific type of oxygenation known as hydroxyperoxidation on polyunsaturated fatty acids (PUFA). LOX are typically iron dependent enzymes which utilise iron for hydrogen abstraction of PUFAs to facilitate a radical reaction of the carbon to a molecular oxygen (Newcomer et al., 2015). Another class of LOX are the manganese-dependent LOX and, in these enzymes, manganese is responsible for hydrogen abstraction instead of iron (Su et al., 1998, Su et al., 2000). PpoA and ppoC are considered to be part the of iron-dependent lipoxygenase family of enzymes (Brodhun et al., 2009). Biochemical characterisation of ppoA and ppoC indicate that both are capable of hydroperoxidising linoleic acids but differ in the placement of a hydroperoxide group in the linoleic acid backbone. However, ppoA favours placement of a hydroperoxide group on the C8 (8-HPODE) and further oxygenate the hydroperoxide lipid intermediate to form a dihydroxylated fatty acids (Brodhun et al., 2009). Furthermore, *ppoA* is also able to carry out isomerisation of linoleic acid. Biochemical evidence indicates *ppoA* is able to act on monounsaturated and polyunsaturated fatty acids (Brodhun et al., 2009). In contrast, ppoC favours C10 placement for hydroperoxide groups (10-HPODE) and no isomerisation activity was detected (Brodhun et al., 2010). It was found that the combination of *ppoA* and *ppoC* can further act on 8-HPODE and 10-HPODE to further oxygenate respective HPODE to form epoxy alcohol groups (Brodhun et al., 2010). It was observed that 10-HPODE appeared to be unstable and can either decompose into a C10 and C8 fragment or can auto-oxidise 10-HPODE to a C10 ketone body (Brodhun et al., 2010). Different fatty acids were incubated with *ppoC,* a diverse profile of product was detected: 1-octanol, 2-undecanal, 1-octen-3-ol, 2-octen-1-ol, 2-octenal, and 3-octanone (Brodhun et al., 2010).

A lyase is required to catalyse a Hock cleavage of the fatty HPODE, which yields a fatty acid with a terminal aldehyde group and an unsaturated aldehyde (Schneider et al., 2001). This reaction does not result in decarbonylation thus the sum length of the two fragments are equal length to the fatty HPODE. A single orthologous cluster and a singlet fit this description of a HPODE lyase (RBHcluster_8002). This is comparable to the hypothesised pathway, in which a single HPODE lyase was annotated in *A. sarcoides* NRRL 50072 dataset. The product of the lyase is an unsaturated alcohol in the form of 1-octen-3-ol, a well characterised product associated with the musky scent of fungi (Rösecke et al., 2000). Dehydration of 1-octen-3-ol by a dehydratase is required to form a hydrocarbon, 1,3-octadiene. We observed four possible clusters that are annotated for this function (RBHcluster_2268 and RBHcluster_7302). Once fatty acid elongation was taken into account, three orthlogous clusters remain annotated for this function. Finally, an enoyl reductase is thought to reduce the unsaturated bond in 1,3-octadiene to form a terminal olefin, 1-octene. In our analysis, we are able to find two orthologous clusters (RBHcluster_3076) and a singlet (RBHcluster_8268 and RBHcluster_12271) that is able to catalyse this step and does not take part in mitochondrial metabolism.

In the final two steps of the lipid hyperoxidase pathway, the steps proposed are highly similar to steps involved in *de novo* fatty acid elongation in fungi. In fact, when gene expression levels are examined for producing and non-producing conditions, there are no significant difference (Gianoulis et al., 2012). The last two steps in the lipid hyperoxidase pathway represent the step with the most uncertainty, as it is ambiguous whether the annotation takes part in elongation of *de novo* fatty acids, hydrocarbon biosynthesis or both. This represents the limitation of our analysis and the published pathway; we are unable to confirm if these orthologous cluster are involved in the elongation of fatty acids or the biosynthesis of terminal olefins.

5.3.2 Hypothetical pathway for linear and cyclic alkanes

Genomes from six isolates of *A. sarcoides* searched for well-characterised alkane genes did not return any significant hits. This indicated that there is a novel basis for alkane biosynthesis. Following this, we endeavoured to elucidate genes that underpin the biosynthesis of alkanes. This led to the discovery of a gene that encodes FAR. Intriguingly, this putative FAR matched significantly with a FAR derived from a plant. Such a distant match suggests that the annotation is potentially novel within the fungal kingdom. Furthermore, this also suggests that each isolate is capable of biosynthesising fatty aldehydes via FAR. For the C(n-1) alkane biosynthesis, aldehyde is required as a substrate for the cleaving. It has been documented that for most fungi the presence of aldehyde is toxic (Aranda et al., 2003, Kunjapur et al., 2014, Kunjapur et al., 2015). This then suggests that aldehyde is a transient intermediate, rather than a specific end product.

We proposed that alkane formation is encoded by one of two paralogs of *fdc*1 (Figure 5.12). This was identified by interrogating the dataset for carbon-carbon lyases and identifying three separate *fdc*1 clusters in each isolate. Interestingly, FDC1 has been biochemically characterised to be able to catalyse a wide range of aryl and aryl cinnamic derivatives under a wide range of conditions (Nagy et al., 2019). The FDC1 enzyme is capable of subjecting arylic acid substrate to a $C_{(n-1)}$ decarboxylation, leading to formation of an unsaturated bond at the site of the cleaved carbon and the formation of a CO_2 by-product (Nagy et al., 2019), similar to that of alkane biosynthesis. It is this wide-ranging substrate acceptance that led us to question whether *A. sarcoides* contains two additional copies of *fdc*1-like genes. We hypothesise that the remaining two paralogs are capable of analogous decarboxylation/decarbonylation reaction, potentially on fatty acids. Indeed this is supported by phylogenetic analysis, in which a single cluster was closely aligned to an *A. niger* FDC1, while the other two clusters are relatively distant, suggesting it to be paralog of *fdc*1. Ultimately, this suggests that FAR and FDC1-like enzymes take part in linear alkane biosynthesis.

For cyclic alkane biosynthesis we were unable to suggest the underpinning genetic element. By reviewing synthetic pathways, the importance of a dicarboxylic acid or a terminal oxygenated fatty acids to be a substrate for cyclisation. This was also supported by retro-synthesis analysis, in which the degradation product of cyclohexane is a terminal oxygenated fatty acid. While this is a speculative pathway, it does highlight the potential of the lipid hyperoxidase pathway to generate dicarboxylic acid or a terminal oxygenated fatty acids (Brodhun et al., 2009). In this hypothetical pathway, terminal oxygenated fatty acids are subjected to cyclisation, in which the polar oxygen guides the cyclisation. This will either form a caprolactam or a cyclic alcohol. This is then reduced, respectively, by either one or two reduction steps to form a cycloalkane. Importantly, this hypothetical pathway details the likely substrate and the number of steps required. If we assume a single gene encodes a single step, this suggests that two or three genes are required for the cyclic alkane pathway.

5.3.3 Even-chain alkane biosynthesis

A. sarcoides contains the expected fatty acid biosynthesis genes which we can organise into comprehensible pathways. From a bioinformatics perspective, FAS Type I pathway was observed to be typical and homologiess for each subunit were comparable to well characterised FAS1 and FAS2. Similarly, we are able to annotate the elongation of fatty acids in *A. sarcoides.* In the previous chapter, alkane of even chain length were detected during the culturing of *A. sarcoides* in our experimental system. For this to occur, one of two theories must be true. Firstly, *A. sarcoides* could have unique odd-chain fatty acid profiles and alkane biosynthesis may occurs via $C_{(n-1)}$ decarbonylation. Secondly, *A. sarcoides* have a typical even-chain fatty acid profiles and alkane biosynthesis might occur via n-2 decarbonylation.

In literature, there are no reports of Type I FAS generating odd-chain fatty acids. Within our annotation and analyses, there were no significant observation that deviate from a typical FAS1/2 homodimeric paradigm that can explain for odd-chain fatty acids. Thus on the balance of evidence and literature, *A. sarcoides* Type I FAS is likely to be responsible for typical even-chain fatty acid biosynthesis. Mitochondrial Type II FAS is thought to have a flexible substrate chain-length preference leading to elongation with either acetyl-CoA or malonyl-CoA (Seubert et al., 1968, Boone et al., 1970, Bloch et al., 1977, Bessoule et al., 1987). This indicates that the alkane biosynthesis may be utilising fatty acids from mitochondrial FAS type I to generate odd-chain alkanes. Moreover, once the odd-chain fatty acids are formed, they can be further elongated by typical n+2 fatty acids elongation mechanism. This would result in a pool of long chain odd fatty acids that can be decarbonylated to form even-chain alkanes. Alternatively, the biochemical potential of the lipid hyperoxidase pathway may explain the generation of even-chain alkane that conforms with the decarboynlation model. The combination of LOX, HPODE, and diverse set of PUFAs represents a route for *A. sarcoides* to generate odd chain aldehyde, which, when catalysed with an ADO, leads to the formation of even chain alkane.

Current understanding from quantum mechanical/molecular mechanical studies of cyanobacteria ADO give insights to its mechanism. The n-1 decarbonylase reaction is only possible due to the polarity of oxygen in the aldo group. This polarity allows for a nucleophilic attack on the aldo C1 carbon by an iron peroxo group to form an intermediate peroxyhemiacetal. This nucleophilic attack transfers an oxygen group to the now radical-formyl group, which induces C1-C2 cleavage to form an alkyl radical. The alkyl radical is pronated to form an alkane product and a formate byproduct (Wang et al., 2016). On the evidence provided by literature, it would seem that a n-2 decarbonylation is highly unlikely and will require a different enzymatic mechanism that that will require a non-aldehyde substrate. Thus, by reviewing literature evidence, it is more probable for the biosynthesis of even-chain alkane to rely on odd-chain aldehyde substrate than on a malonyl-based elongation of fatty acids or the generation of odd-chain aldehyde from the lipid hyperoxidase pathway.

On the balance of evidence, even-chain alkane is most likely derived from a lipoxygenase pathway, in which PUFAs are degraded to form odd-chain aldehydes and then converted to even chain alkanes. The combination of LOX and HPODE lyase represents a route for *A. sarcoides* to generate aldehyde, potentially for decarbonylation. Moreover, LOX pathway also represents *A. sarcoides'* biochemical potential to generate long chain hydroxylated fatty acids. In this hypothetical scenario, a desaturase introduces an unsaturated bond to a fatty acid. Hyperoxidation occurs at the unsaturated bond and coordinates lysis by LOX. This then forms oxygenated acyl compounds, including odd-chain aldehydes. This is then cleaved for the formation of even-chain alkanes.

5.3.4 Alkane degradation

To the best of our knowledge, this is the first reported observation of alkane degradation and alkane biosynthesis in a fungal organism. Here we propose the genetic basis of alkane degradation in *A. sarcoides*. Each isolate contains two cytochrome P450 monooxygenases, an Alk1 ortholog and an Alk2A ortholog. We can ascertain the substrate specificity of alkane degradation by observing assimilation of sole carbon alkanes. Previously, in chapter 3, *A. sarcoides* was observed to assimilate tetradecane (Figure 3.6), a mid-chain alkane that was also observed to be biosynthesised. Compared to a well studied model for alkane degradation, *Yarrowia lypolytica* contain 12 isoforms of ALK genes. In *Y. lipolytica*

Alk1 and Alk3 were characterised to have wide alkane specificity, while Alk2, Alk6, and Alk9 have been characterised and were observed to have specific chain-length substrate.

Alk1 and Alk2A activity has been observed in endoplasmic reticulum (ER) and microsomal fractions (Craft et al., 2003, Thevenieau et al., 2007) indicating that alkane degradation is localised in these cellular regions. Thus, for alkane biosynthesis and alkane degradation to take place in *A. sarcoides*, each pathway must be compartmentalised to avoid futile cycling. We hypothesise the possibility of alkane biosynthesis in the peroxisome compartment as alkane degradation may occupy ER compartment. Pathways such as beta-oxidation have been observed to take place in the peroxisome. The peroxisome is known to partition pathway from the rest of the cell and there is molecular machinery in place to facilitate the availability of fatty acids to drive alkane biosynthesis

As mentioned previously, studies indicate that for alkane degradation to be expressed, alkane must be present and sensed to drive expression the required alkane monooxygenase genes. This, ultimately, is incompatible in *A. sarcoides*. If Yasp-related transcriptional controls the expression of alkane degradation, then there is a possibility of futile expression in *A. sarcoides*. Thus, this leads to an hypothesis for *A. sarcoides* to contain a different regulation for alkane metabolism to alkane degradation in *Y. lipolytica*. Indeed, Yas1p, Yas2p and Yas3p homologs were not detected. This further support the hypothesis of a novel alkane regulation system and represents an interesting regulation for alkane degradation to further explore. If elucidated, this novel regulation system, when coupled to a reporter gene, can act as a biosensor against alkane. This would solve a significant challenge for alkane detection, as current methods requires extraction and expensive low-throughput detection system. A potential biosensor can overcome this and see applications for more complex high-throughput alkane detection or for the detection of in-field environmental alkanes. Questions remain with regards to how *A. sarcoides* is able to reconcile alkane degradation and alkane biosynthesis.

5.3.5 Terpene

With this approach it was possible to identify putative genes involved in isoprene biosynthesis. This elucidated the mevalonate pathway and associated prenyl transferases. Both of these pathways are key for terpene biosynthesis. In our pipeline, we did not observe the expected annotations responsible for sesquiterpene biosynthesis. While many sesquiterpene synthases in plants have been identified and characterised, there is a dearth of description for fungal monoterpene and sesquiterpene synthases in many databases. In this case, when entries are limited for a particular pathway or gene, it severely hampers the possible annotation by homology against novel datasets. Currently, the only known alpha-muurolene synthase in fungi is the *cop*3 gene. The *cop*3 gene was elucidated and functionally characterised from *Coprinopsis cinerea*, a basidiomycete (Agger et al., 2009), and the evolutionarily distance to *A. sarcoides* may hamper comparison by homology. Ultimately, this highlights a need for a more robust annotation approaches that are not solely homology dependent in addition to a need to identify a greater number of terpene biosynthesis pathways.

Prenyl compounds can act as substrates for downstream synthesis of terpene compounds by terpene synthases to produce cyclic or acyclic terpene compounds. Furthermore, modification to synthesised terpenes can lead to the generation of a large variety of compounds. Terpene modification can include (de)esterification, (de)methylation, glycosylation, isomerization, oxidation, reduction, and/or decorations with functional groups to a terpene compound. To date, 40,000 terpenes have been reported. The terpene pathways have been shown to be a suitable pathway for biofuel production. Terpenoid pathways can be manipulated to produce α-farnesene, a highly branched and unsaturated hydrocarbon molecule. α-Bisabolene, a highly branched and unsaturated hydrocarbon with a cyclohexane motif, can also be the product of terpenoid pathways. The biosynthesis of both α-farnesene and α-bisabolene is a one-step conversion of FPP and is achieved enzymatically by α-farnesene synthase and α-bisabolene synthase, respectively. Both reactions dephosphorylate FPP to produce their respective product. However, these unsaturated hydrocarbon molecules are prone to oxidation by air. For fuel transportation usage, both α-farnesene and α-bisabolene must undergo hydrogenation to produce farnesane and bisabolane. Currently, commercial production of farnesane and bisabolane is dependent on the chemical hydrogenation with hydrogen in the presence of a rare metal catalyst. Alternatively, terpenoid C=C unsaturated bonds can be hydrogenated by a reactive proton donor, such as hydrazine, under atmospheric O_2 and the presence of a flavinium catalyst, such as 5-ethyl-3-methyllumiflavinium perchlorate. In such cases, these highly unsaturated terpene compounds can be seen as fuel precursors for biofuel compounds. These saturated terpenes consist entirely of carbon and hydrogen. Terpenoid hydrocarbon compounds are much like aliphatic hydrocarbons in terms of having no compatibility issues with contemporary internal combustion-based transportation. Reports have indicated the suitability of using hydrogenated limonene and myrcene as diesel additives, indicating their ability to blend with fossil fuels. Furthermore, pure or blended terpene fuels can be run on current diesel and gasoline engines. Beyond the alkane biosynthesis pathways therefore, *A. sarcoides* possesses a range of biochemistries that may be of interest to the fuels and chemicals market.

5.4 Conclusion

With this approach, we are able to interrogate the six *A. sarcoides* genomes for biochemical potential. Reciprocal network analysis allow the generation of a non-redundant orthologous *in silico* proteome and the collation of annotation evidence from five annotation pipelines. We are able to propose fatty acid metabolism pathways, which we are able to separate from alkane biosynthesis. With this approach we are able to propose rational hypothetical pathways for alkane biosynthesis via a C(n-1) route via a fatty acyl reductase and a FDC1-like pathway for decarboxylation/decarbonylation (Figure 5.12). This includes a hypothetical pathway for even-chain alkane via a lipid hyperoxidation. Furthermore, we are able to elucidate the genetic component that encodes alkane degradation in *A. sarcoides*. This highlight a more complex interaction for alkane metabolism in *A. sarcoides* than previously thought. With this approach, no evidence of a cycloalkane pathway was found. However, by using retrosynthesis analysis, it suggests that a lipid hyper oxidation plays a key role and also suggest the number of step require for a hypothetical cycloalkane biosynthesis.

CHAPTER 6 HYDROCARBON PROFILING OF *A. SARCOIDES*

6.1 Introduction

Chapter five predicted a carbon-carbon lyase gene that contributes to alkane biosynthesis in all six isolates of *Ascocoryne sarcoides.* Prior to heterologous expression of the gene, the validity of previous reports describing *A. sarcoides* as an alkane producer must be confirmed. This chapter describes the use of two different methods to examine different alkane profiles of *A. sarcoide*s.

Linear alkane biosynthesis has been confirmed and characterised in many life forms including plants, insects, cyanobacteria and many prokaryotes. These are covered in detail in Section 1.2.1. For aliphatic hydrocarbons, the biosynthesis in fungi remains undescribed. These pathways have applications in advanced biofuels; the elucidation of their biosynthesis is important for designing novel synthetic hydrocarbon pathways for advanced biofuel production.

Aliphatic alkanes in the form of linear, branched and cyclic alkanes have previously been reported in filamentous fungi (Griffin et al., 2010, Shaw et al., 2015, Schoen et al., 2017). To date, seven different studies describe the fungus *A. sarcoides* NRRL 50072 (previously identified as *Gliocadium roseum*) as being capable of producing linear and cyclic hydrocarbons (Table 6.1). Of these seven studies (Table 6.1), five are affiliated to one another and only one study in the seven has explored the metabolome of other species and isolates of *Ascocoryne* (Griffin et al., 2010).

Table 6.1. Hydrocarbon metabolites detected from the cultures of *A. sarcoides* by HS-SPME-GC-MS. Metabolic data was collected from seven different studies. (x) = compounds identified by Spectral library analysis. (S) = compounds identified by pure standards. (*) = denotes results that includes volatile of *Ascocoryne* genus.

Compound	Stinson et al., 2001	Strobel et al., 2010	Griffin et al., 2010*	Ahamed et al., 2011	Malette et al., 2012	Gianoulis et al., 2012	Malette et al., 2014
1-Heptene			x				
1-Heptene, 6-methyl-							
1-Methyl cyclohexene			x				
1-Octene	S	S		S		S	
1,3-Octadiene	x						
1,3,5,7-Cyclooctatetraene	x						
1,4-Hexadiene, 3-ethyl-							x
2-Pentene			x				
3-Methyl nonane				S			
3-Nonene			x				
3,4-Dimethyl hexane				S			
3,5-Octadiene			x				
3,5-Octadiene (Z,Z)		x					
4-Decene, 9-methyl-	x						
4-Nonene			x				
Butane, 2-methyl- (isopentane)					x		
Cyclodecane			x				
Cyclodecene		x	x				
Cyclohexene, 4-methyl-		S					
Cyclopropane, propyl-		x					x
Decane, 2,2,6-trimethyl-		x					
Decane, 3,3,5-trimethyl-		x					
Dodecane, 2,7,10-trimethyl-		x					
Dodecane		S		S			x
Heptane			x	S		x	
Heptane, 2-methyl-		S					
Heptane, 4-methyl-					x		

Table 6.1. Hydrocarbon metabolites detected from the cultures of *A. sarcoides* by HS-SPME-GC-MS. Metabolic data was collected from seven different studies. (x) = compounds identified by Spectral library analysis. (S) = compounds identified by pure standards. (*) = denotes results that includes volatile of *Ascocoryne* genus.

Compound	Stinson et al., 2001	Strobel et al., 2010	Griffin et al., 2010*	Ahamed et al., 2011	Malette et al., 2012	Gianoulis et al., 2012	Malette et al., 2014
Heptane, 5-ethyl-2,2,3-trimethyl- (or isomer) Undecane, 4-methyl-		x					
Heptene		S					
Hexadecane				S			
Hexadecane (or isomer)		S					
Hexadecane (possible isomer)		x					
Hexane				S			
Hexane, 2,3,4-trimethyl-		x					
Hexane, 2,4-dimethyl- (possible isomer) Undecane, 2,6-dimethyl-		x					
Hexane, 3-methyl-							x
Hexane, 3,3-dimethyl-							x
Hexane, 3,3-dimethyl- (or isomer) Decane, 2,6,7-trimethyl-		x					
m-Xylene				S			
Nonadecane		S		S			
Nonane		x					
Nonane, 3-methyl-							
Nonane, 4,5-dimethyl-							
Octane		S	x	S		x	
Pentane			x		x		
Pentane, 1-iodo-		x					
Tridecane		S		S			
Undecane 4,7-dimethyl-		x					
Undecane, 3-methyl-							x
Undecane, 4,4,-dimethyl-		x					
Undecane		S					

Importantly, none of these studies specifically focus on alkanes as the compound of interest and have only investigated the volatilome of *A. sarcoides*. This is evident in the methods used by studies to investigate *A. sarcoides*, which only examined the headspace of *A. sarcoides* cultures. These studies relied heavily on extraction by headspace-solid phase micro-extraction (HS-SPME) which is complemented by gas chromatography and mass spectroscopy (GC-MS) to resolve extracted metabolites. HS-SPME is reliant on the physical properties of polydimethylsiloxane (PDMS) polymer, which is able to sorb volatile hydrophobic compounds. Another method that has been used is proton transfer reaction mass spectroscopy (Mallette et al., 2012, Mallette et al., 2014). With this technique, gaseous effluent from the culture headspace is introduced directly into the analytical equipment to facilitate ionisation of captured compounds by hydronium ions and enabled real-time sampling of volatile metabolites. There are disadvantages with these two methods when it comes to examining potential fuel compounds. Most fuel compounds have a boiling point that is higher than the incubation temperature of *A. sarcoides*, therefore would be present in higher quantities in the aqueous phase. Thus previous studies have only explored alkane presence in the gaseous phase of *A. sarcoides*. Investigation of other alkane producing organisms more frequently extract the aqueous phase of the organism by utilising solvent extraction.

