Genome Guided Enzyme Discovery in the Extremophile *Galdieria sulphuraria*

Sarah Cloud Lauren Lock

PhD

University of York

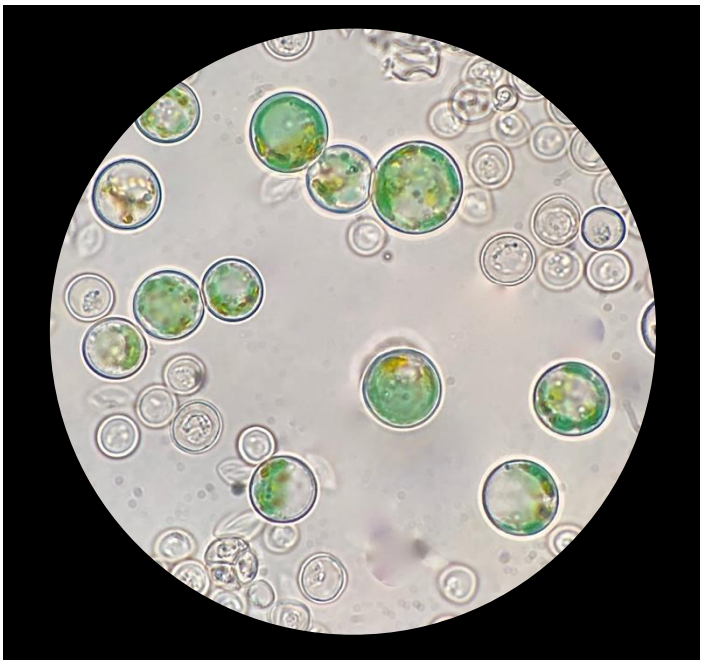
Biology

December 2021

Abstract

Galdieria sulphuraria is a eukaryotic unicellular red alga that predominates geothermal sites with low pH (0-2) and high temperatures (50-56 °C), the absolute limits of eukaryotic life. It can grow photoautotrophically and use a vast array of sugars, organic acids and polyols to support both heterotrophic and mixotrophic growth; making this alga an interesting focus for investigation on novel enzyme discovery. Based on its lifestyle, it seems likely that its enzymes may to be highly thermostable (for a eukaryote) and its secreted enzymes should display high acid tolerance. Consequently, any protein products discovered are likely to be robust and well suited for industrial biotechnology (IB) applications.

Firstly, I resolved the nuclear phylogeny and investigated evolutionary pressures acting on the *Galdieria* genus. This revealed the subdivision of the *G. sulphuraria* into six lineages. Analysis of dN/dS rates showed different evolutionary pressures acting between the strains and revealed a selection of genes under positive selection, one of these had predicted involvement in the degradation of lignocellulosic material. Secondly, I obtained transcriptomic and long read DNA sequence data to annotate the core six genomes, subsequent extensive CAZymes analysis showed ~128 enzymes per strain. Fewer than expected CAZyme families were represented given *G. sulphuraria*'s extraordinary growth capacity. This led to investigating heterotrophic growth on different carbohydrate polymers to identify industrially relevant secreted enzymes. Extraction of proteins from the supernatants and analysis via LCMS showed the presence of potentially interesting enzymes, prompting further investigation. Three target genes were identified and selected for heterologous expression and characterisation. A previously uncharacterised gene was successfully purified and refolded at pH 2 and shown to denature at 92 °C. There is scope to develop *G. sulphuraria*'s acid tolerant, thermostable proteins for industrial use. This thesis has expanded understanding of this extremophile and identified multiple novel enzymes with potential for industrial development.



Light micrograph at 400X magnification of *Galdieria sulphuraria* strain ACUF 074W grown mixotrophically on 2 % xylan with Allen medium pH2.

Table of Contents

List of Figures

List of Tables

Acknowledgments

This PhD has been a roller coaster to say the least, incredibly stressful at times and lots of fun at others. It is by far the most challenging thing I have faced and managing to make it to the end has left me with a lot of people to thank.

Firstly, thank you to Seth for giving me the opportunity to do this PhD especially with the little knowledge I had starting. Thank you for teaching me so much along the way and for all your guidance and support and letting me run with any idea that came to mind. Many thanks also go to Dan for teaching me everything I know about evolution and generally being supportive and excited about my work. I would also like to thank my undergraduate supervisor Marina Knight, who was the main inspiration behind me pursuing a career in academic research and led me onto this path. Simon a great thank you for all the TAP support, always providing good research advice with a positive attitude and making the viva a very enjoyable experience.

Next, I would like to thank all the people that have supported me in the lab over the years. James, for being a huge help and answering all of my incessant questions patiently, Manuela for being my Galdieria fountain of knowledge and making conferences fun. A big thank you to the rest of the Davis lab during my time, for keeping me positive throughout the PhD. Last and most defiantly not least a massive thank you to Mandi for always fighting my corner and being a source of laughter and support.

Throughout this PhD I have met some amazing people who all deserve thanks. The gals that got me through, Jess, Grace, Grace and Annie, we all did it together! Love you all and cannot wait for when we are all done and can go to flares and dance to our hearts content. An extra shoutout for Jess who was an excellent housemate and always there for me turning a bad day into a good one, I really couldn't have done this without you. Lewis for being entertaining and a great friend, for encouraging me to live my most chaotic side with you, shoutout to our nights in 13s where most of our stipend disappeared. I'd also like to give thanks to Charlotte for always making me feel positive and general chats whenever they were needed. For Chole thank you so much for being the wonderful friend that you are, I'm glad to have you by my side as I did through all of this.

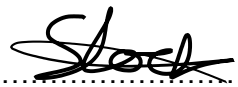
Jamie, you get your very own paragraph. I don't know how I can say thank you enough for your support. Especially in these last few months of writing I have been difficult to be

around at the best of times. You have never failed to bring me up and put a smile on my face when I've been down or didn't believe that I could do it.

Lastly to my family, Mum, Ben, Sam and Zachary thank you for encouraging me and keeping me going across all the degrees especially during this very hard thesis. A very big extra thanks to my Mum for putting up with my dramatic phone calls nearly daily and offering your words of wisdom throughout. I also don't think I could have done this without Todd and Oti.

Declaration

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References.

Signed  ……..