Identifying alkanes is difficult within fungal organisms because fungi are capable of producing a diverse metabolite profile including higher alcohols, ketones, aldehydes and terpenes (Keller et al., 2005). Identification can, therefore, be difficult as a large number of compounds can cause misidentification of alkanes due to co-elution or elution of compound with highly similar spectral compounds. This was highlighted in studies that observed production of biogenic alkanes in biological systems (Schulz et al., 2007, Griffin et al., 2010). This was also the case for an earlier volatilome examination of *A. sarcoides*, in which a retraction was made to correct for some misidentified hydrocarbon metabolite (Strobel et al., 2010b). Alkane identification is confounded by the simplicity and highly repetitive elements associated with alkanes. Similar hydrocarbons and isomers such as octane, 1-octene, cyclooctane, and 1-methylheptane are difficult to resolve as these compounds manifest highly similar spectroscopic features and can lack a distinct molecular ion.

6.1.1 Aims and Objectives

In this chapter, two methods for identifying alkane biosynthesis were explored with the aim of investigating the alkane capabilities of *A. sarcoides*. The methods developed here and the implications for *A. sarcoides* as an alkane producer will be presented and discussed. To test these hypotheses, a suitable method must be developed for the detection for both linear and cyclic alkanes. Here we described:

- Developing a solvent extraction and GC-MS method to identify alkanes in cultures of *A. sarcoides*
- Developing a SBSE and GC-MS method to identify alkanes in cultures of *A. sarcoides*

6.2 Results

This study differs from previously reported metabolic examinations of *A. sarcoides* due to differences in analytical methods. The methods used in this study focus solely on the secretome. The rationale for this is that fuel compounds would have physicochemical properties such as boiling point, melting point, and vapour pressure appropriate for the incubation temperature of *A. sarcoides.* Thus, fuel molecules would be found in higher quantities in the aqueous phase. Two method swere investigated: Firstly, with extraction by solvents and, secondly, by stir bar sorptive extraction (SBSE) method (Baltussen et al., 1999), which is facilitated by a solvent-less Twister polydimethysiloxane-coated stir bar. Both extraction methods were analysed with gas chromatography mass spectroscopy analysis.

6.2.1 Extraction by solvent extraction

For solvent extraction, deuterated-chloroform ($CDCl_3$) was used as the extraction solvent and for resuspension after evaporation. $CDCl_3$ was chosen for its explicit spectroscopy signal when observing the chromatogram of the mass spectra extracts. This is to minimise misidentification between exogenous compound and biologically-derived compounds. A calibration curve was produced to predict retention time (RT) of potential linear alkanes by plotting RT against carbon length for each linear alkane standard (Figure 6.1). This calibration curve was able to produce a proportional calibration graph with a good line of best fit (R^2 = 0.9703). Calibration for cyclic alkane was not possible due to the lack of chemical standards.

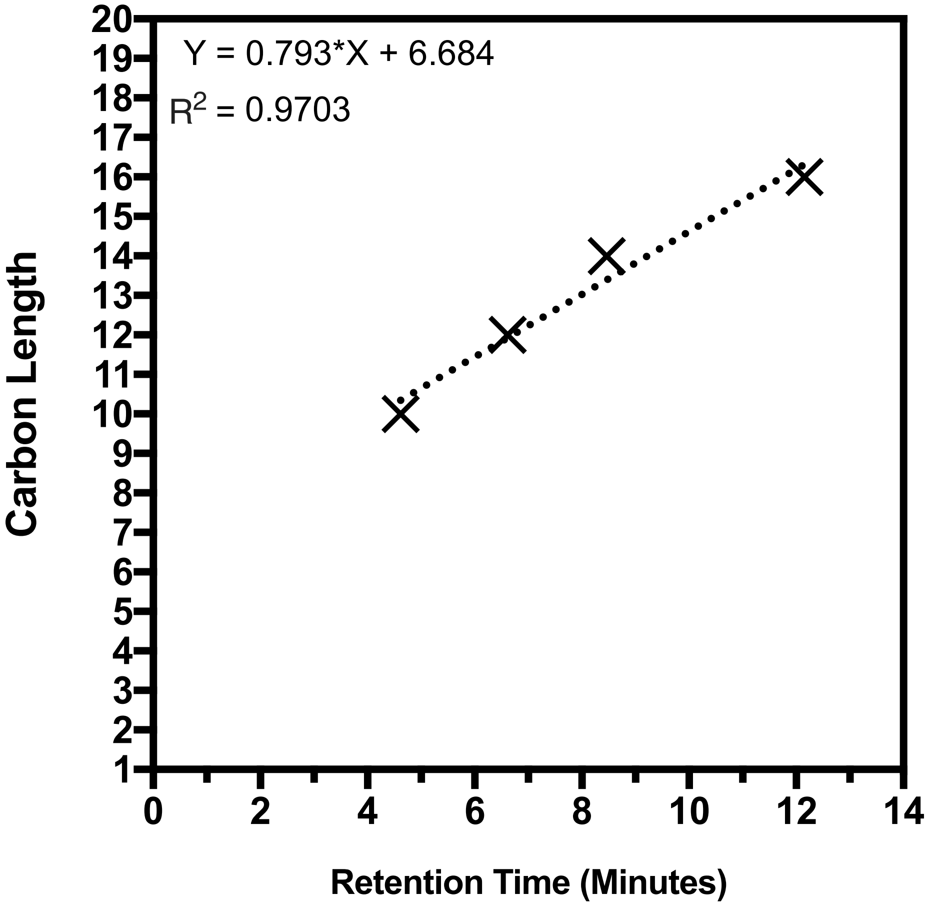


Figure 6.1. Retention time of linear alkane standards. Standards were further identified by NIST spectral library. The retention time was plotted against the carbon length of the standards.

Identification of alkane standards by solvent extraction

A mixture of alkane standards containing cyclodecane, decane, dodecane, tetradecane, and hexadecane was suspended with $CDCl_3$ and were spiked with a range of concentrations from 1.000 mg ml^{-1} to 0.001 mg ml^{-1}. This was to determine the limits of detection and peak RT of each alkane standard (Figure 6.2). Alkane peaks were identified by manually comparing each peak's spectral library to spectral standards. With split-less injection mode, alkane peaks were identifiable in range of concentrations from 1.000 mg ml^{-1} to 0.001 mg ml^{-1}. Lower concentrations at 0.0001mg were observable but peak area calculation was not reliable.

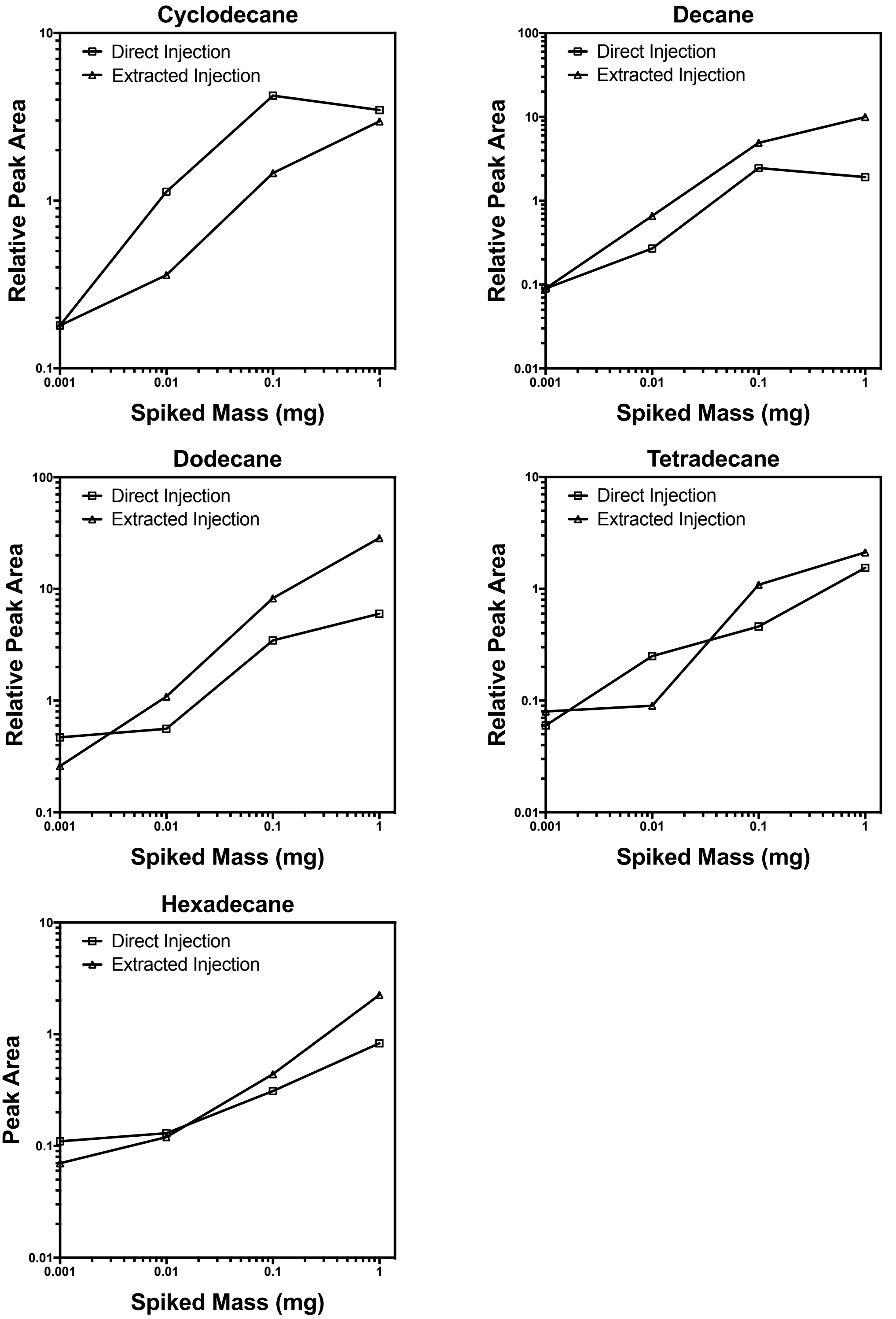


Figure 6.2. Direct injection and extracted injection of alkane standards. A concentration range of 1 mg to 0.001 mg of alkane mix composed of cyclodecane, decane, dodecane, tetradecane, and hexadecane were injected spit-less into the GC-MS. Defined medium was also spiked with the alkane mix and was extracted and injected spit-less into the GC-MS. Standards were further identified by NIST

Extraction efficiency by solvent extraction

To test the extraction efficiency of $CDCl_3$ against sterile medium background, a range of alkane standards from 1.000 mg ml^{-1} to 0.001 mg ml^{-1} were spiked into sterile defined medium. Extraction was performed according to Section 2.4.1. Extracts were injected into the GC-MS by split-less injection and the peak area was calculated (Figure 6.2). The extraction efficiency was calculated from peak areas of appropriate alkane standards from 6.2.1. It is thought that the discrepancy in extraction efficiency over 100% was due to pipetting accuracy and evaporation of the solvent and was omitted from extraction calculations. Nonetheless, the extraction method was able to achieve an efficiency value of 59%.

Preliminary biogenic metabolite screening with solvent extraction

Isolates that were cultured in chemically-defined medium for 14-day culture were extracted as described in Section 2.4.1. Extracts of sterile media samples and isolate 64019 were injected into the GC-MS by split-less injection (Figure 6.3). Under these conditions, no alkane equivalents to chemical standards were detected at the expected RT above the concentration of 0.001 mg ml^{-1} in three isolates of *A. sarcoides* in a screening with six replicates.

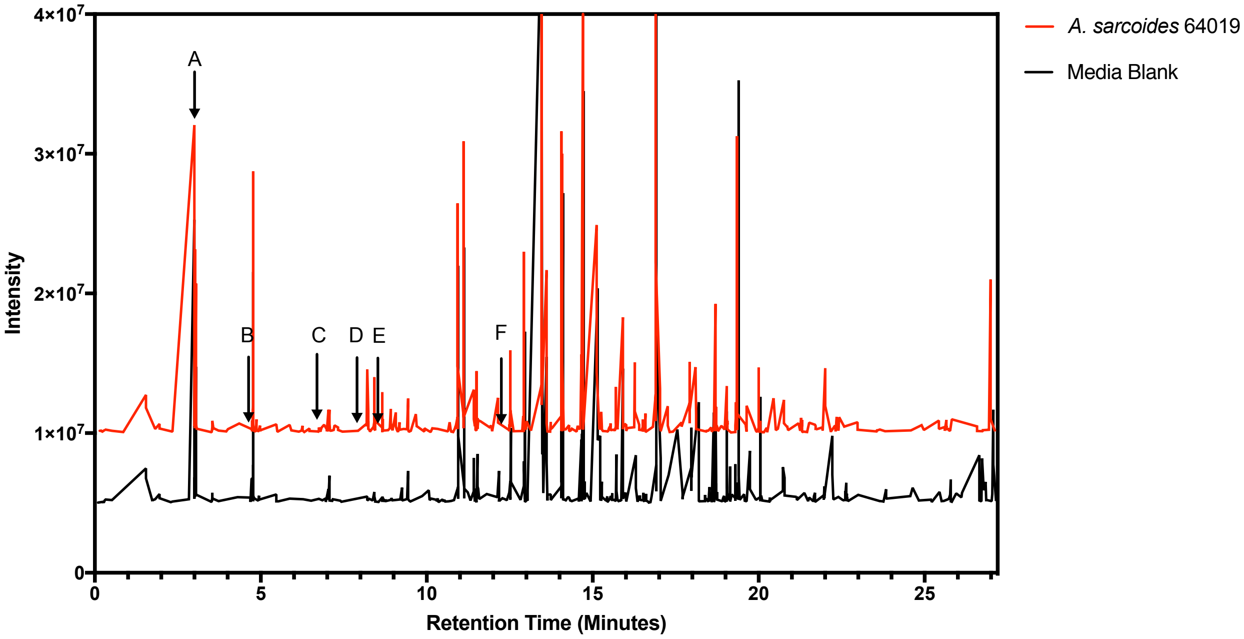


Figure 6.3. Solvent extraction of *A. sarcoides* 64019 cultures and media blank into GC-MS. Expected retention times were derived from chemical standards and are highlighted by the following arrows: A) Solvent peak B) Octane C) Decane, D) Cyclodecane, E) Tetradecane, and F) Hexadecane. Chromatogram is representative of six replicates.

6.2.2 Extraction by SBSE

As the solvent extraction method was not able to detect alkane at the concentration of 1.000 mg ml^{-1} to 0.001 mg ml^{-1} from cultures of *A. sarcoides*, we endeavoured to develop a more sensitive method to further test our hypothesis. The SBSE method relies on a Twister stir bar, a magnetic stir bar coated with PDMS, which is a material that is capable of sorbing hydrophobic and lipophilic molecules from aqueous samples (Baltussen et al., 1999). This was used for the extraction of metabolites from *A. sarcoides* culture. This method was favoured over solvent extraction methods for three reasons. Firstly, this method enables time course extraction that is unachievable with conventional solvent extraction as the producing organism does not survive the solvent extraction processes. Secondly, compared to conventional solvent extraction methods, this extraction method can be considered to be high throughput as extraction is highly parallelised and automation of injection and handling can be automated by the GC-MS. Lastly, the method can be standardised for the extraction and injection process to minimise operator error, thus should improve extraction efficiency. To the best of our knowledge, SBSE has not been used previously to examine metabolite production in biological cultures. However, there are quantitative limitations to the usage of SBSE. Metabolites that are extracted and detected by SBSE reflect the chemical composition adsorbed to the PDMS coating. Ideally, the adsorbed chemical composition should reflect the culture of the medium. Issues such as preferential adsorption may lead to over-saturation of the PDMS fibre which can lead to quantitative biases. Another bias that can manifest is the carry over effect. Carry over in this context can be considered as left over compounds that were present after purging of the SBSE stir bar. Compounds that are strongly adsorbed are more likely to be present after thermal purging.

Identification of alkane standards by SBSE

To examine the performance and sensitivity of the GC-MS, we injected a mixture of alkane standards. Eight newly obtained chemical standards were used for identification of alkanes. Five of these are linear alkane standards: octane, nonane, decane, dodecane, tetradecane, and hexadecane. For cyclic alkanes, cyclooctane and cyclododecane were used. To examine the sensitivity of the GC-MS, a 10μL standard alkane mix at the concentration of 100 ppb was injected using Automated Tube Exchange (ATEX), placed in the thermal desorption unit coupled to the GC-MS and then desorbed thermally to be injected into the GC-MS using the same method used for the Twister stirbars. A chromatogram was produced and the results are shown in Figure 6.4 and 6.5 and with RT and National Institute of Standards and Technology (NIST) spectra prediction in Table 6.2.

Spectral analysis was able to confirm the identity of all alkane standards (Figure 6.4). Initial cutoff for mass ion began at 50. At 100 ppb, the most abundant mass ions were 71 and 85 in all alkane standards. For compounds with a longer chain length, the total mass ion was noticeable. For linear alkanes, the total mass ion was detected for C14 and beyond, while for cyclic alkanes, total mass ion was present at C12 and not at C8. As not all linear alkane standards were available, a calibration graph (Figure 6.6) was constructed to predict the RT of other linear alkanes. An R^2 value of 0.9986 indicated a good fit and proportionality for the calibration curve. The calibration curve was used to predict RT of other linear alkanes in later analysis.

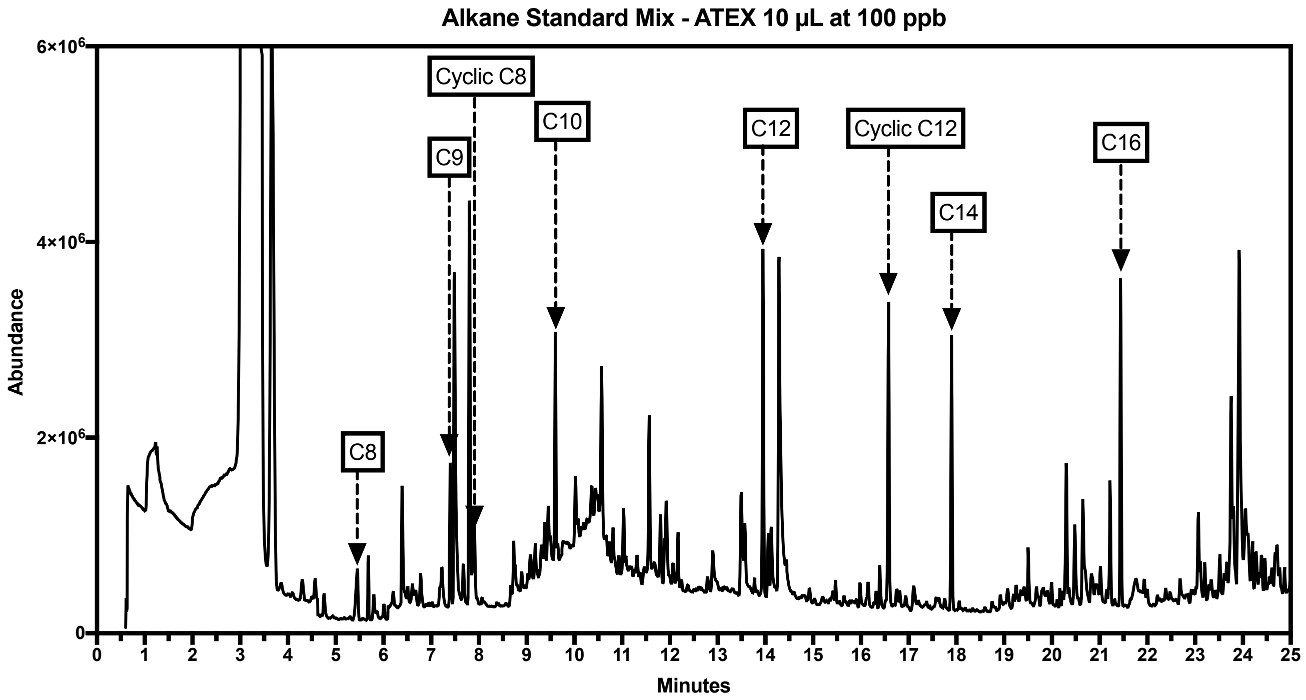


Figure 6.4. Alkane standard mix injected by using Automated Tube Exchange (ATEX) into GC-MS. The standard mix is composed of six linear alkane standards (octane, nonane, decane, dodecane, tetradecane, and hexadecane) and two cyclic alkane standards (cyclooctane and cyclododecane). Concentration of each standard was diluted to 100 ppb.

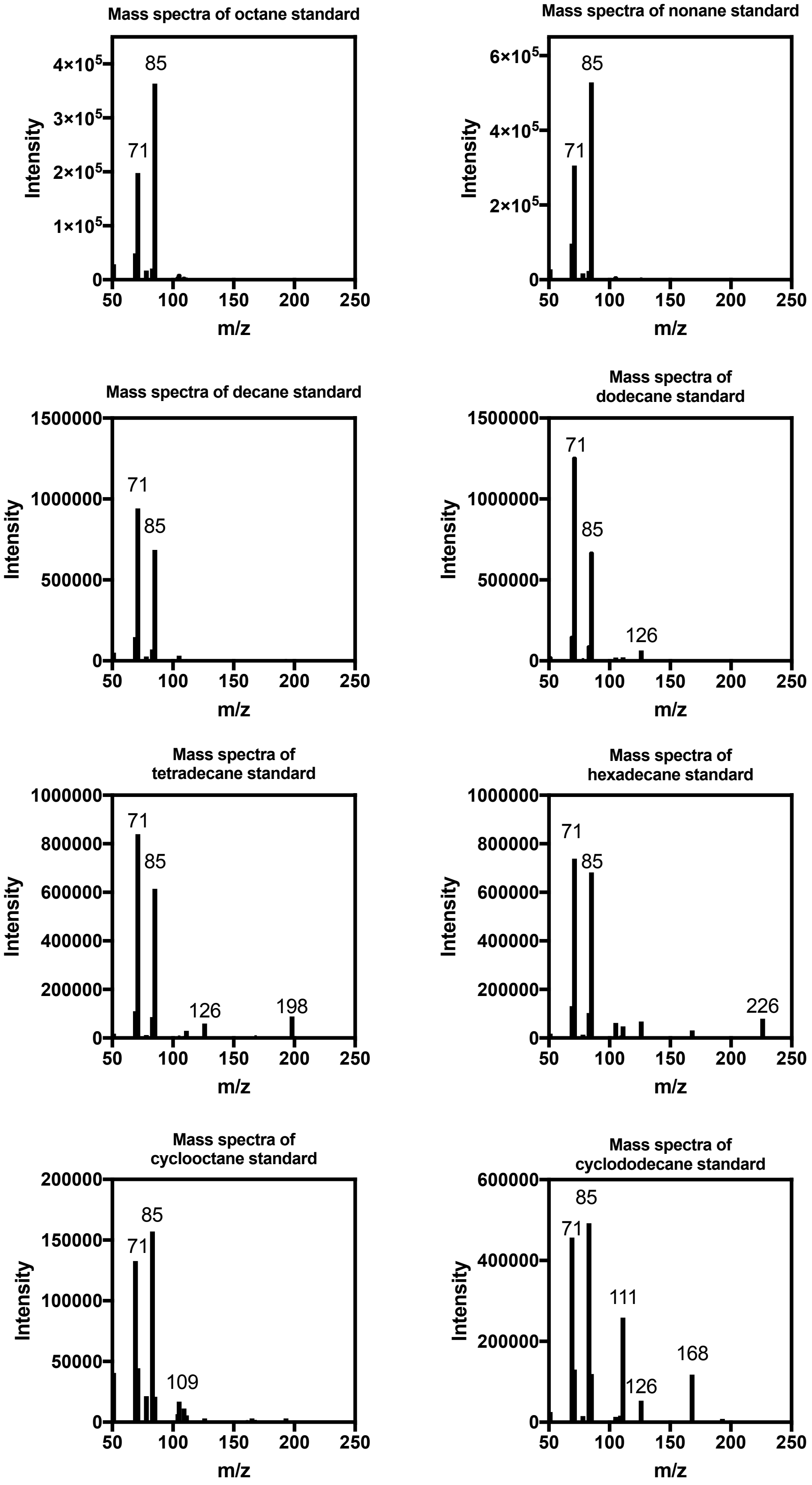


Figure 6.5. Mass spectra of alkane standards. Mass spectral fragmentation of alkane standards. Molecular ion is included when detected.

Table 6.2. Alkane standard mix injected by using Automated Tube Exchange (ATEX) into GC-MS.

NIST Prediction	Formula	Retention Time (Minutes)
Octane	C8H18	5.44
Nonane	C9H20	7.40
Cyclooctane	C8H16	7.88
Decane	C10H22	9.59
Dodecane	C12H26	13.94
Cyclododecane	C12H24	16.57
Tetradecane	C14H30	17.89
Hexadecane	C16H34	21.43

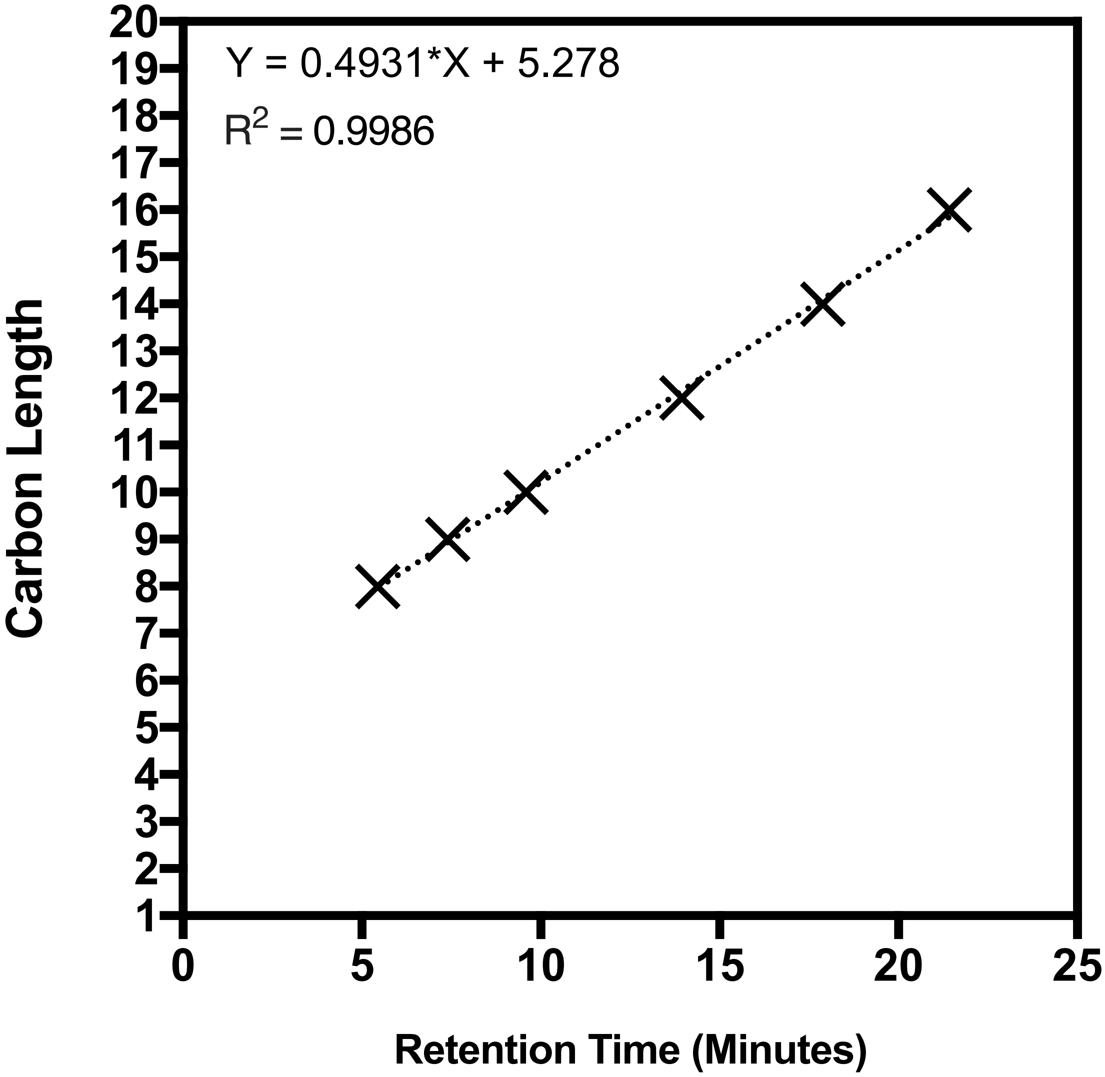


Figure 6.6. Linear alkane retention time calibration curve with a line of best fit. The calibration curve is based on six identified standard linear alkanes from the length of C8 to C16. R^2 value indicate a good fit and proportionality for the line of best fit.

Identification of stir bar-related peaks

The exterior of Twister stir bars are covered with a polymer mainly composed of PDMS. It is a material that is commonly used in GC columns and, when subjected to wear and tear, it can release compounds into the column. When this occurs, unwanted peaks can be detected on the chromatograms. Thus, it is important to identify these contaminant peaks to separate them from biogenic peaks. Peaks were observed in all Twister stir bar related chromatogram. Figure 6.7a is a chromatogram of a stir bar post-purging and did not come into contact with the experimental system, 6.7b is a sterile media sample and 6.7c, 6.7d and 6.7e are representative chromatograms of biological samples at 1, 2, and 3 weeks respectively. As indicated by figure 6.6, C8 to C16 linear alkane were observed to be in the RT range of 0 to 25 minutes. In this region, three peaks have been identified and these peaks are proportional to each other and constant in all sample.

Spectra prediction with the NIST library suggested these peaks to be polyaromatic hydrocarbons with similar mass-ion spectra (Figure 6.8). 6.8a was identified to be a Flourene-derivative, 6.8b is thought to be a Bisazulene-derivative and 6.8c was identify as Pentapentadecafulvalene-derivative. In each of these spectra, the mass-ion 193 is the most significant peak and the ion was associated with contaminant peaks for later analysis. Furthermore, there are unique peaks found across all spectra, which indicate consistent bleeding of polyaromatic hydrocarbons from the stir bar. No silane peaks were identified consistently, suggesting a limited silane bleeding from the Twister stir bar.

As these are Twister stir bar-specific peaks and are present across all SBSE samples, these can be use to quantify shifts in RT across different samples and batches. Figure 6.9 indicated RT shifting in all previously identified peaks in all batches of sterile media samples.

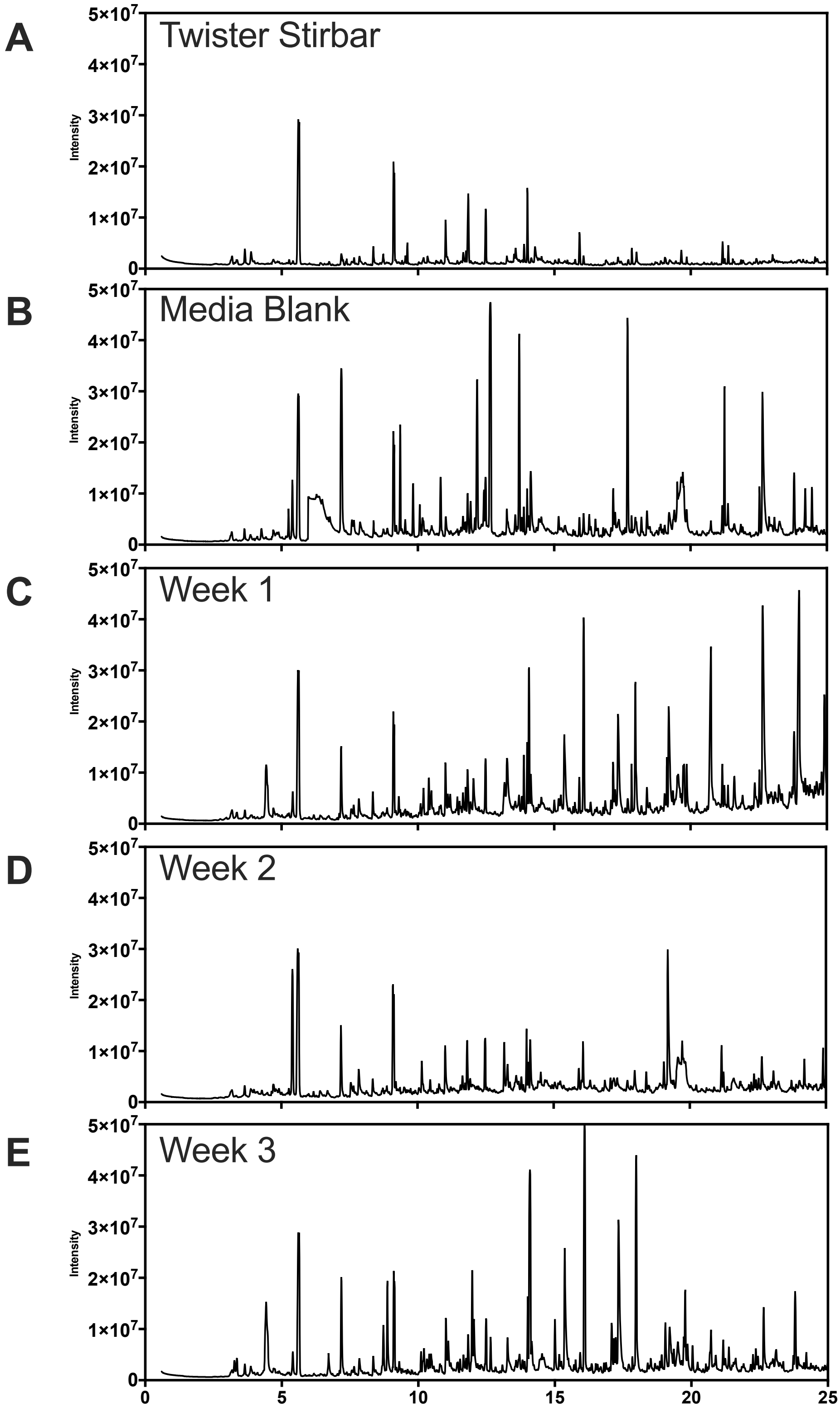


Figure 6.7. Identified twister stir bar contaminant peaks. These three peaks are consistent across all samples and are relatively proportional to each other A. Chromatogram of blank twister stir bar. B. Chromatogram of sterile media sample. C, D, and E correspond to *A. sarcoides* 64019 weekly sample 1, 2, and 3 respectively.

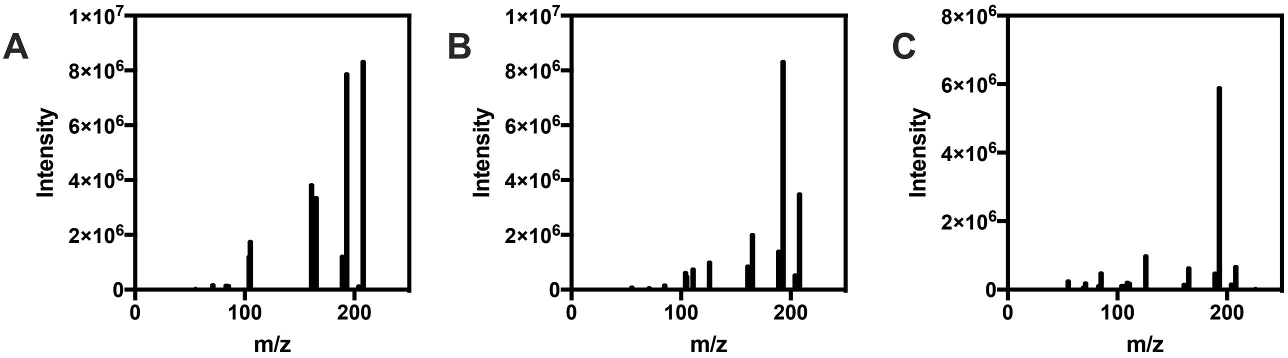


Figure 6.8. Spectra of contaminant peaks. These three peaks are consistent across all samples. NIST spectra prediction indicate that A) Flourene-derivative, B) Bisazulene-derivative and C) Pentapentadecafulvalene-derivative. These spectras are consistent with compounds with highly aromatic rings.

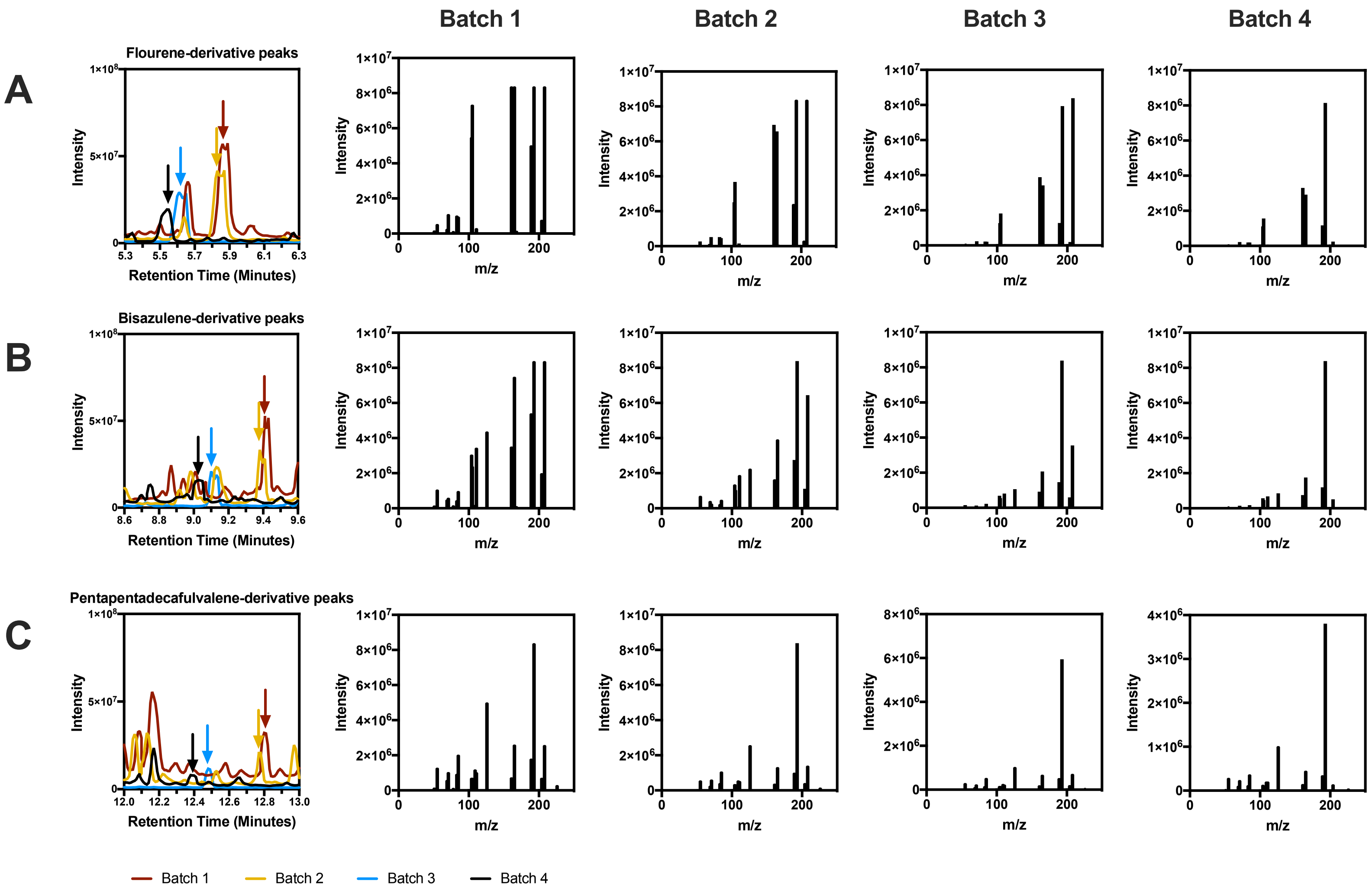


Figure 6.9. Quantification of retention time shift across all SBSE batches. Twister-related contaminants peaks was identified across batches and was used as internal standards to quantify retention time shift. Row indicate A) Flourene-derivative, B) Bisazulene-derivative and C) Pentapentadecafulvalene-derivative, peaks and spectra. Column compares batch differences.

Preliminary screening of A. sarcoides with SBSE

Initial preliminary screening with the SBSE method showed promising results. Three isolates (*A. sarcoides* 170.56, 246.80, and 44013) were cultured for three weeks, with 30 minutes incubation with a Twister stir bar at the start of culturing and with another incubation at the end of three weeks. Spectral comparison to mass spectral libraries indicated that the end of culture sampling contained more linear alkane and cyclic alkane compounds than cultures sampled at the start of incubation (Table 6.3). This suggests that the SBSE method is able to detect alkanes from cultures of *A. sarcoides*. To see if alkanes were extracted, 1 mg of alkane mix containing cyclodecane, decane, dodecane, tetradecane, and hexadecane were spiked into sterile media. When desorbed into the GC-MS, they were recovered in the GC-MS. In sterile media sample chromatograms, no significant amounts of alkane were detected within the RT range of alkane standards. Spectral predictions indicated the presence of hydrocarbons which are thought to be from the background of the medium. Furthermore the number of hydrocarbon compounds present in sterile media sample was less than the numbers of compounds detected in biological and spiked samples (Table 6.3).

Table 6.3. Preliminary screening of *A. sarcoides* isolate 170.52, 309.71, and 44013 by SBSE method. Isolates were sampled by SBSE at the start of culture and at the end of experiments (three weeks). Compound were identified by NIST spectral library analysis.

Sample	Alkanes	Alkenes	Alkynes	Cyclic alk(a/e)nes	Aromatic	Total
Sterile media	8	1	0	1	0	10
Spiked media	13	3	0	31	8	55
309.71 Start	1	3	0	17	9	30
309.71 End	4	3	0	4	4	15
44013 Start	5	1	0	5	3	14
44013 End	10	3	1	13	0	29
170.53 End	0	7	2	15	3	27

Identification of biogenic metabolites by SBSE

A. sarcoides isolates 170.56, 171.56, 246.80, 309.71, 44013, and 64019 were cultured as described in section 2.4.5. Additional care was taken to reduce exogenous alkane contamination and is further detailed in section 2.4.4 and 2.4.5. Additionally, two sterile media samples and hydrocarbon-spiked sterile media samples were cultured with the biological samples. Every seven days, the Twister stir bar were replaced in all samples and cultures were incubated for three weeks. The experiment was repeated three times. Linear alkanes, cyclic alkanes (Figure 6.10 to 6.15) and hydrocarbon-based (Table 6.4) terpenes were observed in all isolates of *A. sarcoides*. Alkanes were identified manually by two different methods. The first method was based on RT and mass spectral fragmentation of alkane standard identified by ATEX and spiked media samples. The second method was based on RT predicted by the calibration curve in figure 6.5. Peaks that do not correlate to RT described previously were identified by spectra analysis against spectral standard libraries. Terpenes were identified by comparing mass ion spectra to a library of spectral standards.

Table 6.4. Examples of sesquiterpenes detected by SBSE methods within biological samples and non-biological samples. Sesquiterpenes were identified via analysis by standard library of spectra. Note that this list is a representative list and is a non-exhaustive list from Batch 2.

Compound Name	Formula	Component RT	Library	Present in non-biological samples	Number of Occurrence
alpha-Cubebene	C15H24	17.2	wiley7n.l	N	3
alpha-Copaene	C15H24	17.65	wiley7n.l	N	9
gamma-Terpinene	C15H24	10.22	NIST17.L	Y	18
beta-Myrcene	C15H24	9.58	NIST17.L	Y	12
alpha-Phellandrene	C15H24	9.98	NIST17.L	Y	18
Germacrene D	C15H24	19.73	NIST17.L	N	4

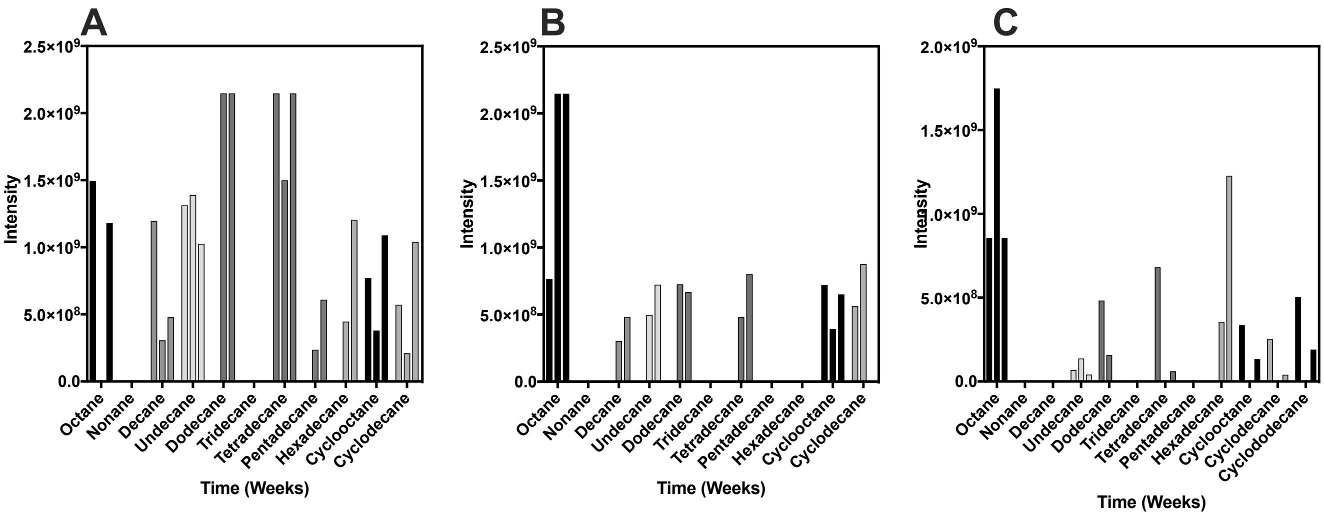


Figure 6.10. Peak area of identified alkanes sampled in the cultures of *Ascocoryne sarcoides* 170.56. Twister stir bar was incubated for a week. In total the cultures were incubated with three twister stir bars. Column represents peak area of identified alkane and columns within grouping indicates weekly differences in relative quantity of identified alkanes. A = Batch 1, B = Batch 2 and C = Batch 3.