Chapter 1 - Introduction

1.1 Overview

This thesis explores the use of biotechnology for the sustainable production of biochemicals and biofuels from a renewable source. There are many difficulties facing the globe today and the expiration date on fossil fuels is fast approaching. Therefore, it is useful to look at nature to identify solutions and alternatives to this problem. Often harsh and extreme environments contain organisms that possess the tools for increasing efficiency of industrial processes. As such this thesis investigates the unusual biology of an extremophilic eukaryotic alga with a view to utilising any relevant enzymes for the degradation of plant material into biofuels.

1.2 Global resource insecurity

Climate change and the consumption and depletion of fossil fuels is an incredible threat facing humanity. Approximately 84% of global energy usage stems from fossil fuels, a significant contributing factor in climate change (bp Statistical Review of World Energy 2021). As global petroleum reserves are depleted, it is imperative a move is made to sources of fuel that promote sustainability of growing energy demands, moving away from damaging fossil fuels and toward a greener future. An ambitious 80% reduction in greenhouse gas emissions released into the atmosphere each year by 2030 was set as a goal in 2017 by the United Nations (United Nations Environment, 2017). Likewise, the European Union has pledged to take drastic action and a global leader in renewable energy, with at least 32 % of its energy originating from renewable sources by 2030 (European Commission, 2020). The population is predicted to grow to over 8.5 billion by 2030, (United Nations World Population Prospects, 2019) bringing an increase in energy consumption with it. Sustainable energy and fuel sources are urgently required to come anywhere close to meeting these goals (United Nations World Population Prospects, 2019).

The growing consciousness of humanity's impact on the globe through harvesting and combustion of fossil fuels has encouraged the search for an alternative fuel source. Clean energy can be harnessed from sustainable, natural sources of energy, for example wind, sun, and water. Biomass is also a key sustainable energy source, providing a liquid fuel that can be used for transportation (Alalwan et al., 2019). As such refining biomass to fuel has become a popular option for the development of alternative fuel sources. There are many advantages to using fuels refined from biomass (known as biofuels), over traditional petroleum feedstocks. Biofuels are not only reliable fuel obtained from renewable sources but can be sustainable, non-polluting, accessible and locally

available (Demirbas, 2009). Long term renewable energy systems, such as biofuels, may be a key requirement in supporting ecosystems and populations across the globe for years to come (Demirbas, 2009; Alalwan *et al*., 2019). Biofuels are recognised as offering a promising solution to reduction in fossil fuel consumption, and as such, play an important role in plans to meet targets set.

1.2.1 Biofuels

'Biofuel' is the umbrella term for any fuel produced by the conversion of biomass into liquid or gas fuel. Such examples are ethanol, lipids, biogas (typically a mixture of methane and carbon dioxide), or hydrogen. These are acquired through biological and/or chemical processes. Several feedstocks can be used to make bioethanol (a liquid biofuel), a highly useful chemical compound that has many uses in industrial applications in addition to an alternative fuel. Bioethanol can be used as a fuel in its pure form, or more typically in combination with petroleum or diesel (Sarkar et al., 2012; Rezania et al., 2020). As the resources from fossil fuels are becoming more limited, bioethanol presents possibly the most attractive alternative when it comes to liquid fuels (Demirbas, 2008; Demirbas, 2009).

Biofuels can be categorised into first, second and third generations, based primarily on the type of feedstock used in the production of the fuel. Summarising, first generation fuels are typically manufactured from edible agricultural feedstocks through fermentation of sugars or starch present in the crops. However, the use of such feedstocks may have a negative impact on global food security (Demirbas, 2009; Dutta et al., 2014; Alalwan et al., 2019; Rezania et al., 2020).

Second generation biofuels are also produced using plant material but use woody, non-edible biomass, known as lignocellulosic biomass. For example, left over crop residues or dedicated biomass. In order to synthesise fuel there are a series of pre-treatment steps, followed by enzymatic hydrolysis and fermentation of the resulting sugars, eventually resulting in bioethanol (Dutta et al., 2014; Rezania et al., 2020). The most recent third generation biofuels are produced by using algal biomass as a feedstock to produce biofuel (Behera et al., 2015).

1.2.2 First generation Biofuel

First-generation biofuels are mainly acquired from consumable biomass, such as starch from corn, barley, wheat and potato, or sugar from sugar beet and sugarcane, or any type of vegetable oil (Alalwan et al., 2019). The biofuels produced such as butanol,

propanol and ethanol are manufactured through the fermentation of such biomass (Dahman et al., 2019). Initially first-generation biofuels showed promising capability in minimising fossil fuel consumption and hence lowering atmospheric levels of CO_2 (Rodionova et al., 2017). However, as these first-generation biofuels were primarily comprised of edible crops as feedstocks, concerns arose that there may be significant impacts on biodiversity, indirect land use change and thus food supply (Alalwan et al., 2019; Dahman et al., 2019). Although first-generation biofuel processes are useful there is a threshold at which they cannot produce enough biofuel without competing for valuable resources (Rodionova et al., 2017; Alalwan et al., 2019). One alternative to limit these impacts could be by using non-edible feedstocks to produce biofuels. Utilising the in edible parts of crops and other crop residues may provide a solution. These fuels are known as second-generation biofuels. As a direct result of the issues presented by first-generation biofuels, the European Union has committed to phasing them out in favour of second-generation biofuels by 2030 (European Union, 2018).

1.2.3 Second-Generation Biofuel

The biomass used to manufacture second-generation biofuels can be broadly classified into four sources. These are agricultural, energy crops, forest residues and cellulosic wastes (Rezania et al., 2020). Lignocellulosic material falls into the cellulosic waste category and makes up the majority of non-edible plant material. It is both abundant and inexpensive. Examples of lignocellulosic biomass include corn stover (the leaves, stalks, husks and cobs left after harvest), and perennial grasses like wheat straw, miscanthus and switchgrass. Wood sourced from forest logging and processing and material from short rotation woody crops are also sources of lignocellulosic biomass suitable for use in biofuel production (Naik et al., 2010; Rezania et al., 2020). The plant cell wall polysaccharides within lignocellulosic biomass have a high sugar content, providing a rich source of fermentable sugars which can be utilised to produce bioethanol. Plan biomass represents one of the largest and yet underutilised biological resources globally (Naik et al., 2010).