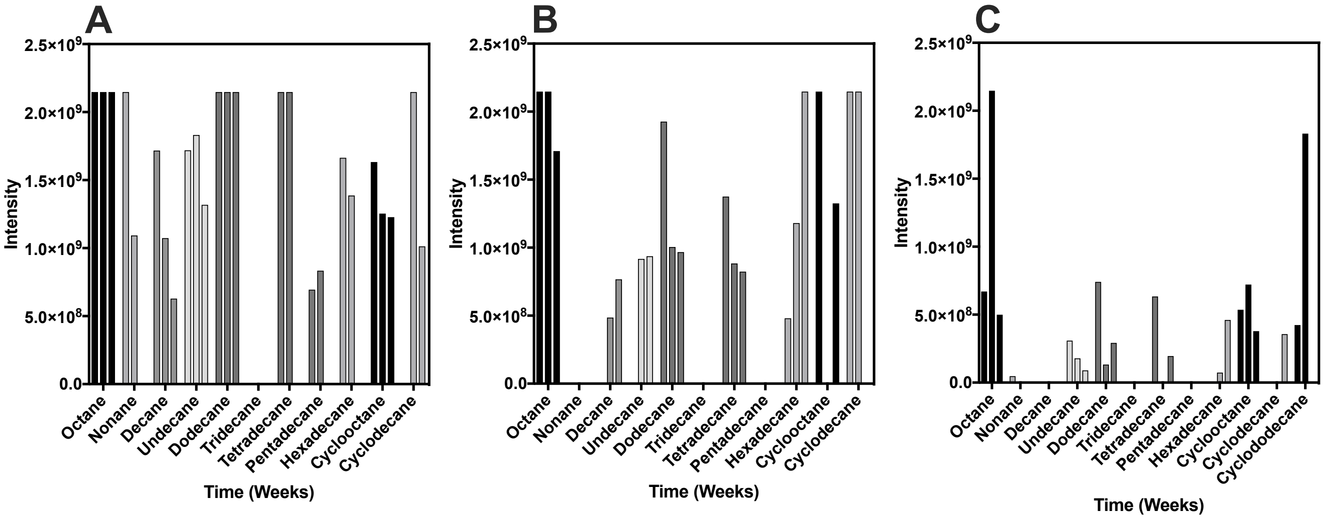


Figure 6.11. Peak area of identified alkanes sampled in the cultures of *Ascocoryne sarcoides* 171.56. Twister stir bar was incubated for a week. In total the cultures were incubated with three twister stir bars. Column represents peak area of identified alkane and columns within grouping indicates weekly differences in relative quantity of identified alkanes. A = Batch 1, B = Batch 2 and C = Batch 3.

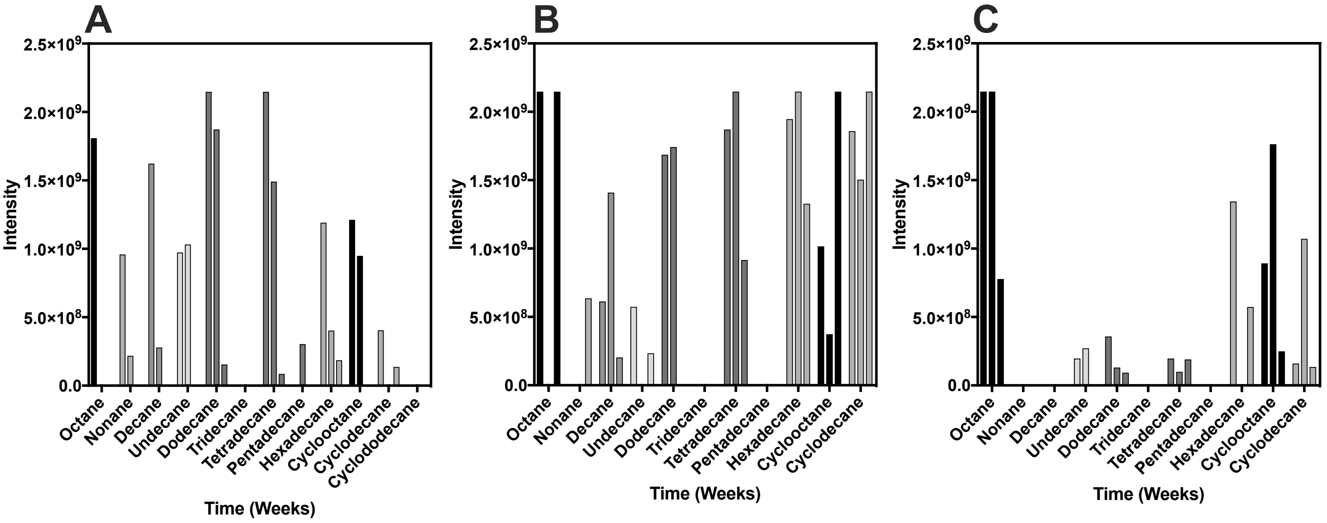


Figure 6.12. Peak area of identified alkanes sampled in the cultures of *Ascocoryne sarcoides* 246.80. Twister stir bar was incubated for a week. In total the cultures were incubated with three twister stir bars. Column represents peak area of identified alkane and columns within grouping indicates weekly differences in relative quantity of identified alkanes. A = Batch 1, B = Batch 2 and C = Batch 3.

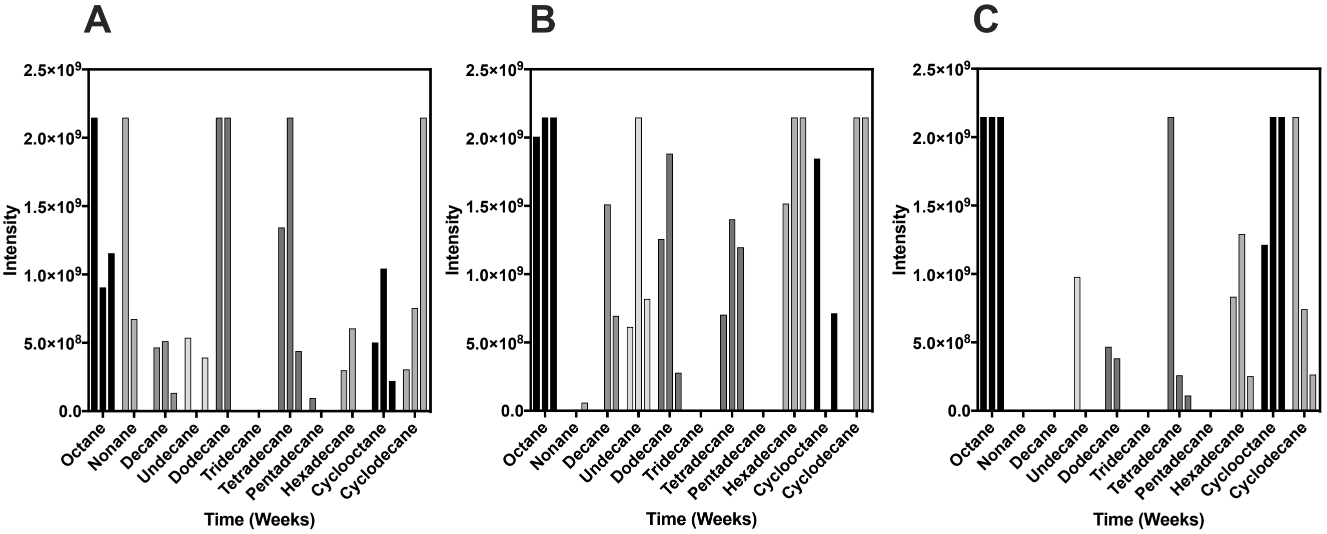


Figure 6.13. Peak area of identified alkanes sampled in the cultures of *Ascocoryne sarcoides* 309.71. Twister stir bar was incubated for a week. In total the cultures were incubated with three twister stir bars. Column represents peak area of identified alkane and columns within grouping indicates weekly differences in relative quantity of identified alkanes. A = Batch 1, B = Batch 2 and C = Batch 3.

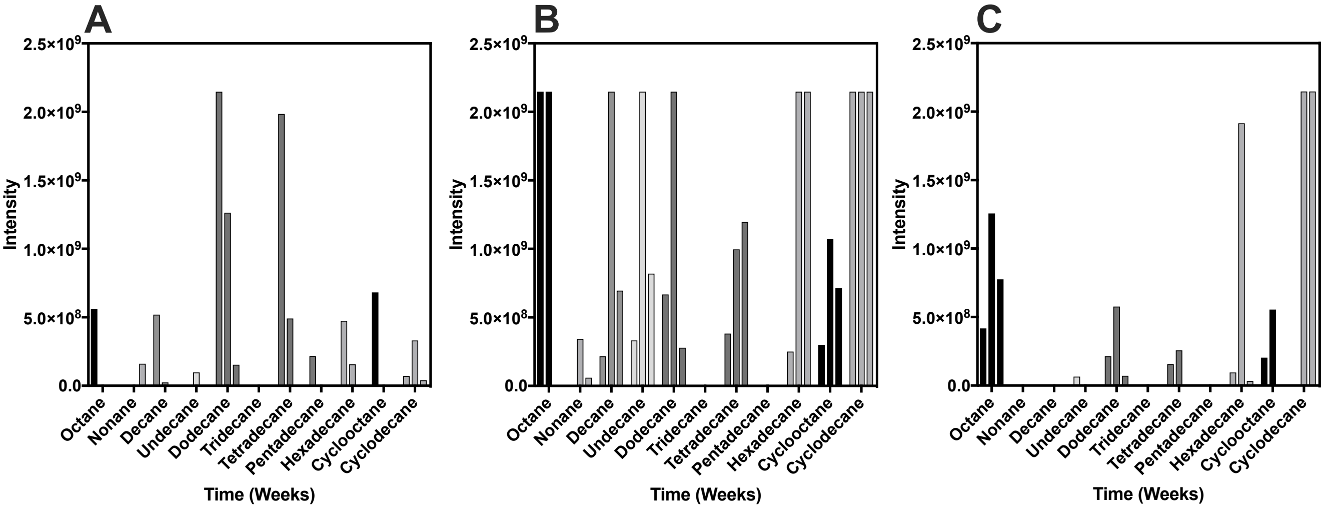


Figure 6.14. Peak area of identified alkanes sampled in the cultures of *Ascocoryne sarcoides* 44016. Twister stir bar was incubated for a week. In total the cultures were incubated with three twister stir bars. Column represents peak area of identified alkane and columns within grouping indicates weekly differences in relative quantity of identified alkanes. A = Batch 1, B = Batch 2 and C = Batch 3.

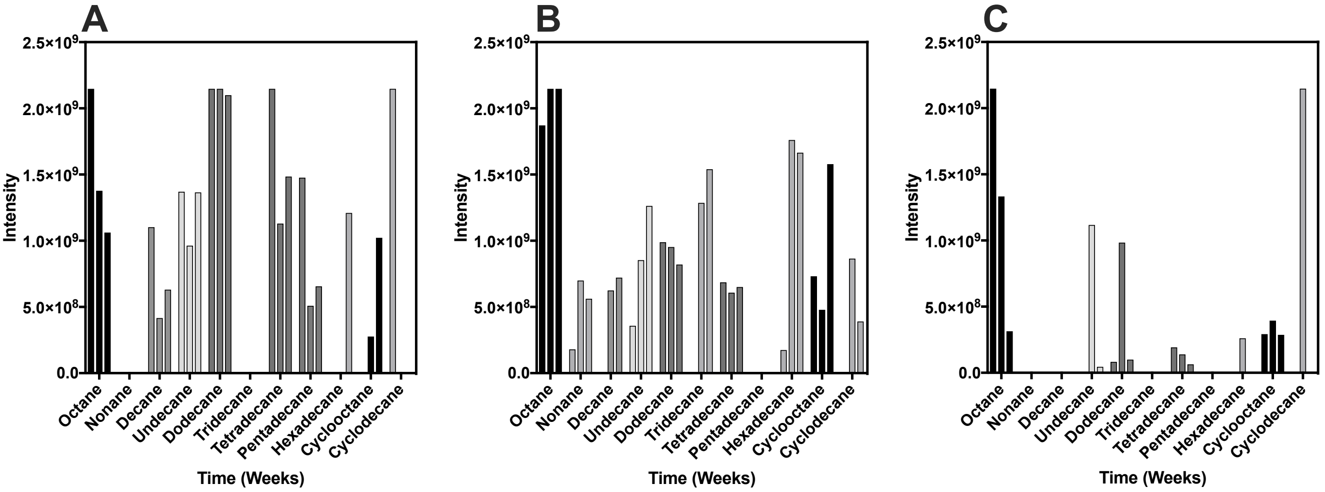


Figure 6.15. Peak area of identified alkanes sampled in the cultures of *Ascocoryne sarcoides* 64019. Twister stir bar was incubated for a week. In total the cultures were incubated with three twister stir bars. Column represents peak area of identified alkane and columns within grouping indicates weekly differences in relative quantity of identified alkanes. A = Batch 1, B = Batch 2 and C = Batch 3.

Identification of biogenic metabolites in non-producers samples

Two fungal linear alkane non-producers *Aspergillus nidulans* FGSCa4 and *Saccharomyces cerevisiae* S288c were also investigated. These fungi were chosen as the metabolism are well characterised and they have well characterised genomes which enabled bioinformatic analysis. To compare the metabolomes with *A. sarcoides*, both organisms were cultured under the same conditions and methods to *A. sarcoides* (Section 2.4.5). The chromatograms of *A. nidulans* and *S. cerevisiae* contained fewer alkane compounds but at a quantity equivalent with *A. sarcoides* samples (Figure 6.16 and 6.17).

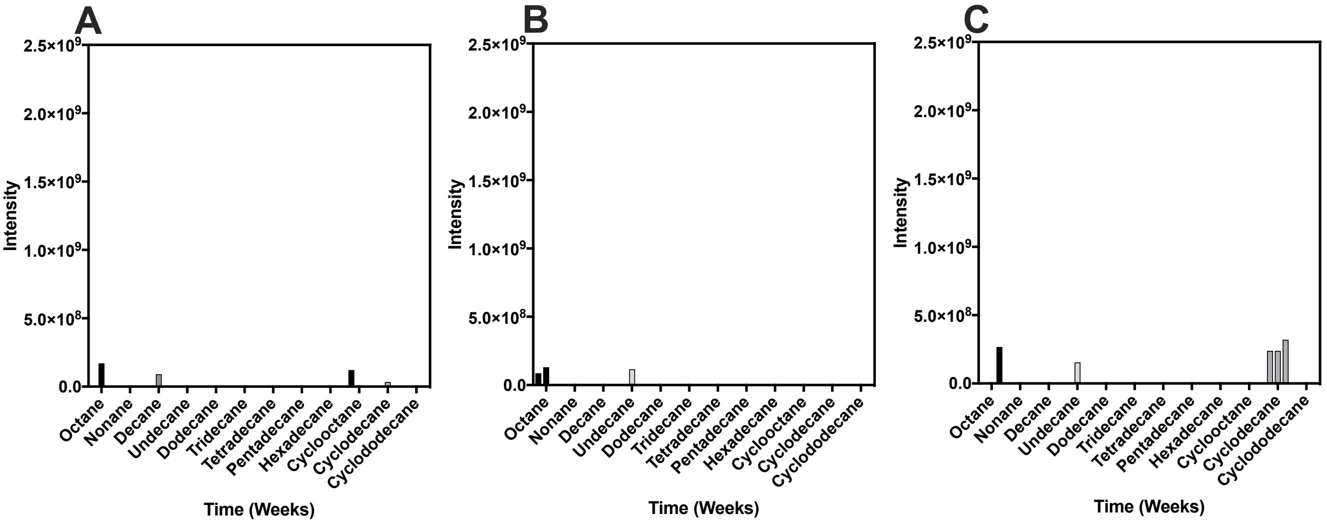


Figure 6.16. Peak areas of identified alkanes sampled in the cultures of alkane non-producer *Aspergillus nidulans FGSCa4.* Twister stir bar incubated for a week. In total the cultures were incubated with three twister stir bars. Column represents peak area of identified alkane and columns within grouping indicates weekly differences in relative quantity of identified alkanes. A = Batch 1, B = Batch 2 and C = Batch 3.

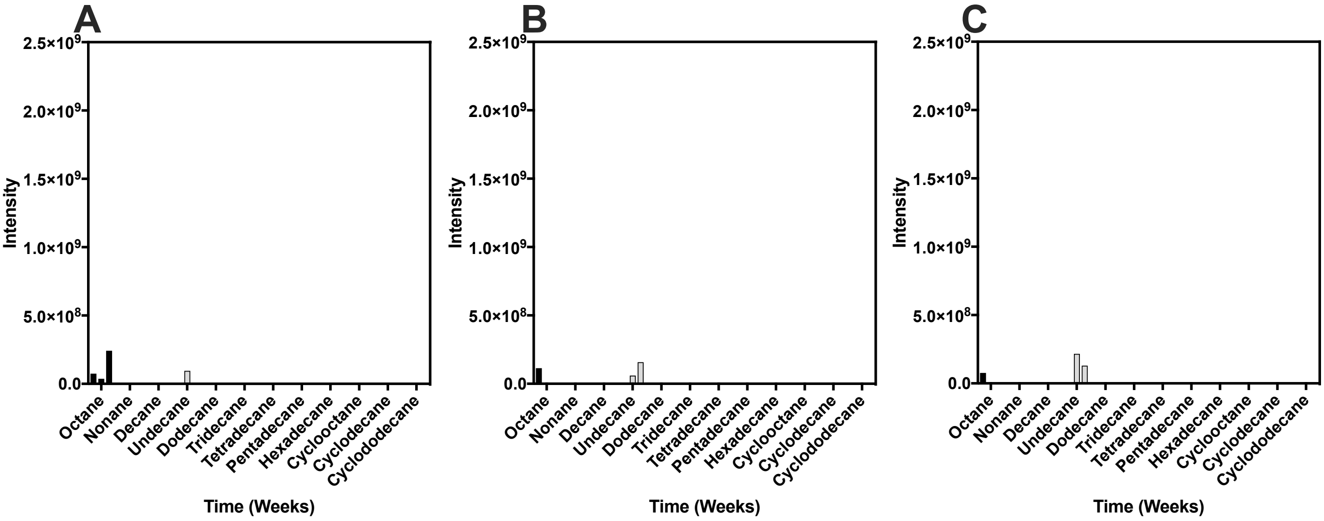


Figure 6.17. Peak area of identified alkanes sampled in the cultures of alkane non-producer *Saccharomyces cerevisiae S288c.* Twister stir bar incubated for a week. In total the cultures was incubated with three twister stir bars. Column represents peak area of identified alkane and columns within grouping indicates weekly differences in relative quantity of identified alkanes. A = Batch 1, B = Batch 2 and C = Batch 3.

Identification of metabolites in sterile media and spiked media samples

Two sterile media samples were included in each batch and treated identically to biological samples. Spiked samples were also included in all batches but were only incubated with a Twister stir bar for a week. The spiked samples contained the comparable level of alkanes in relation to *A. sarcoides* samples of the spiked alkanes (Figure 6.18), however, it was observe that sterile media samples also contained the same level of alkanes to spiked alkanes and *A. sarcoides* cultures (Figure 6.19).

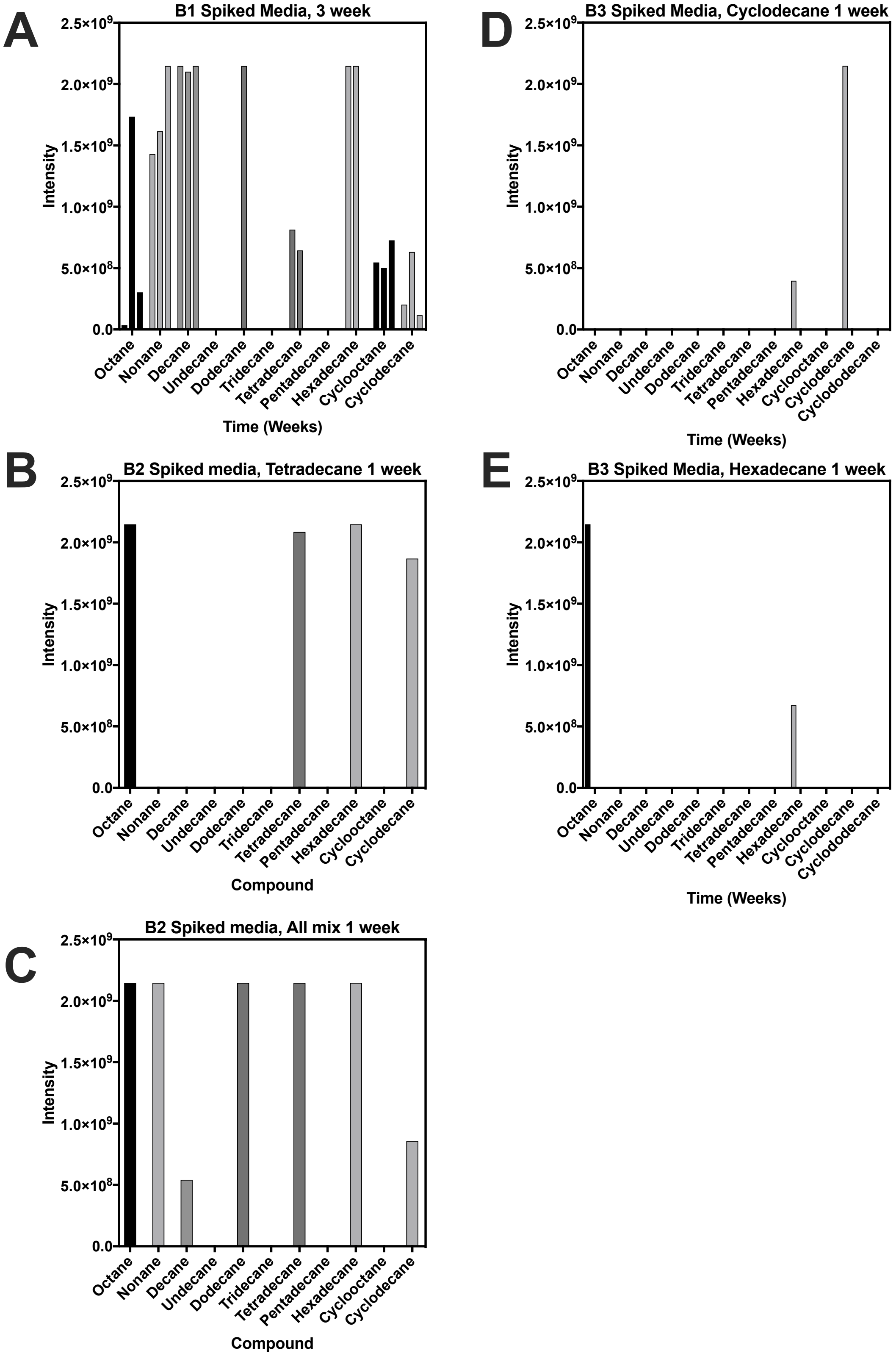

A
B1 Spiked Media, 3 week
Intensity
2.5×10⁹
2.0×10⁹
1.5×10⁹
1.0×10⁹
5.0×10⁸
0.0
Octane
Nonane
Decane
Undecane
Dodecane
Tridecane
Tetradecane
Pentadecane
Hexadecane
Cyclooctane
Cyclodecane
Time (Weeks)
D
B3 Spiked Media, Cyclodecane 1 week
Intensity
Octane
Nonane
Decane
Undecane
Dodecane
Tridecane
Tetradecane
Pentadecane
Hexadecane
Cyclooctane
Cyclodecane
Cyclododecane
Time (Weeks)
B
B2 Spiked media, Tetradecane 1 week
Intensity
Compound
E
B3 Spiked Media, Hexadecane 1 week
Intensity
Time (Weeks)
C
B2 Spiked media, All mix 1 week
Intensity
Compound

Figure 6.18. Peak area of identified alkanes in sterile media spiked with alkanes. Twister stir bar was incubated for a week. In total the cultures were incubated with three twister stir bars. Column represents peak area of identified alkane and columns within grouping indicates weekly differences in relative quantity of identified alkanes. A = Batch 1, spiked standard alkane mix of Octane, Nonane, Decane, Tetradecane, Hexadecane, Cyclooctane, Cyclodecane and Cyclododecane at 1 mg/ml for 3 weeks, B = Batch 2 spiked tetradecane at 1mg/ml for 1 week, C = Batch 3 spiked standard alkane mix of Octane, Nonane, Decane, Tetradecane, Hexadecane, Cyclooctane, Cyclodecane and Cyclododecane at 1 mg/ml for 1 week, D = Batch 2 spiked Cyclodecane at 1mg/ml for 1 week and E = Batch 2 spiked hexadecane at 1mg/ml for 1 week.

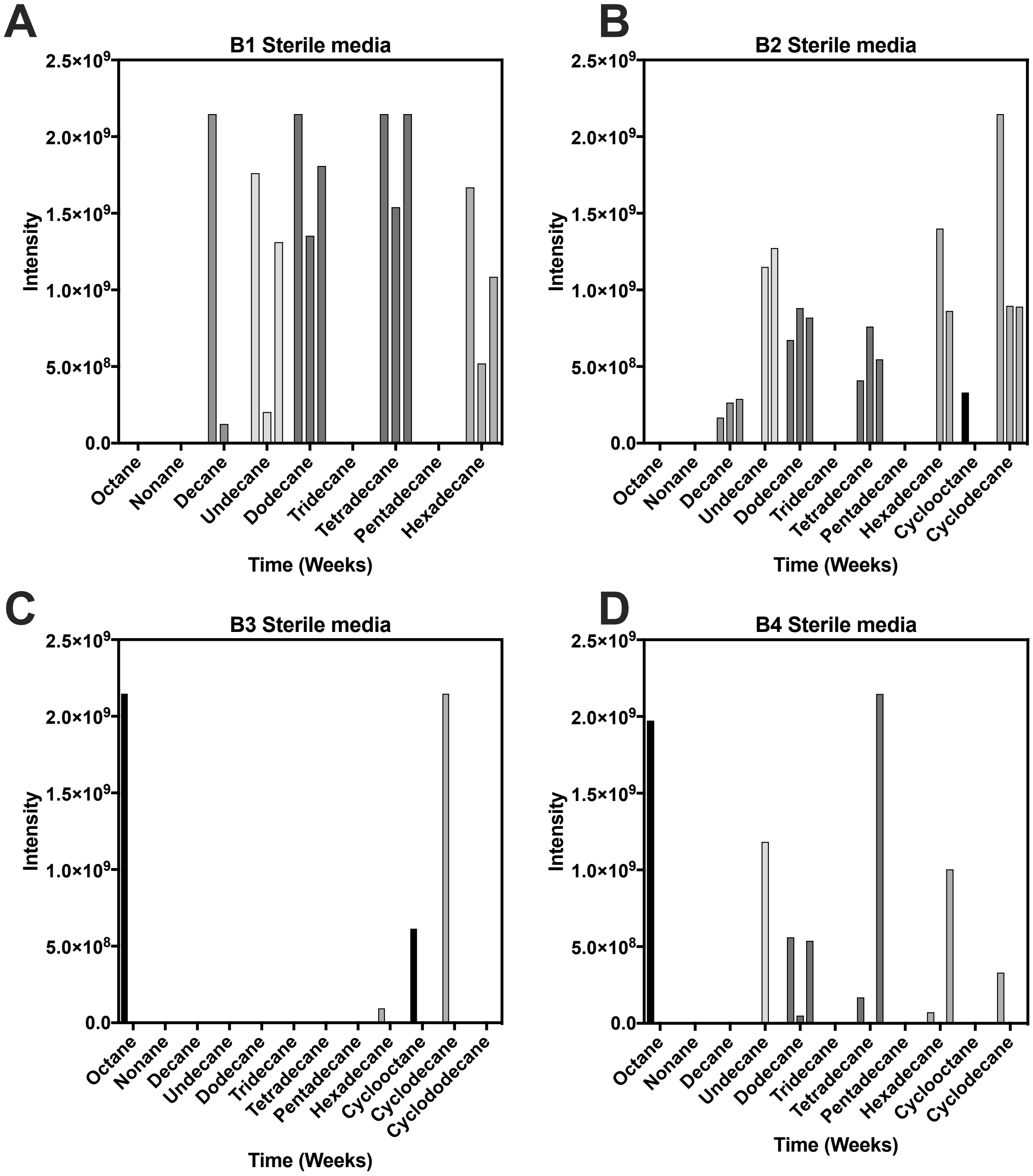


Figure 6.19. Peak area of identified alkanes in sterile media*.* Twister stir bar was incubated for a week. In total the cultures were incubated with three twister stir bars. Column represents peak area of identified alkane and columns within grouping indicates weekly differences in relative quantity of identified alkanes. A = Batch 1, B = Batch 2, C = Batch 3 and D = Batch 4

Identification of metabolites from thermally desorbed stir bars

Suspecting the stir bar as a carrier of contamination, three randomly selected stir bars were sampled by GC-MS. Alkanes were detected at a comparable level of quantity to *A. sarcoides*, spiked media and sterile media samples (Figure 6.20). This suggests that the effects of carry-over contaminated SBSE samples.

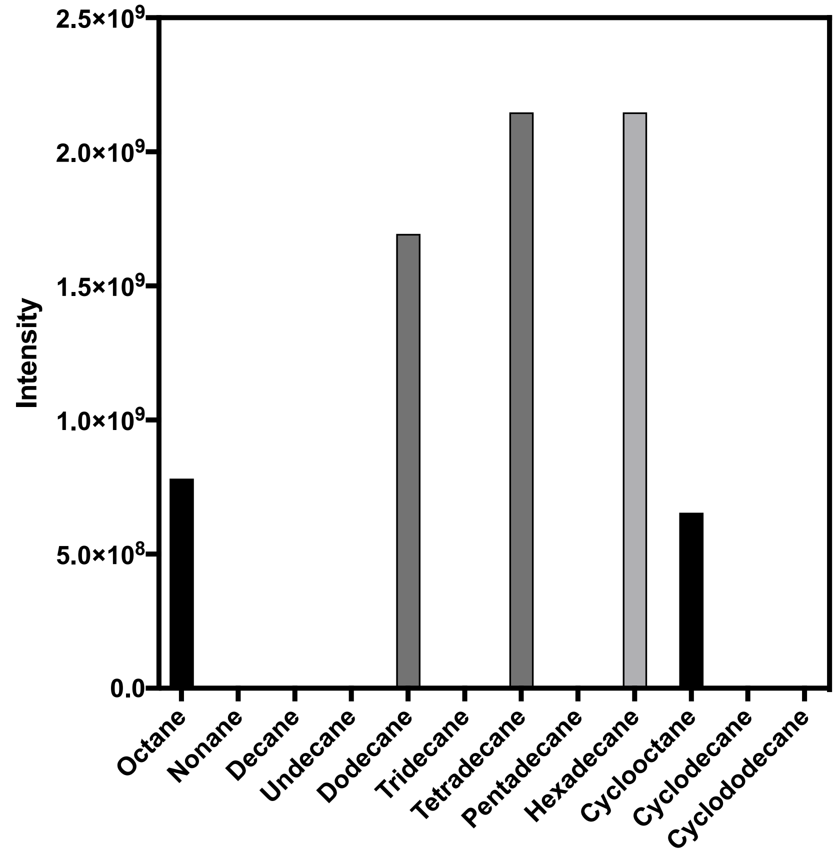


Figure 6.20. Peak area of identified post-purged Twister stirbar*.* Twister stir bar was purged and was analysed for potential carry-over effects.

6.3 Discussions

This study represents the investigation of six unique isolates of *A. sarcoides* for the production of fuel molecules. Previous metabolomic studies have only examined the volatilome of *A. sarcoides*. In contrast, this study investigates the aqueous phase of the secretome of *A. sarcoides*. Here we examined the aqueous secretome of *A. sarcoides* by employing solvent extraction and SBSE methods.

With solvent extraction, we were able to detect spiked alkanes to the concentration of 0.001 mg. No alkane peaks were observed from culture extracts of *A. sarcoides* at the expected RT range in Figure 2. Comparing extracts of *A. sarcoides* 64019 to extracts of sterile media samples did not yield much difference in terms of GC chromatograms. As elucidation of alkanes with chloroform extraction was not able to establish *A. sarcoides* as an alkane producer above the concentration of 0.001 mg ml^{-1}, an alternative method was developed. Stir bar sorptive extraction (SBSE) method was used to directly extract metabolites from *A. sarcoides* culture medium. In conjunction with GC-MS, a method was developed to investigate the presence of linear and cyclic alkanes in cultures of *A. sarcoides*. We designed a method involving SBSE that aims to minimise alkane contamination from exogenous sources and to minimise the escape of biogenic metabolites from the experimental system. To complement this, controls such as sterile media, spiked sterile media, and alkane non-producers were included to evaluate the SBSE method and *A. sarcoides*' ability to biosynthesise linear and cyclic alkanes. The findings indicated that it was possible to resolve peaks for linear and cyclic alkanes at the levels of 100 ppb, which is approximately 10 times more sensitive than previously tested solvent extraction method. Furthermore, linear and cyclic alkanes were detected in samples of sterile medium spiked with alkane standards.

When the compound profile is compared to samples containing sterile medium, there were different alkane compounds present and at a higher level in the profile of the spiked samples (Table 6.2). This suggests that the SBSE method is able to discriminate levels of spiked standard against the background of the medium. Moreover, RT performance of linear and cyclic alkanes in a GC-MS are dependent on the alkanes’ length because the length of the alkane affects its physicochemical properties (Morrison et al., 1988). Thus, it is possible to predict the RT of alkanes by plotting a calibration graph of RT to the function of alkane length. These calibration curve produced is proportional and with a significant line of best fit for the alkane standards. This is useful for predicting RT of other alkanes within the C8 to C16 range and has been used to identify alkanes such as undecane and pentadecane. Further efforts were made to increase confidence in alkane peak identification; the retention shift in each batch was quantified by identifying peaks associated with all SBSE chromatograms.

The production of biogenic organic compounds were further supported by the presence of sesquiterpene, which were detected in all isolates of *A. sarcoides* (Table 6.4). As the mass-ion spectra was complicated to solve manually, elucidation was achieved by relying on spectral library comparisons. The presence of sesquiterpene suggests that the compound is biological origin, as it is a molecule that is commonly associated as a biological secondary metabolite (Schmidt-Dannert, 2015) and further suggests the method is able to detect biogenic compounds. Caution must be applied in interpreting the exact identity of the sesquiterpene compounds as many sesquiterpenes have highly similar chemical structure. Like alkanes, terpenes were detectable in sterile media samples. Observation of sesquiterpene was comparable to previous studies that investigated the metabolome of *A. sarcoides* with methods like HS-SPME or (Proton-Transfer Reaction Mass Spectroscopy) PTR-MS sampling (Griffin et al., 2010, Strobel et al., 2010b, Mallette et al., 2012, Mallette et al., 2014). Like this study, the authors were unable to identify the sesquiterpenes with high confidence.

Within all *A. sarcoides* samples, linear alkanes between the length of C8 to C16 and cyclic alkanes with the length of C8 to C12 were detected. This was achieved by confirming peaks by RT and identification was confirmed by comparing the mass-ion spectra of peaks to alkane standards. The use of a calibration curve was able to identify uncharacterised linear alkanes undecane and pentadecane from the chromatograms of experimental samples. Interestingly, tridecane was not identified but was reported previously (Strobel et al., 2010b, Ahamed et al., 2011). As expected, the chromatograms of *A. nidulans* and *S. cerevisiae* samples indicate low or no levels of linear and cyclic alkanes. This suggests that linear and cyclic alkane biosynthesis is not widespread in ascomycetes and filamentous fungi. However, conclusive proof of alkane biosynthesis was unattainable, as sterile media samples contained levels of alkane at a comparable magnitude to biological *A. sarcoides* samples (Figure 6.10 to 6.15 and 18). This was especially true when comparing within experimental batches. The presence of comparable quantities of alkanes in sterile media samples undermines the confidence of conclusions derived from SBSE. Therefore, with the presence of contaminant alkanes, it is not possible to confirm the hypothesis of linear and cyclic alkane biosynthesis by *A. sarcoides* using the SBSE method.

To recognise the source of the contamination we must interrogate the method itself. At such high quantities, it is unlikely to be contaminated by gaseous transmission due to the closed nature of the experimental system. Furthermore, at the incubation temperature of 23°C, it is highly improbable that alkanes such as hexadecane (boiling point: 287 °C; melting point: 18 °C (NIST)) would exist in a gaseous state for transmission to occur. As the media was composed of inorganic salts and high-purity glucose, it is also unlikely that the growth medium was the source of contamination. The fermentation tube is made up of borosilicate glass. Unlike plastic, it is a material not known for alkane sorption. Additionally, to ensure removal of organic compounds, including alkanes, an alcohol-hydroxide wash (Dong et al., 2019) and heat treatment by baking and autoclaving was used to vaporise leftover organic compounds. The PTFE-lined caps were washed with 98% ethanol to ensure the removal of hydrophobic compounds. Based on these precautions, it is unlikely that alkane remains within the fermentation vessel. Moreover, the preparation method for GC injection was automated and would be unlikely to introduce contamination during injection processes. Therefore, we hypothesised that the presence of exogenous alkane in sterile media samples may be a result of carry-over accumulated by the Twister stir bar.

For the PDMS-based analysis, heat was applied to desorbed hydrophobic compounds from the matrix of the SBS stir bar at the end of each analytical run. There are two functions for applying heat to the stir bar. Firstly, to facilitate injection of sorbed compounds into the GC. Secondly, thermal desorption is a common method for purging compounds from the PDMS matrix. This includes headspace analysis with SPME fibres used in previous metabolic examination of *A. sarcoides* (Table 6.1). It is clear that the thermal desorption was able to desorb and inject extracted compounds to produce chromatograms. As carry-over is an issue, we deduced that the thermal desorption regime was not rigorous enough for the removal of compounds from the Twister bar. Indeed, when a recently purged stir bar was sampled, alkane peaks were identified and detected at levels present in *A. sarcoides* samples (Figure 6.20). This result suggests that the source of exogenous alkane was an effect of stir bar carry-over.

The implication of carry-over reduces our confidence in the presence of alkane within *A. sarcoides* samples. The primary source of exogenous alkane would be from the introduction of alkane standards and therefore, must be discounted from *A. sarcoides* samples. However, the detection of undecane and pentadecane in experimental samples is interesting. If alkane contamination is because of carry-over from spiked alkane standards, then the presence of undecane and pentadecane suggests a biogenic origin. Indeed, undecane and pentadecane were detected in sterile media samples too, presumably from carry-over from the stir bar. Similar conclusions were drawn for sesquiterpenes, which was found in comparable quantities in both biological samples and sterile media samples due to carry-over but may be of biological origin. If undecane and pentadecane was of a biogenic origin, it would be in agreement with $C_{(n-1)}$ linear alkane biosynthesis characterised in other biological systems (Schirmer et al., 2010, Bernard et al., 2012a, Qiu et al., 2012, Rui et al., 2014). Presence of pentadecane suggests the C-C decarboxylation of the ubiquitous palmitic acid. Previous examination of *A. sarcoides* liposomes indicated that C16 and C18 fatty acids were the most abundant fatty acids present in *A. sarcoides* (Müller et al., 1994) and would support this hypothesis. For undecane, however, quantities of C12 fatty acids are minute (0.3 % - 0.4 %) in abundance and may confound this hypothesis (Müller et al., 1994). Another contradiction for this hypothesis was the absence of tridecane. If a fatty acid enzyme is able to convert the C12 and C16, presumably to a corresponding aldehyde or aldehyde decarbonylation, then it would be unlikely to bypass C14 fatty acids or aldehyde.

Conversely, if *A. sarcoides* is an alkane producing organism, we would expect production titre higher than those from carry over. Assuming that carry-over occurs then desorption would lead to a reduction in sorbed alkanes in the stir bar. Therefore, if *A. sarcoides* should produce alkanes it would lead to higher quantities of sorbed alkanes than that of sterile media samples. Indeed, several metabolites reported from previous reports were not consistent and were not reproducible in later studies (Griffin et al., 2010). This includes hydrocarbons, including several linear, branched and cyclic hydrocarbons. A 2001 study investigated the volatilome of NRRL 50072 by HS-SPME and identified a diverse set of hydrocarbons confirmed by chemical standards (Stinson et al., 2003). Hydrocarbons included linear alkanes and branched alkenes were detected. Of interest is the presence of [8]annulene, an unsaturated eight carbon cyclic aromatic ring in the metabolome which exhibited anti-fungal properties (Stinson et al., 2003). Since then, its presence was not observed in studies attempting to replicate it (Griffin et al., 2010). The first report of alkane production in NRRL 50072 was in a 2008 report (Strobel et al., 2010b). HS-SPME examination of the culture in different media was able to produce a range of hydrocarbons including linear, branched, and cyclic alkanes, of which some were identified by inclusion of alkane standards and were subtracted from sterile media controls.

In 2010, HS-SPME sampling detected linear and cyclic alkanes in twelve species from the *Ascocoryne* genus (Griffin et al., 2010). The authors conclude that the genus is capable of producing short- and medium-chain alkenes, ketones, esters and alcohols and several sesquiterpenes. Hydrocarbon and oxygenated hydrocarbon metabolites were not consistent across three independent experiments, suggesting metabolite production to be highly variable within the biological system. The experimental design did not include chemical standards for linear and cyclic alkanes. Instead, identification of compounds were based on comparative spectral analyses against a standard spectral library. The authors and other reports also suggest that identification of alkanes by comparisons with standard spectral libraries are highly similar for many different forms of alkanes (Schulz et al., 2007, Griffin et al., 2010). Even in our study, we recognised discrepancies in automated peak annotation and predicted spectral analysis to manual annotation and spectral analysis, although we are able to resolve identification by utilising RT data provided by chemical standards. Thus, the accurate identification of alkanes in studies involving *A. sarcoides* without alkane standards must be called into question. Furthermore, in the 2010 study, branched alkanes were found in sterile media samples, suggesting that alkanes were present in their experimental system. These results indicated that the previously thought NRRL 50072-derived branched alkanes (Strobel et al., 2010b) may be of exogenous origin instead (Griffin et al., 2010). Indeed the authors from the 2008 study, which reported NRRL 50072 branched alkane production, submitted corrections based on the 2010 report. As there were inaccuracies in identification mentioned previously, these branched alkanes must also be questioned as they may be misidentified linear or cyclic alkanes.

Later volatilome examination of NRRL 50072 compared metabolites captured by HS-SPME against RT-PTR-MS (Mallette et al., 2012, Mallette et al., 2014). By using HS-SPME, only two alkanes, pentane and 4-methylheptane, were identified from the cultures of NRRL 50072. With PTR-MS, no alkanes were detected in NRRL 50072 cultures. The authors stated that the lack of alkanes was due to the ionisation technique, as alkanes are thought to be recalcitrant to hydronium ionisation. Importanntly, this study also lack alkane standards for the identification of alkanes as alkanes were not the compound of interest. Many authors from these studies attribute inconsistent metabolome to biological variability or due to serial microbial passage (Griffin et al., 2010). Findings here may provide explanations to the inconsistencies reported for *A. sarcoides* metabolic profiles and suggests carry-over may be the cause of artefacts and inconsistencies, especially with PDMS-based methodology.