The production of second-generation biofuels uses a more sustainable protocol than the first-generation and have many environmental advantages as well as also being relatively inexpensive. The combustion from second generation biofuels gives a net carbon emission (emitted–consumed) that is neutral or even negative (Alalwan et al., 2019; Geismar et al., 2021). Consequently, it is anticipated that second-generation biofuels could significantly reduce carbon dioxide production and even contribute to offsetting carbon usage elsewhere. Additionally, they do not compete with current food

crops or land use and can offer a beneficial way of recycling otherwise waste material (Naik et al., 2010; Geismar et al., 2021).

To convert lignocellulosic biomass into biofuels, the lignocellulose fermentable sugars need to be released by cleavage of the hydrolytically resistant polymers. This is achieved in four main steps: pre-treatment, saccharification (through acid or enzymatic hydrolysis), fermentation and distillation. Pre-treatment initially aim to breakdown the cell wall structures, reduce the crystallinity and particle size while increasing the porosity and accessibility for the next step (Houfani et al., 2020). This can be achieved in two ways, either though physical treatment like grinding or milling or by chemical means (acids, bases, organic solvents). In practice due to the recalcitrance nature of the biomass a combined approach will yield the best results (Chen et al., 2017; Houfani et al., 2020). After pre-treatment, cellulose and hemicellulose polymers are more accessible for saccharification of polysaccharides into fermentable sugars through enzymatic or acid hydrolysis (Sun and Cheng, 2002; Amiri and Karimi, 2018; Świątek et al., 2020). Lastly the sugars are fermented and distilled into fuels (Liguor et al., 2017).

1.3 Lignocellulosic material

The abundance of lignocellulosic biomass combined with the high quality of polysaccharides make it a rich source for converting into sugars. The conversion of lignocellulosic biomass to bioethanol requires four processes, these are, pre-treatment of the biomass, followed by acid and/or an enzymatic hydrolysis, then fermentation, and finally distillation (Naik et al., 2010; Sims et al., 2010). Although lignocellulose presents a promising feedstock, its naturally durable structure can be challenging to breakdown. Lignocellulose has evolved to be resistant to degradation, protecting the plant against both biological and chemical attacks. This obstructs access to the plant's monosaccharides, hence making the process of degradation difficult (Bajpai, 2016; Alalwan et al., 2019). The conversion of this material into fuel requires a process known as saccharification, where polysaccharides are hydrolysed to monosaccharides (Bajpai, 2016).

Cellulose, hemicellulose and lignin are the three main structural units of lignocellulosic biomass and are characterised through their large, complex structures consisting of repeating cyclic units with differing functional groups (Isikgor and Becer, 2015; Bajpai, 2016). Plant cells contain two cell walls (primary and secondary), which differ in both their chemical composition and physical role. Primary cell walls are found around cells that are elongating and dividing and contain mostly polysaccharides (40-60 % cellulose,

20-40 % hemicellulose and 20-30 % pectin depending on the plant species) (Wang et al., 2017). In some cell types a secondary cell wall is placed once cell expansion is complete (after the primary cell wall), this offers the cell greater mechanical strength. These secondary cell walls are positioned to the interior of the primary cell wall, they are comprised of a cross-linked matrix of mainly lignin, cellulose and hemicellulose (the exact amount of each of the polymers varies depending on the type of plant) (Isikgor and Becer, 2015; Bajpai, 2016). This thick layer stabilises the structure of the plant and gives it strength allowing for resistance against degradation (Naik et al., 2010; Alalwan et al., 2019; Woiciechowski et al., 2020).

1.3.1 Cellulose

Cellulose makes up the main component of plant cell walls, 40 – 60 % of typical lignocellulosic biomass on a dry weight basis (Wang et al., 2017). Cellulose is a polysaccharide composed of linear glucan chains, covalently bonded into a ridged unbranched polymer by β-1,4-glycosidic bonds. This configuration results in stable straight chains and when multiple of these glucan chains align side by side, they form cellulose microfibrils. These straight unbranched cellulose microfibrils are able to interact with one another and are held together by intramolecular hydrogen bonds and intermolecular van der Waals forces (Bai et al., 2019; Satar et al., 2019). Sequential glucose residues along chains in crystalline are rotated 180° to one another meaning the disaccharide cellobiose, is the repeating unit, unlike other polymers of glucan where the disaccharide would be glucose (Ihsan, 2017; Bai et al., 2019; Satar et al., 2019). This results in highly insoluble crystalline structure, rendering the majority of the cell wall cellulose as inaccessible and therefore resistant to microorganisms and enzymatic saccharification (Isikgor et al., 2015; Bai et al., 2019; Satar et al., 2019). Cellulose will usually contain two types of states, well-ordered crystalline regions and disordered regions, formally known as amorphous regions (Ihsan., 2017; Satar et al., 2019). Typically, enzymes can easily digest these amorphous regions and they are hypothesised to be the areas that form the link between cellulose and hemicellulose (Isikgor et al., 2015; Sun et al., 2016; Satar et al., 2019).

1.3.2 Hemicellulose

Hemicellulose represents the third most abundant polysaccharide in plants, between 20 – 40 % of the dry weight plant biomass is made up of hemicelluloses, this could provide a great source of fermentable sugars for use in industry (Houfani et al., 2020; Woiciechowski et al., 2020). Hemicelluloses present a very large, very diverse group of

polysaccharides that are found in both the primary and secondary plant cell walls. Whereas cellulose is derived exclusively from glucose, hemicelluloses are built up by other monosaccharides such as xylose and arabinose (pentoses), or mannose, glucose, and galactose (hexoses) as well as sugar acids (Zoghlami and Paës, 2019). It is these β-1,4 linked subunits that make the polysaccharide backbones (xylan, glucan, mannans) that make up hemicellulose. Notably this also differs from cellulose by containing side chains and branching structures as well as other modifications which as a result prevent the formation of crystalline structures (Sharma et al., 2019; Zoghlami and Paës, 2019). Instead, the nature and structure of hemicellulose means the single chains interact and wrap around the surface of the well-ordered crystalline regions. This in turn means the interaction with cellulose occurs freely through hydrogen and covalent bonds (Sharma et al., 2019). As a result, the lignocellulosic matrix is provided with stability and flexibility, however, this varies greatly according to the plant species (Woiciechowski et al., 2020).