If we are to assume that all reports of *A. sarcoides* contain no contamination, it will also be the first instance of an alkane producing organism capable of producing alkanes with a diverse chain length from C5 to C19. This would also include a range of even-chain alkanes. Collectively, this would represent novel biosynthesis and would be a deviation from the well characterised head-to-head and decarboxylation alkane biosynthesis. If cyclic alkanes were of biogenic origin, its presence would also be unique to biology. There are currently no characterised biological pathway sthat can accommodate cyclic alkane biosynthesis. Cyclohexanoic acid is a comparable biological metabolite to cycloalkanes and it represents a single reduction step to produce cyclohexane. It was observed that *Alicyclobacillus acidocaldarius* is able to produce ω-cyclohexyl fatty acid as the most abundant lipid and cyclohexanoic acid was proposed to be required as a substrate for the formation of cyclohexyl motif (Moore et al., 1993). Notably, the liposome of *A. acidocaldarius* revealed the presence of other cyclic fatty acids with a cyclic motif of (C5-7) (Moore et al., 1993). With a non-producing mutant, feeding isotopic labelled substrate revealed intermediates of cyclohexanoate and elucidated sequential reduction pathway of all hydroxyl groups present in the shikimic acid substrate (Moore et al., 1993). Reduction of the hydroxyl group leads to the formation of an unsaturated bond at the carbon bearing hydroxyl group and the elimination of -OH (Moore et al., 1993). Further reduction by a hydrogen donor saturates the unsaturated bond to a saturated bond (Moore et al., 1993). As three hydroxyl group sare present on shikimic acid, this process occurs three times, leading to the formation of cyclohexanoic acid (Moore et al., 1993). A biosynthetic gene cluster of doramectin was found to encode cyclohexanoic acid for the production of ω-cyclo fatty acids, an antifungal antibiotic which contains a cyclohexyl motif (Cropp et al., 2000). Another interesting class of cyclic metabolites are muscone, a C15 cycloketone from the Musk deer and civetone, a C17 cycloketone from African civets. Although synthetic pathways for both compounds have been proven (Ruzicka, 1926, Choo et al., 1994), the biosynthesis remains unknown and both organisms did not demonstrate the production of cycloketones of other length. Another pathway which can form cyclic hydrocarbons is facilitated by terpene cyclases, a class of enzyme able to catalyse ring closure in terpenes. The ring closure of a terpene would not form an unbranched and aliphatic cycloalkane. An example of this would be germacrene C, a terpene equivalent of a branched cyclodecane. Its biosynthesis is catalysed by germacrene C synthase which converts farnesyl diphosphate to germacrene C (Colby et al., 1998). Much like cyclohexanoic acid, this class of compounds is a few reduction steps to produced cyclic alkanes. If comparable pathway was present in *A. sarcoides* and was responsible for cycloalkane biosynthesis, then it must have a mechanism for ring expansion. This seems unlikely as no ring expansion mechanisms have been characterised in biology.

6.4 Conclusions

Assays with solvent extraction methods and GC-MS were able to detect exogenous alkanes. However, it was concluded that biogenic alkanes were not present at great enough concentrations to facilitate detection by this method. While assaying with SBSE and GC-MS, alkanes were detected in the cultures of *A. sarcoides*, we are unable to confirm the origins of the alkane due to the presence of contaminants. Evidence suggests that the source of the alkanes was due to carry-over from insufficient purging of the Twister stir bars. Therefore, it is not obvious if alkanes are of biogenic origin or from an exogenous source, although undecane and pentadecane were detected and were not used as a spiked standard. On the balance of evidence, it suggests that undecane and pentadecane were derived from *A. sarcoides*. Nonetheless, the presence of alkane within sterile media undermines the confidence of biogenic alkanes and is unable to conclusively establish alkane biosynthesis in *A. sarcoides*.

All studies investigating the metabolome of *A. sarcoides* lack this rigorous approach due to the absence of controls detailing the potential of carry-over or the absence of authentic standards. When they are included, as in this study, the evidence of alkane biosynthesis is mixed. To establish *A. sarcoides* as an alkane producer, this, and previous metabolic examination requires a higher level of proof than is present.

CHAPTER 7 DISCUSSION AND FURTHER RECOMMENDATION

7.1 Summary

This thesis summarises the work to evaluate six isolates of the filamentous fungi *A. sarcoides* with the aim of establishing the species as an alkane producer and elucidating the genetic component that underpins their biosynthesis.

In Chapter 3, we established the basic microbiology of *A. sarcoides*. In this chapter the development of protocols were important for later alkane screening efforts, such as utilising a chemically defined medium to reduce exogenous organic compounds that may interfere with alkane detection, the development of long-term storage protocol to reduce phenotypic drift, and a growth assay method using microtitre plates. The chapter also detailed alkane degradation in *A. sarcoides*, which was not reported in previous studies. Interestingly, only linear alkanes were able to support growth, while results suggest that cyclic alkanes may be toxic to the fungi.

In Chapter 4, six novel genomes were assembled successfully from Illumina-seq libraries. The work here succeeded in generating high quality bioinformatics resources, such as the genome and a predicted proteome, for interrogation of potential pathways in later studies. While we are able to assemble six high quality genomes, we were unable to attain 100% coverage for all genomes. To achieve complete coverage with high accuracy, a combination of different sequencing is required. By using high-frequency short reads library derived from Illumina-seq and the long reads library from Pacbio or Nanopore sequencing methods, it is possible to map the longer reads as a guide for the shorter and more accurate reads. In achieving 100% coverage for each genomes, it would be possible to elucidate the complete chromosomes of each isolate. This is important, as this would allow spatial interrogation of genes on a chromosomal level and enable synteny analysis of the genomes. By using a publicly available transcriptome of NRRL 50072 (Gianoulis et al., 2012), it was possible to elucidate the genes within other isolates by HMM machine learning. The accuracy and fidelity of gene elucidation processes can be increased further by developing the transcriptome for each isolate. This work will depend on Chapter 3, which investigated different growth conditions, and be used to elicit different transcriptomic responses. Each dataset was able to achieve over 90% BUSCO annotation, a benchmark that suggests that the draft genome and predicted genes are of high quality. We are also able to annotate the predicted genes to pathways on KEGG, *Aspergillus nidulans* PANTHERS ontologies, and proteome of *Sacchromyces cerevisiae* S288c and *A. nidulans* FGSCa4.

In Chapter 5, we were able to describe the biological relevance of the six genomes of *A. sarcoides*. Four different annotation evidences were integrated to inform putative annotations of all genes predicted by HMM machine learning algorithms. Importantly, by forming non-redundant orthologous clusters, we were able to conduct comparative analysis at the species level and against *S. cerevisiae* S288c and *A. nidulans* FGSCa4. This approach allowed us to propose a rational pathway to describe the biochemical potential of all six isolates. We proposed six pathways relating to fatty acid synthase I & II, fatty acid elongation, isoprene biosynthesis, prenyl elongation, lipid oxygenation and a hypothetical linear alkane pathway. Importantly, we were able to propose an alkane degradation pathway, which supports the alkane degradation observation noted in Chapter 3. In order to increase the sensitivity in detecting unique genes within *A. sarcoides*, a genome from an organism with a closer phylogenetic relation is required. This can be facilitated by the inclusion of *Botrytis cinerea*, for which the genome was fully published. The lipid hyperoxidase pathway proposed will be of interest to synthetic biologists aiming to expand the product profile of advanced biofuels. The product profile of alkane in synthetic pathways is dependent on controlling the characteristics of the substrate, such as branched motifs and length (Howard et al., 2013, Sheppard et al., 2016). With the lipid hyperoxidase pathways, it is a pathway that is capable of generating C8 aldehydes from linoleic acid (Brodhun et al., 2010). This opens up potential bioengineering opportunities to use this pathway to generate varying lengths of aldehydes and provides further alternatives to pathway designed to synthesise short and even chain alkanes.

In Chapter 6, alkane production was investigated with two different extraction methods. No alkanes were detected from the cultures of *A. sarcoides* using a solvent extraction method at the levels of 0.001 mg ml^{-1}. An alternative method with SBSE was developed to detect alkanes from the cultures of *A. sarcoides*. This would be the first report of utilising SBSE for extracting metabolites directly from biological cultures. The SBSE method was able to detect alkanes from *A. sarcoides* cultures. However, alkanes were also detected in sterile media control. These were thought to be a result of carry-over due to the lack of thermal purging efficacy. This decreases the confidence in claiming that alkanes detected from biological samples were of biogenic origin. To improve on the SBSE method developed in chapter 6, the focus will be to reduce the effects of carry-over. To remedy this, the Twister stir bar must be further investigated for purging time and cleaning by solvent. Furthermore, it would be of interest to establish the partition co-efficient for different length and types of alkanes within the growth conditions of *A. sarcoides*. This would allow the quantification of alkanes that are sorbed on to the Twister stir bar.

It must be highlighted that without certain controls, alkane carry-over would not have been detected. This raises some concerns about the reliability of other studies that primarily relied on HS-SPME, also PDMS-based extraction method (Table 6.1). In these studies, an alkane was considered to be of a biogenic origin by subtracting compounds for controls. However, often subtracted metabolic and the carry-over of HS-SPME were under reported or not considered. Indeed, branched alkanes were once considered to be part of the alkane repertoire that was biosynthesised by *A. sarcoides* (Strobel et al., 2010b), however, this was disproven by later HS-SPME study and was concluded to be exogenous alkane contaminants (Griffin et al., 2010), which resulted in a retraction. This highlights the need to take caution with methods that are reliant on sorb sampling. It also highlights the need for further confirmation by other sampling methods which is needed to establish *A. sarcoides* as an alkane producer. To this date, only one other method, which used PTR-MS, has been used for the sampling of volatile metabolite of *A. sarcoides*. Within two PTR-MS studies of *A. sarcoides*, only one study reported the identification of one alkane compound, in the form of pentane (Mallette et al., 2012, Mallette et al., 2014). If *A. sarcoides* was a prolific alkane producer, we would expect a wider variety of aliphatic hydrocarbon compounds than pentane. However, the authors argued that the lack of alkanes was because of low proton affinity associated with alkanes, which would impede alkane ionisation and, by extension, detection by PTR-MS. This is not in agreement with previous PTR-MS methods which detailed the identification of volatile alkanes, within the range of C8-C16, from combustion exhaust by a similar ionisation method (Jobson et al., 2005). Ultimately, *A. sarcoides* should be subjected to a higher level of proof to establish the organism as an alkane producer.

7.2 Further recommendations

We report that *A. sarcoides* is an alkane degrader, a novel finding not reported in previous studies of the fungi. Linear alkane assimilation was noted in Chapter 3 and the genetic basis for alkane degradation was elucidated in Chapter 5, which suggests a homology to a family of alkane hydroxylating, P450 ALK enzymes. This is a well characterised pathway in another fungal species, *Yarrowia lipolytica*, in which alkane is hydroxylated by a P450 ALK and is assimilated by the beta-oxidation pathway facilitated by downstream enzymes (Fukuda et al., 2013). The putative genes should be heterologously expressed to confirm alkane hydroxylating activity. If *A. sarcoides* is an alkane producer, it is unclear how the fungus is able to consolidate both pathways. The regulation of alkane metabolism would be of interest for synthetic biologists to design transcriptional controls for designing alkane biosensors. If we are to assume that expression of alkane biosynthesis represses alkane degradation and vice versa, by interrogating the transcriptome, we may elucidate candidates that are linked to the regulation of alkane metabolism and, by extension, an alkane biosynthetic gene.

To seek further proof for *A. sarcoides* as an alkane producer must depart from *A. sarcoides*. As mentioned in Section 1.3, secondary metabolite production can be varied and be unpredictable. This reason was suggested to be the case in studies that examine the highly varied alkane production in *A. sarcoides* (Table 6.1). Thus, it is important to explore methods to heterologously express candidate genes in production hosts. Although there are inherent limitations to this, as the size of genes may not be practical to facilitate transformation. Heterologous expression methods were explored in studies that aim to elucidate the biosynthetic genes of antibiotic metabolites by heterologous expression (Greunke et al., 2018, Kawahara et al., 2018, Duell et al., 2019). Candidates identified in Chapter 5 support this kind of exploratory pipeline. Another option is to use more speculative and exploratory methods like transformation-associated recombination (TAR) cloning (Kouprina et al., 2008) in which sections of the genomes are spliced and expressed in yeast. Draft genomes detailed in Chapter 4 enabled the exploration of TAR cloning as a viable method for prospecting alkane producing genes. This is limited by the assumption that alkane biosynthetic genes are arranged in biosynthetic gene clusters (BGCs) in fungi and that the spliced fragments contained the necessary genes to evoke alkane biosynthesis. Either method can then be assayed for alkane production by either assaying the cultures of the transforming hosts directly or by assaying the lysates of the transformants. If alkanes were detected within the cultures of transformant host, the genetic element responsible will be identified and be attributed to heterologously expressed gene or genes. This non-targeted TAR method was used for the identification of the hydrocarbon biosynthesis gene (Rui et al., 2014, Rui et al., 2015).

For heterologous expression approaches to succeed, it is necessary to have an efficient method to assay for alkanes. Unlike bioactive compounds, such as antibiotics, the presence of alkane must be directly analysed and cannot be inferred. This often requires solvent extraction methods, which are time-consuming and cannot handle a large number of samples. In Chapter 6, attempts to overcome these limitations were explored with the development of the SBSE methodology. While the method has its disadvantages, which have been discussed, it also has many traits that suit non-targeted approaches. It is a method that is high-throughput, is capable to sorbed hydrocarbon compounds, and concentrating the extracts, which makes it suitable for handling high sample numbers and detecting a wide range of compounds generated by non-targeted approaches. Deviating from the usage of GC-MS for assaying alkanes, a biosensor that is sensitive and specific to alkane detection will suit a non-targeted screening.

Biosensors may enable rapid detection of alkanes within microbial cultures and by coupling to a fluorescent signal, allows the usage of conventional methods, such as plate readers, for real-time alkane quantification. This has the inherent advantage of throughput when compared to GC-MS related methods for identifying alkanes. Recent development of alkane biosensors indicate the possibility of detecting intracellular alkanes by co-expressing the alkane biosensors with an alkane pathway (Lehtinen et al., 2017).

^{13}C radio-labelled substrate feeding may also be a viable route for investigating alkane biosynthesis. From Chapter 3 we were able to indicate the substrate preferences of *A. sarcoides.* This includes preference for acetic acid, a compound that is commonly used for ^{13}C radio-labelling. In this recommendation, the cell extracts of *A. sarcoides* will be incubated with ^{13}C-acetic acid and assay for ^{13}C-alkane. This has the advantage of identifying potential intermediates in alkane formation and confirming the production of alkane within *A. sarcoides*. A similar method was used to elucidate the *Chlorella* fatty acid photodecarboxylase pathway, in which the lysate was further purified and incubated with ^{13}C-palmitic acid to elucidate the protein responsible for alkane production (Sorigué et al., 2017).

7.3 Conclusion

In conclusion, this thesis has shown that it is possible to elucidate genes relating to phenotypic observations, we were able to elucidate genes involved in alkane degradation by combining observation of alkane assimilation with bioinformatic analysis. This resulted in the annotation of alkane degradation pathway. With the same methodology, we also proposed a rational alkane pathway for both odd and even chain alkanes. This alkane biosynthesis route is based on *fdc*1 decarboxylation/decarbonylation. No alkane was detected at the range of C8 to C14 with conventional solvent extraction methods to the detection limits of 0.001 mg ml-1. However, it also highlights the lack of reliable data with PDMS-related methodology in establishing *A. sarcoides* as an alkane producer. In summary, the case for alkane degradation in *A. sarcoides* is greater than the case for alkane biosynthesis.

Bibliography

1994. United Nations Framework Convention on Climate Change : resolution / adopted by the General Assembly, UN General Assembly.

2009. Directive 2009/28/EC of the European Parliament and of the Council of 23 April 2009 on the promotion of the use of energy from renewable sources and amending and subsequently repealing Directives 2001/77/EC and 2003/30/EC (Text with EEA relevance), BG, ES, CS, DA, DE, ET, EL, EN, FR, IT, LV, LT, HU, MT, NL, PL, PT, RO, SK, SL, FI, SV.

2016. Paris Agreement. COP Report No. 21, Addenum, at 21, U.N. Doc. FCCC/CP/2015/10/Add, 1. UNFCC: United Nations.

2018. Directive (EU) 2018/2001 of the European Parliament and of the Council of 11 December 2018 on the promotion of the use of energy from renewable sources. *PE/48/2018/REV/1,* OJ L 328**,** 82–209.

AGGER, S., LOPEZ-GALLEGO, F. & SCHMIDT-DANNERT, C. 2009. Diversity of sesquiterpene synthases in the basidiomycete Coprinus cinereus. *Molecular Microbiology,* 72**,** 1181-1195.

AHAMED, A. & AHRING, B. K. 2011. Production of hydrocarbon compounds by endophytic fungi Gliocladium species grown on cellulose. *Bioresource Technology,* 102**,** 9718-9722.

AIRD, D., ROSS, M. G., CHEN, W.-S., DANIELSSON, M., FENNELL, T., RUSS, C., JAFFE, D. B., NUSBAUM, C. & GNIRKE, A. 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology,* 12**,** R18.

ALBRO, P. W. & DITTMER, J. C. 1969. Biochemistry of long-chain, nonisoprenoid hydrocarbons. I. Characterization of the hydrocarbons of *Sarcina lutea* and the isolation of possible intermediates of biosynthesis. *Biochemistry,* 8**,** 394-405.

ALDRICH, J. R., KHRIMIAN, A. & CAMP, M. J. 2007. Methyl 2, 4, 6-decatrienoates attract stink bugs and tachinid parasitoids. *Journal of Chemical Ecology,* 33**,** 801.

ANDREWS, S. 2010. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.

ANFAVEA 2012. Carta da ANFAVEA. *In:* (BRASIL), A. N. D. F. D. V. A. (ed.). Brazil: ANFAVEA.

ARANDA, A. & DEL OLMO, M. L. 2003. Response to acetaldehyde stress in the yeast *Saccharomyces cerevisiae* involves a strain-dependent regulation of several

ALD genes and is mediatedby the general stress response pathway. *Yeast,* 20**,** 747-759.

BALTUSSEN, E., SANDRA, P., DAVID, F. & CRAMERS, C. 1999. Stir bar sorptive extraction (SBSE), a novel extraction technique for aqueous samples: theory and principles. *Journal of Microcolumn Separations,* 11**,** 737-747.

BANKEVICH, A., NURK, S., ANTIPOV, D., GUREVICH, A. A., DVORKIN, M., KULIKOV, A. S., LESIN, V. M., NIKOLENKO, S. I., PHAM, S. & PRJIBELSKI, A. D. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology,* 19**,** 455-477.

BAO, L., LI, J.-J., JIA, C., LI, M. & LU, X. 2016. Structure-oriented substrate specificity engineering of aldehyde-deformylating oxygenase towards aldehydes carbon chain length. *Biotechnology for Biofuels,* 9**,** 185.

BECKER, E. W. 1994. Microalgae: biotechnology and microbiology, Vancouver, Cambridge University Press.

BELLER, H. R., GOH, E.-B. & KEASLING, J. D. 2010. Genes involved in long-chain alkene biosynthesis in *Micrococcus luteus*. *Appl. Environ. Microbiol.,* 76**,** 1212-1223.

BENNETT, J. W. & CIEGLER, A. 1983. Secondary metabolism and differentiation in fungi.

BERLA, B. M., SAHA, R., MARANAS, C. D. & PAKRASI, H. B. 2015. Cyanobacterial Alkanes Modulate Photosynthetic Cyclic Electron Flow to Assist Growth under Cold Stress. *Scientific Reports,* 5**,** 14894-14894.

BERNARD, A., DOMERGUE, F., PASCAL, S., JETTER, R., RENNE, C., FAURE, J.-D., HASLAM, R. P., NAPIER, J. A., LESSIRE, R. & JOUBÈS, J. 2012a. Reconstitution of Plant Alkane Biosynthesis in Yeast Demonstrates That *Arabidopsis* ECERIFERUM1 and ECERIFERUM3 Are Core Components of a Very-Long-Chain Alkane Synthesis Complex. *The Plant Cell,* 24**,** 3106-3118.

BERNARD, A., DOMERGUE, F., PASCAL, S., JETTER, R., RENNE, C., FAURE, J.-D., HASLAM, R. P., NAPIER, J. A., LESSIRE, R. & JOUBÈS, J. 2012b. Reconstitution of plant alkane biosynthesis in yeast demonstrates that *Arabidopsis* ECERIFERUM1 and ECERIFERUM3 are core components of a very-long-chain alkane synthesis complex. *The Plant Cell***,** tpc. 112.099796.

BESSOULE, J.-J., LESSIRE, R., RIGOULET, M., GUERIN, B. & CASSAGNE, C. 1987. Fatty acid synthesis in mitochondria from *Saccharomyces cerevisiae*. *FEBS letters,* 214**,** 158-162.

BIELESKI, R. & FERGUSON, I. 1983. Physiology and metabolism of phosphate and its compounds. *Inorganic Plant Nutrition.* Springer.

BLOCH, K. & VANCE, D. 1977. Control mechanisms in the synthesis of saturated fatty acids. *Annual Review of Biochemistry,* 46**,** 263-298.

BOKULICH, N. A., SUBRAMANIAN, S., FAITH, J. J., GEVERS, D., GORDON, J. I., KNIGHT, R., MILLS, D. A. & CAPORASO, J. G. 2013. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nature Methods,* 10**,** 57.

BOLGER, A. M., LOHSE, M. & USADEL, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics,* 30**,** 2114-2120.

BONNETT, S. A., PAPIREDDY, K., HIGGINS, S., DEL CARDAYRE, S. & REYNOLDS, K. A. 2011. Functional characterization of an NADPH dependent 2-alkyl-3-ketoalkanoic acid reductase involved in olefin biosynthesis in *Stenotrophomonas maltophilia*. *Biochemistry,* 50**,** 9633-9640.

BOONE, S. C. & WAKIL, S. J. 1970. In vitro synthesis of lignoceric and nervonic acids in mammalian liver and brain. *Biochemistry,* 9**,** 1470-1479.

BORATYN, G. M., SCHÄFFER, A. A., AGARWALA, R., ALTSCHUL, S. F., LIPMAN, D. J. & MADDEN, T. L. 2012. Domain enhanced lookup time accelerated BLAST. *Biology Direct,* 7**,** 12.

BOURDENX, B., BERNARD, A., DOMERGUE, F., PASCAL, S., LÉGER, A., ROBY, D., PERVENT, M., VILE, D., HASLAM, R. P., NAPIER, J. A., LESSIRE, R. & JOUBÈS, J. 2011. Overexpression of *Arabidopsis* ECERIFERUM1 Promotes Wax Very-Long-Chain Alkane Biosynthesis and Influences Plant Response to Biotic and Abiotic Stresses. *Plant Physiology,* 156**,** 29.

BRASIL, A. 2008. "ANP: consumo de álcool combustível é 50% maior em 2007" (in Portuguese) [Online]. Invertia. [Accessed 2008].

BREDEHOEFT, M. & FARBER-DEANDA, M. 2014. *E85 fueling station availability is increasing* [Online]. https://www.eia.gov/todayinenergy/detail.php?id=15311: U.S. Energy Information Administration. [Accessed 24/08/2018 2018].

BRODHUN, F., GÖBEL, C., HORNUNG, E. & FEUSSNER, I. 2009. Identification of PpoA from *Aspergillus nidulans* as a fusion protein of a fatty acid heme dioxygenase/peroxidase and a cytochrome P450. *Journal of Biological Chemistry,* 284**,** 11792-11805.

BRODHUN, F., SCHNEIDER, S., GÖBEL, C., HORNUNG, E. & FEUSSNER, I. 2010. PpoC from *Aspergillus nidulans* is a fusion protein with only one active haem. *Biochemical Journal,* 425**,** 553-565.

BUER, B. C., PAUL, B., DAS, D., STUCKEY, J. A. & MARSH, E. N. G. 2014. Insights into substrate and metal binding from the crystal structure of cyanobacterial aldehyde deformylating oxygenase with substrate bound. *ACS Chemical Biology,* 9**,** 2584-2593.

CAMACHO, C., COULOURIS, G., AVAGYAN, V., MA, N., PAPADOPOULOS, J., BEALER, K. & MADDEN, T. L. 2009. BLAST+: architecture and applications. *BMC Bioinformatics,* 10**,** 421.

CAMPBELL, C. D. & VEDERAS, J. C. 2010. Biosynthesis of lovastatin and related metabolites formed by fungal iterative PKS enzymes. *Biopolymers,* 93**,** 755-763.

CHAN, E.-C., KUO, J., LIN, H.-P. & MOU, D.-G. 1991. Stimulation of n-alkane conversion to dicarboxylic acid by organic-solvent-and detergent-treated microbes. *Applied microbiology and biotechnology,* 34**,** 772-777.

CHANG, P.-K. & EHRLICH, K. C. 2013. Genome-wide analysis of the Zn (II) 2 Cys 6 zinc cluster-encoding gene family in *Aspergillus flavus*. *Applied Microbiology and Biotechnology,* 97**,** 4289-4300.

CHANGE, C. O. C. 2018. Reducing UK emissions (2018) Progress Report to Parliament. *In:* CHANGE, C. O. C. (ed.). Committee on Climate Change.

CHEESBROUGH, T. M. & KOLATTUKUDY, P. E. 1984. Alkane biosynthesis by decarbonylation of aldehydes catalyzed by a particulate preparation from *Pisum sativum*. *Proceedings of the National Academy of Sciences of the United States of America,* 81**,** 6613-6617.

CHEN, X., GOODWIN, S. M., BOROFF, V. L., LIU, X. & JENKS, M. A. 2003. Cloning and Characterization of the *WAX2*; Gene of *Arabidopsis* Involved in Cuticle Membrane and Wax Production. *The Plant Cell,* 15**,** 1170.

CHOI, Y. J. & LEE, S. Y. 2013. Microbial production of short-chain alkanes. *Nature,* 502**,** 571.

CHOO, Y. M., OOI, K. E. & OOI, I. H. 1994. Synthesis of civetone from palm oil products. *Journal of the American Oil Chemists' Society,* 71**,** 911-913.

CHOY, J. S., WEI, S., LEE, J. Y., TAN, S., CHU, S. & LEE, T.-H. 2010. DNA methylation increases nucleosome compaction and rigidity. *Journal of the American Chemical Society,* 132**,** 1782-1783.

CHRISTENSON, J. K., RICHMAN, J. E., JENSEN, M. R., NEUFELD, J. Y., WILMOT, C. M. & WACKETT, L. P. 2017a. β-Lactone synthetase found in the olefin biosynthesis pathway. *Biochemistry,* 56**,** 348-351.

CHRISTENSON, J. K., ROBINSON, S. L., ENGEL, T. A., RICHMAN, J. E., KIM, A. N. & WACKETT, L. P. 2017b. OleB from bacterial hydrocarbon biosynthesis is a β-lactone decarboxylase that shares key features with haloalkane dehalogenases. *Biochemistry,* 56**,** 5278-5287.

COLBY, S. M., CROCK, J., DOWDLE-RIZZO, B., LEMAUX, P. G. & CROTEAU, R. 1998. Germacrene C synthase from *Lycopersicon esculentum* VFNT Cherry tomato: cDNA isolation, characterization, and bacterial expression of the multiple product sesquiterpene cyclase. *Proceedings of the National Academy of Sciences,* 95**,** 2216-2221.

COMPEAU, P. E., PEVZNER, P. A. & TESLER, G. 2011. How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology,* 29**,** 987.

CRAFT, D. L., MADDURI, K. M., ESHOO, M. & WILSON, C. R. 2003. Identification and characterization of the CYP52 family of *Candida tropicalis* ATCC 20336, important for the conversion of fatty acids and alkanes to α, ω-dicarboxylic acids. *Appl. Environ. Microbiol.,* 69**,** 5983-5991.

CRÉPIN, L., LOMBARD, E. & GUILLOUET, S. E. 2016. Metabolic engineering of Cupriavidus necator for heterotrophic and autotrophic alka(e)ne production. *Metabolic Engineering,* 37**,** 92-101.

CROPP, T. A., WILSON, D. J. & REYNOLDS, K. A. 2000. Identification of a cyclohexylcarbonyl CoA biosynthetic gene cluster and application in the production of doramectin. *Nature Biotechnology,* 18**,** 980-983.

DONG, V. M. & KIM, D. 2019. Organic Chemistry II. Cleaning Glassware. *JoVE*.

DOU, C., MARCONDES, W. F., DJAJA, J. E., BURA, R. & GUSTAFSON, R. 2017. Can we use short rotation coppice poplar for sugar based biorefinery feedstock? Bioconversion of 2-year-old poplar grown as short rotation coppice. *Biotechnology for Biofuels,* 10**,** 144-144.

DUELL, E. R., D'AGOSTINO, P. M., SHAPIRO, N., WOYKE, T., FUCHS, T. M. & GULDER, T. A. 2019. Direct pathway cloning of the sodorifen biosynthetic gene cluster and recombinant generation of its product in E. coli. *Microbial Cell Factories,* 18**,** 32.

EDDY, S. R. 1998. Profile hidden Markov models. *Bioinformatics (Oxford, England),* 14**,** 755-763.

EDWARDS, R., PADELLA, M., GIUNTOLI, J., KOEBLE, R., CONNELL, A., BULGHERONI, C. & MARELLI, L. 2017. *Definition of input data to assess GHG default emissions from biofuels in EU legislation. Version 1c*.

EMISSIONS, U. G. G. 2017. 2017 UK Greenhouse Gas Emissions. *In:* DEPARTMENT FOR BUSINESS, E. I. S. (ed.).

FINN, R. D., CLEMENTS, J. & EDDY, S. R. 2011. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research,* 39**,** W29-W37.

FONG, Y., ANUAR, S., LIM, H., THAM, F. & SANDERSON, F. 2000. A modified filter paper technique for long-term preservation of some fungal cultures. *Mycologist,* 14**,** 127-130.

FRIAS, J. A., RICHMAN, J. E., ERICKSON, J. S. & WACKETT, L. P. 2011. Purification and characterization of OleA from *Xanthomonas campestris* and demonstration of a non-decarboxylative Claisen condensation reaction. *Journal of Biological Chemistry,* 286**,** 10930-10938.

FUKUDA, R. & OHTA, A. 2013. Utilization of hydrophobic substrate by *Yarrowia lipolytica.* Springer.

GHOBADIAN, B. & RAHIMI, H. 2004. Biofuels-past, present and future perspective. *In International Iran and Russian Congress of Agricultural and Natural Science.* Shahre-Kord University, Shahre Kord, Iran.: Vancouver.

GIANOULIS, T. A., GRIFFIN, M. A., SPAKOWICZ, D. J., DUNICAN, B. F., SBONER, A., SISMOUR, A. M., KODIRA, C., EGHOLM, M., CHURCH, G. M. & GERSTEIN, M. B. 2012. Genomic analysis of the hydrocarbon-producing, cellulolytic, endophytic fungus Ascocoryne sarcoides. *PLoS Genetics,* 8**,** e1002558.

GIBBS, P., SEVIOUR, R. & SCHMID, F. 2000. Growth of filamentous fungi in submerged culture: problems and possible solutions. *Critical Reviews in Biotechnology,* 20**,** 17-48.

GRANT, J. L., HSIEH, C. H. & MAKRIS, T. M. 2015. Decarboxylation of fatty acids to terminal alkenes by cytochrome P450 compound I. *Journal of the American Chemical Society,* 137**,** 4940-4943.

GREUNKE, C., DUELL, E. R., D'AGOSTINO, P. M., GLÖCKLE, A., LAMM, K. & GULDER, T. A. M. 2018. Direct Pathway Cloning (DiPaC) to unlock natural product biosynthetic potential. *Metabolic Engineering,* 47**,** 334-345.

GRIFFIN, M. A., SPAKOWICZ, D. J., GIANOULIS, T. A. & STROBEL, S. A. 2010. Volatile organic compound production by organisms in the genus *Ascocoryne* and a re-evaluation of myco-diesel production by NRRL 50072. *Microbiology,* 156**,** 3814-3829.

GSCHWEND, P., ZAFIRIOU, O. C. & GAGOSIAN, R. B. 1980. Volatile organic compounds in seawater from the Peru upwelling region 1, 2. *Limnology and Oceanography,* 25**,** 1044-1053.

GÜR, T. M. 2018. Review of electrical energy storage technologies, materials and systems: challenges and prospects for large-scale grid storage. *Energy & Environmental Science,* 11**,** 2696-2767.

GUREVICH, A., SAVELIEV, V., VYAHHI, N. & TESLER, G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics,* 29**,** 1072-1075.

HALL, D., MOULTAK, M. & LUTSEY, N. 2017. ELECTRIC VEHICLE CAPITALS OF THE WORLD DEMONSTRATING THE PATH TO ELECTRIC DRIVE. *ICCT White Paper*.

HARFORD, T. 2016. *How Rudolf Diesel's engine changed the world* [Online]. BBC World Service. Available: https://www.bbc.co.uk/news/business-38302874 [Accessed].

HARGER, M., ZHENG, L., MOON, A., AGER, C., AN, J. H., CHOE, C., LAI, Y.-L., MO, B., ZONG, D. & SMITH, M. D. 2012. Expanding the product profile of a microbial alkane biosynthetic pathway. *ACS Synthetic Biology,* 2**,** 59-62.

HAYNES, W., WICKERHAM, L. & HESSELTINE, C. 1955. Maintenance of cultures of industrially important microorganisms. *Applied Microbiology,* 3**,** 361.

HEATON, E. A., DOHLEMAN, F. G. & LONG, S. P. 2008. Meeting US biofuel goals with less land: the potential of *Miscanthus*. *Global Change Biology,* 14**,** 2000-2014.

HERN, A. & DORN, S. 2004. A female-specific attractant for the codling moth, *Cydia pomonella*, from apple fruit volatiles. *Naturwissenschaften,* 91**,** 77-80.

HOFFMEISTER, D. & KELLER, N. P. 2007. Natural products of filamentous fungi: enzymes, genes, and their regulation. *Natural Product Reports,* 24**,** 393-416.

HOLT, C. & YANDELL, M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics,* 12**,** 491.

HOWARD, R. W. & BLOMQUIST, G. J. 1982. CHEMICAL ECOLOGY AND BIOCHEMISTRY OF INSECT HYDROCARBONS. *Annual Review of Entomology,* 27**,** 149-172.

HOWARD, T. P., MIDDELHAUFE, S., MOORE, K., EDNER, C., KOLAK, D. M., TAYLOR, G. N., PARKER, D. A., LEE, R., SMIRNOFF, N. & AVES, S. J. 2013. Synthesis of customized petroleum-replica fuel molecules by targeted modification of free fatty acid pools in *Escherichia coli*. *Proceedings of the National Academy of Sciences,* 110**,** 7636-7641.

HSIEH, C. H., HUANG, X., AMAYA, J. A., RUTLAND, C. D., KEYS, C. L., GROVES, J. T., AUSTIN, R. N. & MAKRIS, T. M. 2017. The enigmatic P450 decarboxylase OleT is capable of, but evolved to frustrate, oxygen rebound chemistry. *Biochemistry,* 56**,** 3347-3357.

HSIEH, C. H. & MAKRIS, T. M. 2016. Expanding the substrate scope and reactivity of cytochrome P450 OleT. *Biochemical and Biophysical Research Communications,* 476**,** 462-466.

HSIEH, F.-L., CHANG, T.-H., KO, T.-P. & WANG, A. H.-J. 2011. Structure and mechanism of an *Arabidopsis* medium/long-chain-length prenyl pyrophosphate synthase. *Plant Physiology,* 155**,** 1079-1090.

HUELIN, F. & MURRAY, K. 1966. α-Farnesene in the natural coating of apples. *Nature,* 210**,** 1260.

HUIJBERS, M. M., ZHANG, W., TONIN, F. & HOLLMANN, F. 2018. Light-Driven Enzymatic Decarboxylation of Fatty Acids. *Angewandte Chemie International Edition,* 57**,** 13648-13651.

INGLIS, D. O., BINKLEY, J., SKRZYPEK, M. S., ARNAUD, M. B., CERQUEIRA, G. C., SHAH, P., WYMORE, F., WORTMAN, J. R. & SHERLOCK, G. 2013. Comprehensive annotation of secondary metabolite biosynthetic genes and gene clusters of *Aspergillus nidulans, A. fumigatus, A. niger and A. oryzae*. *BMC Microbiology,* 13**,** 91.

JENKS, M. A., EIGENBRODE, S. D. & LEMIEUX, B. 2002. Cuticular waxes of Arabidopsis. *The arabidopsis book,* 1**,** e0016-e0016.

JIA, C., LI, M., LI, J., ZHANG, J., ZHANG, H., CAO, P., PAN, X., LU, X. & CHANG, W. 2015. Structural insights into the catalytic mechanism of aldehyde-deformylating oxygenases. *Protein & Cell,* 6**,** 55-67.

JIN, H., PFEFFER, P., DOUDS, D., PIOTROWSKI, E., LAMMERS, P. & SHACHAR-HILL, Y. 2005. The uptake, metabolism, transport and transfer of nitrogen in an arbuscular mycorrhizal symbiosis. *New Phytologist,* 168**,** 687-696.

JOBSON, B. T., ALEXANDER, M. L., MAUPIN, G. D. & MUNTEAN, G. G. 2005. On-line analysis of organic compounds in diesel exhaust using a proton transfer reaction mass spectrometer (PTR-MS). *International Journal of Mass Spectrometry,* 245**,** 78-89.

JOHNSON, A. D., HANDSAKER, R. E., PULIT, S. L., NIZZARI, M. M., O'DONNELL, C. J. & DE BAKKER, P. I. 2008. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics,* 24**,** 2938-2939.

JORDAN, M. 2004. The Encyclopedia of Fungi of Britain and Europe, London, Frances Lincoln.

JOUBÈS, J., RAFFAELE, S., BOURDENX, B., GARCIA, C., LAROCHE-TRAINEAU, J., MOREAU, P., DOMERGUE, F. & LESSIRE, R. 2008. The VLCFA elongase gene family in Arabidopsis thaliana: phylogenetic analysis, 3D modelling and expression profiling. *Plant Molecular Biology,* 67**,** 547.

KANCHARLA, P., BONNETT, S. A. & REYNOLDS, K. A. 2016. *Stenotrophomonas maltophilia* OleC-Catalyzed ATP-Dependent Formation of Long-Chain Z-Olefins from 2-Alkyl-3-hydroxyalkanoic Acids. *Chembiochem,* 17**,** 1426-1429.

KATOH, K. & STANDLEY, D. M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution,* 30**,** 772-780.

KAWAHARA, T., IZUMIKAWA, M., KOZONE, I., HASHIMOTO, J., KAGAYA, N., KOIWAI, H., KOMATSU, M., FUJIE, M., SATO, N. & IKEDA, H. 2018. Neothioviridamide, a polythioamide compound produced by heterologous expression of a *Streptomyces sp.* cryptic RiPP biosynthetic gene cluster. *Journal of natural products,* 81**,** 264-269.

KELLER, N. P. & HOHN, T. M. 1997. Metabolic pathway gene clusters in filamentous fungi. *Fungal Genetics and Biology,* 21**,** 17-29.

KELLER, N. P., TURNER, G. & BENNETT, J. W. 2005. Fungal secondary metabolism —from biochemistry to genomics. *Nature Reviews Microbiology,* 3**,** 937.

KHALDI, N., SEIFUDDIN, F. T., TURNER, G., HAFT, D., NIERMAN, W. C., WOLFE, K. H. & FEDOROVA, N. D. 2010. SMURF: genomic mapping of fungal secondary metabolite clusters. *Fungal Genetics and Biology,* 47**,** 736-741.

KHARA, B., MENON, N., LEVY, C., MANSELL, D., DAS, D., MARSH, E. N. G., LEYS, D. & SCRUTTON, N. S. 2013. Production of propane and other short-chain alkanes by structure-based engineering of ligand specificity in aldehyde-deformylating oxygenase. *ChemBioChem,* 14**,** 1204-1208.

KINGHORN, J. & PATEMAN, J. 1973. NAD and NADP L-glutamate dehydrogenase activity and ammonium regulation in *Aspergillus nidulans*. *Microbiology,* 78**,** 39-46.

KOUPRINA, N. & LARIONOV, V. 2008. Selective isolation of genomic loci from complex genomes by transformation-associated recombination cloning in the yeast *Saccharomyces cerevisiae*. *Nature Protocols,* 3**,** 371.

KRIEG, T., SYDOW, A., FAUST, S., HUTH, I. & HOLTMANN, D. 2018. CO2 to Terpenes: Autotrophic and Electroautotrophic α-Humulene Production with *Cupriavidus necato*r. *Angewandte Chemie International Edition,* 57**,** 1879-1882.

KRUEGER, F. 2015. Trim galore. A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files.

KUNJAPUR, A. M. & PRATHER, K. L. 2015. Microbial engineering for aldehyde synthesis. *Appl. Environ. Microbiol.,* 81**,** 1892-1901.

KUNJAPUR, A. M., TARASOVA, Y. & PRATHER, K. L. 2014. Synthesis and accumulation of aromatic aldehydes in an engineered strain of *Escherichia coli*. *Journal of the American Chemical Society,* 136**,** 11644-11654.

KURATA, T., KAWABATA-AWAI, C., SAKURADANI, E., SHIMIZU, S., OKADA, K. & WADA, T. 2003. The YORE-YORE gene regulates multiple aspects of epidermal cell differentiation in *Arabidopsis*. *The Plant Journal,* 36**,** 55-66.

LEA-SMITH, D. J., ORTIZ-SUAREZ, M. L., LENN, T., NÜRNBERG, D. J., BAERS, L. L., DAVEY, M. P., PAROLINI, L., HUBER, R. G., COTTON, C. A. R., MASTROIANNI, G., BOMBELLI, P., UNGERER, P., STEVENS, T. J., SMITH, A. G., BOND, P. J., MULLINEAUX, C. W. & HOWE, C. J. 2016. Hydrocarbons Are Essential for Optimal Cell Size, Division, and Growth of Cyanobacteria. *Plant Physiology,* 172**,** 1928-1940.

LEAHY, J. G. & COLWELL, R. R. 1990. Microbial degradation of hydrocarbons in the environment. *Microbiology and Molecular Biology Reviews,* 54**,** 305-315.

LEE, E.-H. & CHO, K.-S. 2008. Characterization of cyclohexane and hexane degradation by *Rhodococcus sp. EC1*. *Chemosphere,* 71**,** 1738-1744.

LEHTINEN, T., SANTALA, V. & SANTALA, S. 2017. Twin-layer biosensor for real-time monitoring of alkane metabolism. *FEMS Microbiology Letters,* 364.

LI, R., FAN, W., TIAN, G., ZHU, H., HE, L., CAI, J., HUANG, Q., CAI, Q., LI, B. & BAI, Y. 2010. The sequence and de novo assembly of the giant panda genome. *Nature,* 463**,** 311.

LI, Z., CHEN, Y., MU, D., YUAN, J., SHI, Y., ZHANG, H., GAN, J., LI, N., HU, X. & LIU, B. 2012. Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph. *Briefings in Functional Genomics,* 11**,** 25-37.

LIU, Y., WANG, C., YAN, J., ZHANG, W., GUAN, W., LU, X. & LI, S. 2014. Hydrogen peroxide-independent production of α-alkenes by OleT JE P450 fatty acid decarboxylase. *Biotechnology for biofuels,* 7**,** 28.

LOGAN, B. E. 2009. Scaling up microbial fuel cells and other bioelectrochemical systems. *Applied Microbiology and Biotechnology,* 85**,** 1665-1671.

LU, F. & TEAL, P. 2001. Sex pheromone components in oral secretions and crop of male Caribbean fruit flies, *Anastrepha suspensa* (Loew). *Archives of Insect Biochemistry and Physiology: Published in Collaboration with the Entomological Society of America,* 48**,** 144-154.

M.M. MÜLLER, R. KANTOLA & V. KITUNEN 1994. Combining sterol and fatty acid profiles for the characterization of fungi. *Mycological Research,* 98**,** 593-603.

MAJOR, M. A. & BLOMQUIST, G. J. 1978. Biosynthesis of hydrocarbons in insects: Decarboxylation of long chain acids ton-alkanes in *Periplaneta*. *Lipids,* 13**,** 323-328.

MALLETTE, N. D., KNIGHTON, W. B., STROBEL, G. A., CARLSON, R. P. & PEYTON, B. M. 2012. Resolution of volatile fuel compound profiles from *Ascocoryne sarcoides*: a comparison by proton transfer reaction-mass spectrometry and solid phase microextraction gas chromatography-mass spectrometry. *AMB Express,* 2**,** 23.

MALLETTE, N. D., PANKRANTZ, E., BUSSE, S., STROBEL, G. A., CARLSON, R. P. & PEYTON, B. M. 2014. Evaluation of cellulose as a substrate for hydrocarbon fuel production by *Ascocoryne sarcoides* (NRRL 50072). *Journal of Sustainable Bioenergy Systems,* 4**,** 33-49.

MARZLUF, G. A. 1970. Genetic and biochemical studies of distinct sulfate permease species in different developmental stages of *Neurospora crassa*. *Archives of Biochemistry and Biophysics,* 138**,** 254-263.

MATHUR, H. & BHATTACHARYYA, S. 1963. 658. Macrocyclic musk compounds. Part V. New syntheses of exaltone, exaltolide, dihydroambrettolide, and Δ 9-

isoambrettolide from aleuritic acid. *Journal of the Chemical Society (Resumed)*, 3505-3509.

MAUGERI, L. 2006. The Age of Oil: the Mythology, History, and Future of the World's Most Controversial Resource, Praeger.