Hemicellulose is more susceptible to degradation and can be enzymatically broken down a consequence of its amorphous and branched structure (Woiciechowski et al., 2020). Once broken down into its monosaccharides, these sugars are then able to be fermented into ethanol. For the effective degradation of hemicelluloses multiple classes of enzymes are necessary (Houfani et al., 2020). Enzymes like but not limited to, glycoside hydrolases, carbohydrate esterases, polysaccharide lyases and endo-hemicellulases contribute to the breakdown of hemicellulose by the collective actions which hydrolyse glycosidic bonds, ester bonds and remove side chains (Houfani et al., 2020). These include α-L-arabinofuranosidase, acetylxylan esterase endo-1,4-β-xylanase, β,-xylosidase, β-mannanase, β-mannosidase, (Zoghlami and Paës, 2019; Houfani et al., 2020)

1.3.3 Lignin

Lignin is a complex polymer that makes up a part of the secondary cell wall, along with both cellulose and hemicellulose that are entwined into this polymer. After cellulose, lignin is the second most abundant naturally occurring polymer, corresponding to 15–40% of dry lignocellulosic biomass weight (Hassan et al., 2018; Bajwa et al., 2019). It is hypothesised that lignin will play an important role in future and efficacy of the production of biofuel as a raw material and potential for untapped fermentable sugars (Brosse et al., 2011; Li and Zheng, 2020). Lignin is an amorphous phenolic polymer, comprised mostly of three aromatic alcohol monomers (*p*-coumaryl, coniferyl, and sinapyl) depending on the plant species and tissue the amounts these vary (Alalwan et al., 2019, Houfani et al., 2020). As these monomers are combined into a lignin molecule they form units, these

monomers become known as, p - hydroxyphenyl (H), guaiacyl (G) and syringyl (S) retrospectively. The number of these units vary between different plant species as when building their lignin's, they will use different proportions of each unit (Zoghlami and Paës, 2019). Furthermore, lignin is responsible for binding hemicelluloses to cellulose within the cell wall through covalent, ester and hydrogen bonds, giving the whole lignocellulosic network more rigidity.

Lignin acts as a crucial barrier within the breakdown of plant biomass, within its structure the lack of repetitive pattern gives both rigour and strength to the internal cell wall (Woiciechowski et al., 2020). Additionally, a hydrophobic coat is formed surrounding the polysaccharides due to lignin's aromatic nature. For these reasons microorganisms and enzymes have difficulty in being able to directly degrade lignin and so lignin protects the plant (Hassan et al., 2018; Bajwa et al., 2019).

1.4 CAZymes

As a result of the complex structure of lignocellulosic biomass the enzymatic deconstruction is achieved though the combination of several carbohydrate-active enzymes (CAZymes). Typically, various CAZymes will act together synergistically in order to breakdown the polysaccharides into monosaccharides (Cantarel et al., 2009). The CAZy database is the up-to-date collection of enzymes that act on glycosidic bonds by either creating, modifying or degrading them (http://www.cazy.org/; Lombard et al., 2014). These enzymes are classified into families based on the sequence similarity of their amino acids, linking specificity and structurally related enzymes together (Henrissat, 1991; Lombard et al., 2014). The families can also be used where an unidentified protein has high sequence similarity to a known experimentally characterised protein within a family to conservatively classify a putative function (Cantarel et al., 2009).

1.4.1 CAZyme families

At present, the CAZy database covers 5 modules containing over 350 protein families for enzymes associated to catalysing the synthesis, modification or breakdown of carbohydrates and glycoconjugates (Lombard et al., 2014). Glycoside hydrolases (GHs), containing 171 protein families, these enzymes either hydrolyse or catalyse trans-glycosylation reaction of glycosidic bonds. These GH enzymes are widespread and observed in most genomes they are important enzymes in the breakdown of cellulose and hemicellulose. For these reasons they are important for biotechnological applications and are thus far the most biochemically characterised set of enzymes in the

database (Lombard et al., 2014). Next Glycosyltransferases (GTs), with 114 families contain the enzymes that by using phospho-activated sugar donors synthesise glycosidic bonds (Lairson et al., 2008; Lombard et al., 2014). Polysaccharide lyases (PLs) are presently found in 42 families. This class of enzymes act on certain activated glycosidic bonds in acid-containing polysaccharides by cleaving linkages by a non-hydrolytic mechanism (β-elimination) (Lombard et al., 2014; Chakraborty et al., 2017). Multiple PLs have been shown to have applications in the biotechnological and biomedical sectors, such as applications in the food processing and textile industries (Chakraborty et al., 2017). Carbohydrate esterases (CEs) are presently classified from the CAZy database into 19 families. CEs represent a class of esterases that hydrolytically removed ester-based modifications from carbohydrates by catalysing the de-O or de-N-acylation of mono-, oligo- and polysaccharides (Lombard et al., 2014; Nakamura et al., 2017).

Auxiliary activities (AAs) at present contains 9 families of ligninolytic enzymes and eight families of lignin degradation enzymes such as lytic polysaccharide mono-oxygenases (LPMOs), giving 17 total families in this class (Levasseur et al., 2013; Lombard et al., 2014). The AAs contain redox-active enzymes, the different families contain catalytic enzymes that have been shown to be involved in plant biomass degradation (Lombard et al., 2014). Enzymes that break down lignin may not exclusively act on carbohydrates, however, as lignin is habitually found together and closely associated with carbohydrates within the secondary cell wall in plants, the lignin acting enzymes do allow and assist the other CAZymes by allowing them to gain access to the carbohydrates contained within the secondary plant cell wall (such as GHs, PLs and CEs) (Levasseur et al., 2013; Lombard et al., 2014).

Furthermore, carbohydrate binding modules (CBMs) are an additional module, which are presently classified into 88 families (Lombard et al., 2014). CBMs do not intrinsically exhibit catalytic activity but are assigned though amino acid sequences that have carbohydrate-binding activity within a CAZyme. Most often CBMs are associated to other CAZyme catalytic modules in the same polypeptide, they generally bind to carbohydrate ligands and boost the catalytic efficiency of other CAZymes (Shoseyov et al., 2006). Functionally, this catalytic improvement is implemented by inducing changes in the shape of the polysaccharide chains and by changing the position of the substrate, moving it closer to the site of catalytic domain (Armenta et al., 2017). Hence it has been shown that degradation of substrates is more efficient with enzymatic complexes bearing CBMs (Shoseyov et al., 2006).

1.5 Adaptation to extreme environments

The exact definition of an extreme environment can be quite complex. There are a high number of variables involved, and it is dependent on whether the measure of 'extremeness' comes from an objective position. Generally, an extreme environment is described as any environment in which the conditions are considered significantly difficult for the majority of life forms to survive in (Merino et al., 2019). Typically, there are two types of extremes: physical extremes (pressure, temperature, radiation) and geochemical extremes (pH, desiccation, salinity) (Rothschild and Mancinelli, 2001). Extremophile or polyextremophile (more than one extreme), is the label given to organisms able to thrive in an extreme environment. Most commonly, the label extremophile is associated with prokaryotes, however extremophile archaea and eukaryotes also exist. In order to survive in harsh conditions, extremophiles have evolved numerous strategies and mechanisms. Thus, unsurprisingly the study of these organisms has led to the discovery of thousands of novel enzymes now being used in biotechnology and industry (Eichler, 2001; Raddadi et al., 2015; Coker, 2016; Dumorne et al., 2017). In addition to this, information has been gained about the function of important proteins under stressful conditions and the potential of extra-terrestrial life (Rothschild and Mancinelli, 2001; Van Den Burg, 2003; Laksanalamai and Robb 2004; Nicolaus 2010; Gabani and Singh, 2013).

1.5.1 pH

pH, defined as $-\log_{10}[H^+]$, determines the availability of inorganic ions and metabolites making it arguably one of the most essential factors that affects an organism's life (Oarga 2009). Highly acidic or highly alkaline conditions pose problems for the average organism; most species only survive in the middle range of the pH spectrum. To ensure that all metabolic activities of the cell can remain active, microorganisms are required to maintain a cytoplasmic pH near neutral ($5 \leq pH \leq 8.5$) (Oarga, 2009; Krulwich et al., 2011; Jin and Kirk, 2018; Merino et al., 2019).

When pH values are extremely low, the majority of organisms cannot survive, due to denaturing of their proteins making life impossible (Yang and Honig, 1993; De Oliveria and Martiinez, 2020). However, some organism can survive in low pH environments such as hot springs (e.g. Yellow stone national park) and anthropogenic waste sites (acid mine waste) (Skorupa et al., 2013; Simate and Nflovu, 2014). These organisms are called acidophiles, made up of mainly archaea and bacteria with fewer eukaryotes (Rampelotto, 2013). These organisms evolved proton pumps capable of removing excess internal

protons, therefore maintaining the important neutral intercellular pH (Kristjansson and Hreggyidsson, 1995; Serrano et al., 2004; Dhakar and Pandey, 2016).

Conversely, alkaliphiles favour high pH (pH >8.5), e.g alkaline soda lakes, and will outcompete rivals and succeed in these environments (Horikoshi, 2016). For prokaryotes respiring aerobically with a membrane-bound ATP synthase the limited protons available creates challenging conditions (Krulwich et al., 1998; Rothschild and Mancinelli, 2001).

1.5.2 Temperature

There is a wide range of temperatures recorded on the Earth's surface, from deep-sea hydrothermal vents (up to 495°C) (McDermott et al., 2018) to East Antarctica (reaching -110.9°C) (Zhao et al., 2021). Without the influence of geothermal activity (hydrothermal or magmatic) or the influence of high pressure (or a combination), the highest reported land temperature is 80.8 °C, in the Lut Desert, Iran (Zhao et al., 2021). Extreme temperatures can create a variety of challenges for organisms. In low temperatures, organisms can suffer a complete reduction of metabolism, or fatal disruption to cellular structures due to the formation of ice crystals. In high temperatures, organisms can suffer lethal dehydration and the pushing of metabolic processes to their limit (Merino et al., 2019; Zhao et al., 2021). As temperatures approach 100 °C, proteins and nucleic acids denature (Rothschild and Mancinelli, 2001) and chlorophyll degradation (75 °C+) prevents organisms from photosynthesising (Rothschild and Mancinelli, 2001). Despite this life is still found spanning a wide range of temperatures, as to date life has been found everywhere where water is in its liquid state.

Cold adapted organisms known as psychrophiles, have tailored themselves to survive in low temperature environments such as deep sea, permafrost, ice lakes and glaciers (Hamdan, 2018). These harsh environments have successfully been colonised by diverse communities of archaea, bacteria, insects, algae and fish that are able to thrive at these sub-zero temperatures (Siddiqui et al., 2013; Bhatia et al., 2021). In order to overcome limitations imposed by low temperatures and to maintain metabolic activity these organisms have adapted and developed necessary mechanisms and implemented crucial changes to their cellular structures and functional organisation (Bhatia et al., 2021).

Organisms adapted to succeed in higher temperatures are known as thermophiles (growth temperature ~70 °C) or hyperthermophiles (up to 110 °C) (Bala and Singh, 2019). There are many thermophiles among prokaryotic groups, such as bacteria and

archaea and some in eukaryotic microorganisms, such as protozoa, algae and filamentous fungi (Bala and Singh, 2019). These organisms can produce enzymes able to function at extreme high temperatures. Hyperthermophiles, consists mostly of archaea and bacteria, and can produce enzymes with temperature optimums of up to 142 °C, such as amylopullulanase, which far exceeds the optimal living temperature of the organism itself (Schuliger et al., 1993).

In contrast to low temperatures, eukaryotes do not survive high temperatures. Due to the limited membrane adaptation ~56 °C is the upper limit for eukaryotes survival (Zeldes et al., 2015). However, in some cases thermophile eukaryotes can produce stable proteins functioning at temperatures, higher than the survival temperature of the organism. It has been shown that extremophile eukaryote red alga *Cyanidium caldarium* produces a temperature resistant C-phycocyanin that is more stable than its mesophilic counterpart (Kao et al., 1975; Eisele et al., 2000). Additionally, another extremophilic eukaryote red alga *Galdieria sulphuraria* has a secreted peroxidase, shown to exhibit 100 % activity at 60 °C, reduced to 50 % activity at 80 °C (Oesterhelt et al., 2008). Therefore, investigations into extremophile eukaryotes can be useful for uncovering heat stable proteins.

1.5.3 Pressure

Environments with an extreme high or low pressure present challenges to survival because it forces volume changes, this can change the fluidity of cellular membranes namely though compression of lipids (Rothschild and Mancinelli, 2001). An increase in pressure can directly inhibit an increase in volume as a result of a chemical reaction. Numerous organisms have adapted to very high-pressure environments, though rapid changes in pressure can still be harmful. Organisms living under these high-pressure conditions have adapted through various strategies, e.g. their cell membranes contain an increased number of unsaturated fatty acids, increasing the membrane fluidity at higher pressures (Merino et al., 2019).