MEDEMA, M. H., BLIN, K., CIMERMANCIC, P., DE JAGER, V., ZAKRZEWSKI, P., FISCHBACH, M. A., WEBER, T., TAKANO, E. & BREITLING, R. 2011. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Research,* 39, W339-W346.

MELETIADIS, J., MEIS, J. F., MOUTON, J. W. & VERWEIJ, P. E. 2001. Analysis of growth characteristics of filamentous fungi in different nutrient media. *Journal of Clinical Microbiology,* 39, 478-484.

METZLER, B. 1997. Quantitative assessment of fungal colonization in Norway spruce after green pruning. *European Journal of Forest Pathology,* 27, 1-11.

MEYER, M. & KIRCHER, M. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols,* 2010, pdb. prot5448.

MI, H., MURUGANUJAN, A., CASAGRANDE, J. T. & THOMAS, P. D. 2013. Large-scale gene function analysis with the PANTHER classification system. *Nature Protocols,* 8, 1551.

MOHANRAM, S., AMAT, D., CHOUDHARY, J., ARORA, A. & NAIN, L. 2013. Novel perspectives for evolving enzyme cocktails for lignocellulose hydrolysis in biorefineries. *Sustainable Chemical Processes,* 1, 15.

MOORE, S. B., PORALLA, K. & FLOSS, H. G. 1993. Biosynthesis of the cyclohexanecarboxylic acid starter unit of .omega.-cyclohexyl fatty acids in *Alicyclobacillus acidocaldarius*. *Journal of the American Chemical Society,* 12, 5267-5274.

MORIYA, Y., ITOH, M., OKUDA, S., YOSHIZAWA, A. C. & KANEHISA, M. 2007. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Research,* 35, W182-W185.

MORRISON, R. & BOYD, R. 1988. Organic chemistry, 1992. Prentice Hall, Inc.

MUSAT, F., WILKES, H., BEHRENDS, A., WOEBKEN, D. & WIDDEL, F. 2010. Microbial nitrate-dependent cyclohexane degradation coupled with anaerobic ammonium oxidation. *The ISME journal,* 4, 1290.

NAGY, E. Z. A., NAGY, C. L., FILIP, A., NAGY, K., GÁL, E., TŐTŐS, R., POPPE, L., PAIZS, C. & BENCZE, L. C. 2019. Exploring the substrate scope of ferulic acid decarboxylase (FDC1) from *Saccharomyces cerevisiae*. *Scientific Reports,* 9**,** 647.

NAHAR, K. & OZORES-HAMPTON, M. 2011. *Jatropha*: An Alternative Substitute to Fossil Fuel. *Horticultural Sciences Department, UF/IFAS Extension,* HS1193.

NAN, X., MEEHAN, R. R. & BIRD, A. 1993. Dissection of the methyl-CpG binding domain from the chromosomal protein MeCP2. *Nucleic Acids Research,* 21**,** 4886-4892.

NEWCOMER, M. E. & BRASH, A. R. 2015. The structural basis for specificity in lipoxygenase catalysis. *Protein Science,* 24**,** 298-309.

OGUNDERO, V. W. 1987. Temperature and aflatoxin production by *Aspergillus flavus* and *A. parasiticus* strains from Nigerian groundnuts. *Journal of Basic Microbiology,* 27**,** 511-514.

OLIYNYK, M., SAMBORSKYY, M., LESTER, J. B., MIRONENKO, T., SCOTT, N., DICKENS, S., HAYDOCK, S. F. & LEADLAY, P. F. 2007. Complete genome sequence of the erythromycin-producing bacterium *Saccharopolyspora erythraea NRRL23338*. *Nature Biotechnology,* 25**,** 447.

PANDELIA, M. E., LI, N., NØRGAARD, H., WARUI, D. M., RAJAKOVICH, L. J., CHANG, W.-C., BOOKER, S. J., KREBS, C. & BOLLINGER JR, J. M. 2013. Substrate-triggered addition of dioxygen to the diferrous cofactor of aldehyde-deformylating oxygenase to form a diferric-peroxide intermediate. *Journal of the American Chemical Society,* 135**,** 15801-15812.

PAUL, B., DAS, D., ELLINGTON, B. & MARSH, E. N. G. 2013. Probing the mechanism of cyanobacterial aldehyde decarbonylase using a cyclopropyl aldehyde. *Journal of the American Chemical Society,* 135**,** 5234-5237.

PAVLIDIS, T., ILIEVA, M., BENCHEVA, S. & STANCHEVA, J. 2005. Researches on wood-destroying fungi division Ascomycota, classis *Ascomycetes*. *Matica Srpska Proceedings for Natural Sciences*.

PEDRINI, N., JUÁREZ, M. P., CRESPO, R. & DE ALANIZ, M. J. 2006. Clues on the role of *Beauveria bassiana* catalases in alkane degradation events. *Mycologia,* 98**,** 528-534.

PERALTA-YAHYA, P. P., ZHANG, F., DEL CARDAYRE, S. B. & KEASLING, J. D. 2012. Microbial engineering for the production of advanced biofuels. *Nature,* 488**,** 320.

POTTER, N. I. 2008. How Brazil achieved energy independence and the lessons the United States should learn from Brazil's experience. *Wash. U. Global Stud,* L.

QIU, Y., TITTIGER, C., WICKER-THOMAS, C., LE GOFF, G., YOUNG, S., WAJNBERG, E., FRICAUX, T., TAQUET, N., BLOMQUIST, G. J. & FEYEREISEN, R. 2012. An insect-specific P450 oxidative decarbonylase for cuticular hydrocarbon biosynthesis. *Proceedings of the National Academy of Sciences,* 109**,** 14858-14863.

QUACK, W., SCHOLL, H. & BUDZIKIEWICZ, H. 1980. Ascocorynin, a terphenylquinone from *Ascocoryne sarcoides*. *Phytochemistry,* 21**,** 2921-2923.

RADWAN, S. & SOLIMAN, A. H. 1988. Arachidonic acid from fungi utilizing fatty acids with shorter chains as sole sources of carbon and energy. *Microbiology,* 134**,** 387-393.

REED, J. R., VANDERWEL, D., CHOI, S., POMONIS, J. G., REITZ, R. C. & BLOMQUIST, G. J. 1994. Unusual mechanism of hydrocarbon formation in the housefly: cytochrome P450 converts aldehyde to the sex pheromone component (Z)-9-tricosene and CO2. *Proceedings of the National Academy of Sciences of the United States of America,* 91**,** 10000-10004.

REGUEIRA, T. B., KILDEGAARD, K. R., HANSEN, B. G., MORTENSEN, U. H., HERTWECK, C. & NIELSEN, J. 2011. Molecular basis for mycophenolic acid biosynthesis in *Penicillium brevicompactum*. *Appl. Environ. Microbiol.,* 77**,** 3035-3043.

RENNINGER, N. S. & MCPHEE, D. J. 2008. Fuel compositions comprising farnesane and farnesane derivatives and method of making and using same. Google Patents.

RENNINGER, N. S. & MCPHEE, D. J. 2010. Fuel compositions comprising farnesane and farnesane derivatives and method of making and using same. Google Patents.

ROLL-HANSEN, F. & ROLL-HANSEN, H. 1979. *Ascocoryne* species in living stems of Picea species A literature review. *European Journal of Forest Pathology,* 9**,** 275-280.

RÖSECKE, J., PIETSCH, M. & KÖNIG, W. A. 2000. Volatile constituents of wood-rotting basidiomycetes. *Phytochemistry,* 54**,** 747-750.

RUDE, M. A., BARON, T. S., BRUBAKER, S., ALIBHAI, M., DEL CARDAYRE, S. B. & SCHIRMER, A. 2011. Terminal olefin (1-alkene) biosynthesis by a novel P450 fatty acid decarboxylase from *Jeotgalicoccus species*. *Appl. Environ. Microbiol.,* 77**,** 1718-1727.

RUI, Z., HARRIS, N. C., ZHU, X., HUANG, W. & ZHANG, W. 2015. Discovery of a family of desaturase-like enzymes for 1-alkene biosynthesis. *ACS Catalysis,* 5**,** 7091-7094.

RUI, Z., LI, X., ZHU, X., LIU, J., DOMIGAN, B., BARR, I., CATE, J. H. D. & ZHANG, W. 2014. Microbial biosynthesis of medium-chain 1-alkenes by a nonheme iron oxidase. *Proceedings of the National Academy of Sciences,* 111**,** 18237-18242.

RUZICKA, L. 1926. Zur Kenntnis des Kohlenstoffringes VII. Über die Konstitution des Muscons. *Helvetica Chimica Acta,* 9**,** 715-729.

SAMSON, R., MANI, S., BODDEY, R., SOKHANSANJ, S., QUESADA, D., URQUIAGA, S., REIS, V. & HO LEM, C. 2005. The Potential of C4 Perennial Grasses for Developing a Global BIOHEAT Industry. *Critical Reviews in Plant Sciences,* 24**,** 461-495.

SCHIRMER, A., RUDE, M. A., LI, X., POPOVA, E. & DEL CARDAYRE, S. B. 2010. Microbial Biosynthesis of Alkanes. *Science,* 329**,** 559-562.

SCHMIDT-DANNERT, C. 2015. NextGen microbial natural products discovery. *Microbial biotechnology,* 8**,** 26.

SCHMITT, E. K. & KÜCK, U. 2000. The fungal CPCR1 protein, which binds specifically to β-lactam biosynthesis genes, is related to human regulatory factor X transcription factors. *Journal of Biological Chemistry,* 275**,** 9348-9357.

SCHNEIDER, C., TALLMAN, K. A., PORTER, N. A. & BRASH, A. R. 2001. Two distinct pathways of formation of 4-hydroxynonenal mechanisms of nonenzymatic transformation of the 9-and 13-hydroperoxides of linoleic acid to 4-hydroxyalkenals. *Journal of Biological Chemistry,* 276**,** 20831-20838.

SCHOEN, H. R., HUNT, K. A., STROBEL, G. A., PEYTON, B. M. & CARLSON, R. P. 2017. Carbon chain length of biofuel-and flavor-relevant volatile organic compounds produced by lignocellulolytic fungal endophytes changes with culture temperature. *Mycoscience,* 58**,** 338-343.

SCHULZ, S. & DICKSCHAT, J. S. 2007. Bacterial volatiles: the smell of small organisms. *Natural product reports,* 24**,** 814-842.

SCHWARZENBACH, R. P., BROMUND, R. H., GSCHWEND, P. M. & ZAFIRIOU, O. C. 1978. Volatile organic compounds in coastal seawater. *Organic Geochemistry,* 1**,** 93-107.

SEUBERT, W., LAMBERTS, I., KRAMER, R. & OHLY, B. 1968. On the mechanism of malonyl-CoA-independent fatty acid synthesis: I. The mechanism of elongation of long-chain fatty acids by acetyl-CoA. *Biochimica et Biophysica Acta (BBA)-Lipids and Lipid Metabolism,* 164**,** 498-517.

SHAW, J. J., SPAKOWICZ, D. J., DALAL, R. S., DAVIS, J. H., LEHR, N. A., DUNICAN, B. F., ORELLANA, E. A., NARVÁEZ-TRUJILLO, A. & STROBEL, S. A. 2015. Biosynthesis and genomic analysis of medium-chain hydrocarbon production by the endophytic fungal isolate *Nigrograna mackinnonii E5202H. Applied Microbiology and Biotechnology,* 99**,** 3715-3728.

SHEPPARD, M. J., KUNJAPUR, A. M. & PRATHER, K. L. 2016. Modular and selective biosynthesis of gasoline-range alkanes. *Metabolic Engineering,* 33**,** 28-40.

SIMÃO, F. A., WATERHOUSE, R. M., IOANNIDIS, P., KRIVENTSEVA, E. V. & ZDOBNOV, E. M. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics,* 31**,** 3210-3212.

SINGHANIA, R. R., PATEL, A. K., SUKUMARAN, R. K., LARROCHE, C. & PANDEY, A. 2013. Role and significance of beta-glucosidases in the hydrolysis of cellulose for bioethanol production. *Bioresource technology,* 127**,** 500-507.

ŠOBOTNÍK, J., JIROŠOVÁ, A. & HANUS, R. 2010. Chemical warfare in termites. *Journal of Insect Physiology,* 56**,** 1012-1021.

SOCCOL, C., VANDENBERGHE, L., COSTA, B., WOICIECHOWSKI, A., DE CARVALHO, J., MEDEIROS, A., FRANCISCO, A. M. & BONOMI, L. J. 2005. Brazilian biofuel program: An overview. *Journal of Scientific and Industrial Research,* 64**,** 897-904.

SONG, J., SHI, L., LI, D., SUN, Y., NIU, Y., CHEN, Z., LUO, H., PANG, X., SUN, Z. & LIU, C. 2012. Extensive pyrosequencing reveals frequent intra-genomic variations of internal transcribed spacer regions of nuclear ribosomal DNA. *PloS One,* 7**,** e43971.

SORIGUÉ, D., LÉGERET, B., CUINÉ, S., BLANGY, S., MOULIN, S., BILLON, E., RICHAUD, P., BRUGIÈRE, S., COUTÉ, Y. & NURIZZO, D. 2017. An algal photoenzyme converts fatty acids to hydrocarbons. *Science,* 357**,** 903-907.

STANKE, M., STEINKAMP, R., WAACK, S. & MORGENSTERN, B. 2004. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic acids research,* 32**,** W309-W312.

STINSON, M., EZRA, D., HESS, W., SEARS, J. & STROBEL, G. 2003. An endophytic *Gliocladium sp.* of *Eucryphia cordifolia* producing selective volatile antimicrobial compounds. *Plant Science,* 4**,** 913-922.

STIRLING, L. A., WATKINSON, R. & HIGGINS, I. 1977. Microbial metabolism of alicyclic hydrocarbons: isolation and properties of a cyclohexane-degrading bacterium. *Microbiology,* 99**,** 119-125.

STROBEL, G., TOMSHECK, A., GEARY, B., SPAKOWICZ, D., STROBEL, S., MATTNER, S. & MANN, R. 2010a. Endophyte strain *NRRL 50072* producing volatile organics is a species of *Ascocoryne*. *Mycology,* 1**,** 187-194.

STROBEL, G. A., KNIGHTON, W. B., KLUCK, K., REN, Y., LIVINGHOUSE, T., GRIFFIN, M., SPAKOWICZ, D. & SEARS, J. 2010b. The production of myco-diesel hydrocarbons and their derivatives by the endophytic fungus *Gliocladium roseum* (NRRL 50072). *Microbiology,* 156**,** 3830-3833.

SU, C. & OLIW, E. H. 1998. Manganese lipoxygenase purification and characterization. *Journal of Biological Chemistry,* 273**,** 13072-13079.

SU, C., SAHLIN, M. & OLIW, E. H. 2000. Kinetics of manganese lipoxygenase with a catalytic mononuclear redox center. *Journal of Biological Chemistry,* 275**,** 18830-18835.

SUGIHARA, S., HORI, R., NAKANOWATARI, H., TAKADA, Y., YUMOTO, I., MORITA, N., YANO, Y., WATANABE, K. & OKUYAMA, H. 2010. Possible Biosynthetic Pathways for all cis-3, 6, 9, 12, 15, 19, 22, 25, 28-Hentriacontanonaene in Bacteria. *Lipids,* 45**,** 167-177.

SUKOVICH, D. J., SEFFERNICK, J. L., RICHMAN, J. E., HUNT, K. A., GRALNICK, J. A. & WACKETT, L. P. 2010. Structure, function, and insights into the biosynthesis of a head-to-head hydrocarbon in Shewanella oneidensis strain MR-1. *Appl. Environ. Microbiol.,* 76**,** 3842-3849.

TANG, X., LI, J., MILLÁN-AGUIÑAGA, N., ZHANG, J. J., O'NEILL, E. C., UGALDE, J. A., JENSEN, P. R., MANTOVANI, S. M. & MOORE, B. S. 2015. Identification of thiotetronic acid antibiotic biosynthetic pathways by target-directed genome mining. *ACS Chemical Biology,* 10**,** 2841-2849.

TARAILO-GRAOVAC, M. & CHEN, N. 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics,* 25**,** 4.10. 1-4.10. 14.

TEMPLE, S. J., VANCE, C. P. & GANTT, J. S. 1998. Glutamate synthase and nitrogen assimilation. *Trends in plant science,* 3**,** 51-56.

TESLA. 2019. *On the Road* [Online]. Tesla. Available: https://www.tesla.com/en_GB/supercharger?redirect=no [Accessed 02/05/2019 2019].

THEVENIEAU, F., LE DALL, M.-T., NTHANGENI, B., MAUERSBERGER, S., MARCHAL, R. & NICAUD, J.-M. 2007. Characterization of *Yarrowia lipolytica* mutants affected in hydrophobic substrate utilization. *Fungal Genetics and Biology,* 44**,** 531-542.

THOMAS, G. 2000. Overview of storage development DOE hydrogen program. *Sandia National Laboratories*.

THOMAS, P. D., CAMPBELL, M. J., KEJARIWAL, A., MI, H., KARLAK, B., DAVERMAN, R., DIEMER, K., MURUGANUJAN, A. & NARECHANIA, A. 2003. PANTHER: a library of protein families and subfamilies indexed by function. *Genome research,* 13**,** 2129-2141.

TSUJI, G., KENMOCHI, Y., TAKANO, Y., SWEIGARD, J., FARRALL, L., FURUSAWA, I., HORINO, O. & KUBO, Y. 2000. Novel fungal transcriptional activators, Cmr1p of *Colletotrichum lagenarium* and Pig1p of *Magnaporthe grisea*, contain Cys2His2 zinc finger and Zn (II) 2Cys6 binuclear cluster DNA-binding motifs and regulate transcription of melanin biosynthesis genes in a developmentally specific manner. *Molecular Microbiology,* 38**,** 940-954.

UNFCCC 1997. Kyoto Protocol to the United Nations Framework Convention on Climate Change adopted at COP3 in Kyoto. *In:* UN (ed.). Kyoto.

VAN BEILEN, J. B. & FUNHOFF, E. G. 2007. Alkane hydroxylases involved in microbial alkane degradation. *Applied Microbiology and Biotechnology,* 74**,** 13-21.

VAN DEN BERG, M. A., ALBANG, R., ALBERMANN, K., BADGER, J. H., DARAN, J.-M., DRIESSEN, A. J., GARCIA-ESTRADA, C., FEDOROVA, N. D., HARRIS, D. M. & HEIJNE, W. H. 2008. Genome sequencing and analysis of the filamentous fungus *Penicillium chrysogenum*. *Nature Biotechnology,* 26**,** 1161.

VARJANI, S. J. 2017. Microbial degradation of petroleum hydrocarbons. *Bioresource Technology,* 223**,** 277-286.

VAZ, A. H., JURENKA, R. A., BLOMQUIST, G. J. & REITZ, R. C. 1988. Tissue and chain length specificity of the fatty acyl-CoA elongation system in the American cockroach. *Archives of Biochemistry and Biophysics,* 267**,** 551-557.

WANG, C., ZHAO, C., HU, L. & CHEN, H. 2016. Calculated mechanism of cyanobacterial aldehyde-deformylating oxygenase: asymmetric aldehyde activation by a symmetric diiron cofactor. *The journal of physical chemistry letters,* 7**,** 4427-4432.

WANG, Q., BAO, L., JIA, C., LI, M., LI, J.-J. & LU, X. 2017. Identification of residues important for the activity of aldehyde-deformylating oxygenase through investigation into the structure-activity relationship. *BMC biotechnology,* 17**,** 31.

WARUI, D. M., LI, N., NØRGAARD, H., KREBS, C., BOLLINGER JR, J. M. & BOOKER, S. J. 2011. Detection of formate, rather than carbon monoxide, as the stoichiometric coproduct in conversion of fatty aldehydes to alkanes by a cyanobacterial aldehyde decarbonylase. *Journal of the American Chemical Society,* 133**,** 3316-3319.

WIDÉN, K.-G., ALANKO, P. & UOTILA, M. *Thymus serpyllum L.× vulgaris L.*, morphology, chromosome number and chemical composition. *Annales Botanici Fennici*, 1977. *JSTOR*, 29-34.

WIELOCH, W. 2006. Chromosome visualisation in filamentous fungi. *Journal of microbiological methods,* 67**,** 1-8.

WIJFFELS, R. H., KRUSE, O. & HELLINGWERF, K. J. 2013. Potential of industrial biotechnology with cyanobacteria and eukaryotic microalgae. *Current Opinion in Biotechnology,* 24**,** 405-413.

WILLIAMS, R. B., HENRIKSON, J. C., HOOVER, A. R., LEE, A. E. & CICHEWICZ, R. H. 2008. Epigenetic remodeling of the fungal secondary metabolome. *Organic & Biomolecular Chemistry,* 6**,** 1895-1897.

WINTERS, K., PARKER, P. & VAN BAALEN, C. 1969. Hydrocarbons of blue-green algae: geochemical significance. *Science,* 163**,** 467-468.

WORLEY, K. C. & GIBBS, R. A. 2010. Genetics: Decoding a national treasure. *Nature,* 463**,** 303.

XU, L., DONG, Z., FANG, L., LUO, Y., WEI, Z., GUO, H., ZHANG, G., GU, Y. Q., COLEMAN-DERR, D. & XIA, Q. 2019. OrthoVenn2: a web server for whole-genome comparison and annotation of orthologous clusters across multiple species. *Nucleic Acids Research*.

YAKIMOV, M. M., TIMMIS, K. N. & GOLYSHIN, P. N. 2007. Obligate oil-degrading marine bacteria. *Current Opinion in Biotechnology,* 18**,** 257-266.

YEH, H.-H., AHUJA, M., CHIANG, Y.-M., OAKLEY, C. E., MOORE, S., YOON, O., HAJOVSKY, H., BOK, J.-W., KELLER, N. P. & WANG, C. C. 2016. Resistance gene-guided genome mining: serial promoter exchanges in *Aspergillus nidulans* reveal the biosynthetic pathway for fellutamide B, a proteasome inhibitor. *ACS Chemical Biology,* 11**,** 2275-2284.

ZEMACH, A. & GRAFI, G. 2003. Characterization of *Arabidopsis thaliana* methyl-CpG-binding domain (MBD) proteins. *The Plant Journal,* 34**,** 565-572.

ZHANG, X. & ELLIOT, M. A. 2019. Unlocking the trove of metabolic treasures: activating silent biosynthetic gene clusters in bacteria and fungi. *Current Opinion in Microbiology,* 51**,** 9-15.

ZULETA, E. C., BAENA, L., RIOS, L. A. & CALDERÓN, J. A. 2012. The oxidative stability of biodiesel and its impact on the deterioration of metallic and polymeric materials: a review. *Journal of the Brazilian Chemical Society,* 23**,** 2159-2175.