In contrast low-pressure environments, for example the high altitude in mountains, are less likely to affect microbial survival in the same way as high pressure does. The vacuum in space has the lowest pressure possible where even gravitational effects are reduced, some organisms have been shown to survive exposure to space conditions ranging from months to years. Examples of these are several prokaryotes, fungi, and lichen (Onofri et al., 2018; Yamagishi et al., 2018; Merino et al., 2019). In this atmosphere, it is more likely that cosmic radiation, low temperatures and desiccation

would have a more influential role in the microbial diversity than decreasing pressure (DasSarma and DasSarma, 2018; Merino et al., 2019). It is hypothesised, sporulation, resting stages and the formation of biofilms can have an effect. It is proposed that the top layer of the biofilm is exposed to conditions and thus protects the inner layers, therefore facilitating the survival of microorganisms under space conditions (Delort et al., 2010; Frösler et al., 2017; Merino et al., 2019).

1.5.4 Radiation

An important parameter known to affect mutagenic events is radiation. This is defined as the emission or transmission of energy either as electromagnetic waves (such as X-rays, gamma rays, visible light, ultraviolet (UV) radiation, infrared radiation, microwaves or radio waves) or as particles (such as alpha particles, heavy ions, electrons, neutrons and protons) (Rothschild and Mancinelli, 2001; Oarga, 2009). Different types of radiation have a variety of effects on organisms ranging in severity. Less severe effects include a reduction of motility, more severe include inhibition of photosynthesis, and in extreme cases, the mutation of nucleic acids. Extreme damage could lead to modified bases along with strand breakages through direct damage to the DNA. Alternatively, indirect damage can occur when reactive oxygen species are produced, resulting in structural changes. (Rothschild and Mancinelli, 2001)**.**

Radiation can affect all ecosystems, therefore the development of suitable resistance to ionising radiation and UV radiation has been necessary for numerous organisms (Merino et al., 2019). Adaptations for microorganisms are reported to cause changes to DNA repair functions and increase in genome copies for genome redundancy among others (Byrne et al., 2014; Merino et al., 2019).

1.5.5 Desiccation

Water is essential for basic metabolic processes and therefore for life. Environments lacking in water are considered extreme, and organisms able to survive air-drying close to absolute dehydration are described as having desiccation tolerance (Billi and Potts, 2002; Merino et al., 2019). Occurring very early on in the evolution of terrestrial life, desiccation tolerance is commonly observed in multiple cyanobacteria and green algae (Holzinger and Karsten, 2013; Singh, 2018; Oliver et al., 2020) among others. Removing water through air drying can have severe consequences including damage to proteins, nucleic acids and membranes, and more severely, organism death (Billi and Potts, 2002).

Anhydrobiosis is a survival mechanism which can be employed by organisms experiencing extreme desiccation. The organism will enter a state of suspended animation, depicted by no metabolic activity and little intracellular water (Rothschild and Mancinelli, 2001; Oarga, 2009). During anhydrobiosis cellular death can occur, often due to denaturation of proteins and nucleic acids, structural breakages and accumulation of reactive oxygen species, therefore anhydrobiosis is not always viable. A multitude of organisms such as bacteria, yeast, fungi, plants, insects, tardigrades, microphagous nematodes show examples where they can become anhydrobiotic (Rothschild and Mancinelli, 2001). Key survival factors during desiccation are cellular recovery and cellular protection (Merino et al., 2019).

1.5.6 Salinity

Levels of salinity vary across environments. In marine environments, salinity measures 3-4%,10.5 % in hot springs and up to 37.1 % in soda lakes (Last, 2002; Mamayey, 2012; Merino et al., 2019). Organisms live in varying degrees of salinity, ranging from distilled water to saturated salt solutions. Protein biosynthesis, the uptake of nutrients and enzymatic reactions are highly influenced by salinity in the environment (Oarga, 2009; Telesh et al., 2013; Gunde-Cimerman, 2018; Oren, 2020). For an organism's survival in high salt concentrations, adaptations to the osmotic alterations around them are required (Gunde-Cimerman, 2018).

Organisms able to survive in environments characterised by high salinity (hypersaline) are described as halophiles, and include archaea, bacteria, and eukaryotes (Gunde-Cimerman, 2018; Oren, 2020). Sodium ions are essential for the growth and metabolism of halophilic organisms; therefore, a high salt concentration is required for survival (Rothschild and Mancinelli, 2001; Telesh et al., 2013). The osmotic potential experienced in hypersaline environments causes challenging effects such as cellular dehydration, loss of turgor pressure, and desiccation. Therefore, it is essential halophiles are able to withstand extreme osmotic stress (Gunde-Cimerman, 2018; Merino et al., 2019; Oren, 2020). Hypersaline tolerance developed as novel traits, evolving from organisms able to survive these environments (Madern et al., 2000; Edbeib et al., 2016).

The production of organic solutes known as osmoprotectants (e.g polyols, amino acids, sugars, and betaines) allows many microorganisms to tolerate a wide range of salt concentrations Termed the salt-out strategy, the production of osmoprotectants counteracts the concentration of salts, by way of expulsion (Oren, 2011; Edbeib et al., 2016). Other adaptations to aid survival in high salt environments include the

accumulation of inorganic ions intracellularly. Known as the salt-in strategy where unique transporter pumps allow for accumulation of salt in the cytoplasm to create a state of equilibrium between the inside and outside of cells, therefore, eliminating the osmotic gradient (Glenn et al., 1999; Oren, 2011; Edbeib et al., 2016).

Without appropriate adaptations to allow survival in these conditions, organisms would likely go through osmotic stress caused by a change in the solute concentration around them. Increased salt would create osmotic potential and water would be drawn from cells via osmosis, putting the cells into a state of 'shock,' preventing usual function and eventually leading to cell death (Oren, 2011; Edbeib et al., 2016; Gunde-Cimerman et al., 2018).

1.5.7 Oxygen

Throughout the majority of the Earth's existence, it has been an anaerobic environment (Weber, 2006). At present day organisms inhabit environments both anaerobic and aerobic (Rothschild and Mancinelli, 2001). Anaerobic metabolism is considerably more inefficient than aerobic, nevertheless the exploitation of aerobic respiration does have its costs (Oarga, 2009).

Reactive forms of oxygen are recognised as superoxide radicals, the hydroxyl radical (OH) and hydrogen peroxide (H_20_2) and singlet oxygen (O_2) (Mallick and Mohn, 2000; Nosaka and Nosaka, 2017). They can be produced photochemically as a result of UV-A radiation (320–400 nm) such as H_2O_2 within cells resulting from photosynthesis and can be formed also during mitochondrial respiration, during production of uric acid and due to the cytochrome P450 metabolism of hydroperoxides in eukaryotic cells (Oarga, 2009).

Oxidative damage resulting from any of these reactive oxygen molecules is extremely serious and can present significant danger to cells. This can affect organisms in many ways, from physiological changes such as ageing, through to cancer development (Rothschild and Mancinelli, 2001; Oarga, 2009). Functionally the reactive oxygen species can interact with certain biomolecules by modifying or completely inactivating their biochemical activities (Nosaka and Nosaka, 2017). Despite this there are environmental conditions where organisms are found to have adapted and survive fatally low oxygen concentrations (Oarga, 2009).

1.6 Acidic hot springs and polyextremophile algae

Acidic hot springs are the result of secondary volcanic activity, they are produced as a result of geothermally heated groundwater emerging onto the surface of the Earth. The groundwater is typically heated by small bodies of magma cooling from contact with water infiltrated deep within the Earth's crust (Fouke, 2011). Due to the surfacing of these waters through the layers of the earth these hot springs will often contain dissolved minerals in high quantities. There is a high variability in the chemistry of hot springs. With acidic springs that are dominated by sulphates the pH can reach as low as 0 (Brock, 1978).

Typically identified by hot sulphureous mines, fumaroles, hot muds, and geysers (Gonsior et al., 2018) (Figure 1.1). Acidic hot springs have a presence across the globe, however, only appear at particular geological niches. Yellowstone National Park in USA is possibly the most well-studied and well-known geothermal area (Brock, 1978; Brock, 2001; Toplin et al., 2008; Skorupa et al., 2013) though others include Iceland (Claudia Ciniglia et al., 2014), Japan (Toplin et al., 2008), Russia (Sentsova, 1991), New Zealand (Toplin et al., 2008), Italy (Yoon et al., 2004), Turkey (Iovinella et al., 2018, 2020) and Taiwan (Hsieh et al., 2015).

Different organisms dominate at these pH and temperature extremes. These acidic hot springs are home to a variety of thermophilic and acidophilic organisms (Gonsior et al., 2018). These organisms are subject to more than one of the features described above concurrently and thus are classified as polyextremophiles (Seckbach and Rampelotto, 2015). Extraordinarily polyextremophiles are prospering in environments previously thought inhospitable to life. They have adapted and evolved to not just survive but to in fact thrive and dominate these extreme environments by being permanently exposed to these harsh conditions (Seckbach and Rampelotto, 2015). If extra-terrestrial life exists is has been hypothesised and generally accepted that it would be in the form of an extremophile (Seckbach and Chapman, 2010; Lage et al., 2012).

Figure 1.1: Examples of hot springs in the Phlegraean Fields, Italy (left; Seth Davis 2016), Yellowstone National Park, USA (middle; National Geographic, 2019) and Reykjavik, Iceland (right; Iovinella, 2018).

It is for these reasons that the ecological study of acidic hot springs and even more the organisms inhabiting this environment is of interest, to better understand how organisms have the capacity to withstand more than one harsh environment (Dhakar and Pandey 2016; Dodds and Whiles, 2018). These types of hot spring environments are usually fatal to most eukaryotes so there are limited group that can tolerate such extremes, alternatively prokaryotes have been shown to host a diverse group of organisms that are able to survive in such conditions (Dodds and Whiles, 2018). In these environments, there are examples of eukaryotic microalgae living at the limit of their potentiality, adapting their metabolism and biological processes to this extreme life (Seckbach and Rampelotto, 2015**;** Dodds and Whiles, 2018). These microalgae are considered important, as not only are they both thermophiles and acidophiles, but they also have higher photosynthetic abilities than that found in terrestrial plants and thus are themselves a renewable resource (Rodolfi et al., 2009; Katayama et al., 2020).

Currently there are five genera of known acidophilic microalgae, Dunaliella (usually known for its ability to thrive in hypersaline environments) has only one species *Dunaliella acidophila* (Gimmler and Weis, 1992). Similarly, another genus *Coccomyxa* (Fuentes et al*.,* 2016; Navarro et al., 2017). The remaining three genera belong to the class of *Cyanidiophyceae Cyanidioschyzon*, *Cyanidium* and *Galdieria* (Yoon et al., 2006; Varshney et al., 2015).

1.6.1 Cyanidiophyceae

It has been estimated that red algae (Rhodophyta) diverged into seven major lineages. The earliest divergence is shown to be the Cyanidiophyceae class (Pinto et al., 2003; Ciniglia et al., 2004; Yoon et al., 2006). This divergence has been calculated at approximately 1.3 billion years ago and is separated from the remainder of the red algal lineages. Members of Cyanidiophyceae are unicellular microalgae that are typically found colonising acidic (pH 0-4) and thermal (25-56°C) sites world-wide (Ciniglia et al., 2004; Yoon et al., 2006; Yang et al., 2016). The species within this class can be difficult to distinguish via morphological and physiological traits due to their simple morphology structure showing few diagnostic features (Ciniglia et al., 2014). Phylogenetic analysis using the plastid encoded rbcL (ribulose-1,5-bisphosphate carboxylase/oxygenase) gene was used to establish three genera, *Cyanidium* (the only clade containing a mesophilic species), *Galdieria* and *Cyanioschyzon*. These were identified as containing eight species in total (*C. chilense, C. caldarium, G. sulphuraria, G. daedala, G. partita, G. phlegrea, G. maxima* and *C. merolae*) (Sentsova 1991; Albertano et al., 2000; Pinto et al., 2003; Ciniglia et al., 2004; Yoon et al., 2006; Toplin et al., 2008).

Cyanidium is the genus comprised of two species, *C. caldarium* first described by Tilden 1898 found in thermal acidic areas in Yellowstone National Park is a polyextremophilic species. The second species is *C. chilense* (Hoffmann, 1994) the only mesophilic species in this class is found in caves around pH 7 and temperatures between 20-25°C (Ciniglia et al., 2019). Morphologically the key identifiable features are, round shaped cells between 2-6 µm in size, no vacuole present and typically only one mitochondrion (Merola et al., 1981). These cells divide asexually via endospores, the usual pigments found within the cells are allophycocyanin, chlorophyll a, carotenoids and C-phycocyanin (Allen, 1959). The genus *Cyanidioschyzon* only contains one species, *C. merolae*, this alga was isolated from sulfuric hot spring water in Italy (De Luca et al., 1978). It is smaller compared to *Cyanidium* with a typical cell diameter of 2-3 µm, additionally it displays simple cellular architecture containing just one chloroplast and one mitochondrion per cell along with the same pigments observed in Cyanidium (De Luca et al., 1978; Suzuki et al., 1994; Toda, 1995). A notable feature of this species compared to the rest of the Cyanidiophyceae genus is its lack of a ridged cell wall. Additionally, its mode of cellular division also differs, whereby it will undergo binary fission instead of using endospores (Suzuki et al., 1994; Toda et al., 1995; Nishida et al., 2005).

The *Galdieria* genus before 1981 was referred to under the *Cyanidium* genus due to similar morphological traits (Merola, et al., 1981). For this reason, numerous studies

carried out on *G. sulphuraria* were attributed to *C. caldarium* (Seckbach, 1991; Albertanol et al., 2000). The originally defined *G. sulphuraria* was a spherical cell between 3-11 µm, it reproduces via 3-11 endospores (as *C. caldarium*), it contains one mitochondrion and one chloroplast inside its cell wall. Differently to *Cyanidium* and *Cyanidiophyceae* there is the presence of a vacuole (De Luca et al., 1978; Merola et al., 1981). Based on key morphological characteristics such as the number of endospores and cell size but namely the shape and number of plastids during the cell cycle allowed for the classification of three further species, *G. partita, G. daedala, G. maxima* (Sentsova, 1991). Development of molecular tools lead to the establishment of the fifth *Galdieria* species, *G. phlegrea* originating Pisciarelli in the Phlegraean Fields in Italy (Ciniglia et al., 2004; Pinto et al., 2007; Qiu et al 2013). Recent comparative genome analysis also revealed a new genus with in the *Cyanidiophyceae* class called *Cyanidiocoocus*, this new genus contains one species, *Cyanidiococcus yangmingshanensis* (Liu et al., 2020).

1.6.2 Growth capacity of *G. sulphuraria*

G. sulphuraria unlike many other eukaryotic organisms can grow in autotrophic conditions through photosynthesis, in heterotrophic conditions utilising multiple carbon sources and in mixotrophic conditions, a mixture of the two (Barbier et al., 2005; Sloth et al., 2006; Curien et al., 2021). As shown in Figure 1.2 adapted from Schönknecht et al., 2013, when grown autotrophically cells are a deep green colour due to the chlorophyll production via photosynthesis. Once switched to heterotrophic growth, in this case 200 mM glucose, cells become more yellow in colour. Many of this species have often been found inhabiting endolithic areas, by growing both on rocks and soil as well as forming endolithic algal mats where 0.1-1% of sunlight penetrated (Gross et al., 1998; Gross and Oesterhelt, 1999; Yoon et al., 2006). A step further than that G. sulphuraria is also a cryptoendolithic, by colonising empty pores inside a rock (Gross et al., 1998). It is these habitats where the availability of light is minimal that heterotrophy is vital for the alga's survival (Gross et al., 1998; Gross and Oesterhelt, 1999; Yoon et al., 2006; Oesterhelt et al., 2007). Often in these habitats the biggest substrate available will be themselves and thus they often will be using consumption of their own cell walls as an energy source (Gross et al., 1998; Jain et al., 2015).

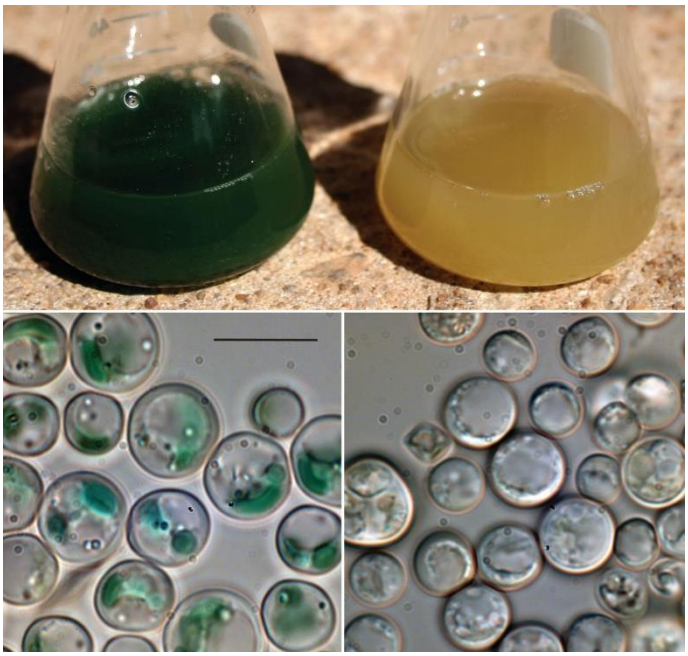


Figure 1.2: Adapted from Schonknecht et al., 2013 *G. sulphuraria* cells grow in photoautotrophic (constant light) (left) and heterotrophic (constant darkness, 200 mM glucose) (right) conditions. Bar shown in light microscope represents 10 µm.

G. sulphuraria has a vast array of metabolic properties that allow it to grow vigorously on a wide range of carbon sources as well as displaying high tolerance to heavy metals (Gross and Oesterhelt; 1999, Jain et al., 2014). Currently, over 50 have been identified as supporting growth, including several sugars, sugar alcohols, amino acids, and organic acids (Rigano et al., 1976; Rigano et al., 1977; Gross and Schanarrenbergeer, 1995; Oesterhelt et al., 1999; Oesterhelt and Gross; 2002; Qiu *et al.,* 2013). This is achieved through a complex uptake system for polyols and sugars, consisting of at least 14 transporters (Oesterhelt and Gross; 2002; Oesterhelt et al., 1999). These transporters have been shown not to act in a way of one substrate per transporter but rather all transporters for the same sugar uptake are induced all together under heterotrophic conditions (Oesterhelt et al., 1999; Oesterhelt and Gross, 2002; Barbier et al., 2005). It is the sugar sensing mechanism that is crucial, whereby the availability of sugars up-regulates heterotrophic metabolism and in turn will down-regulate photosynthesis (Oesterhelt et al., 1999; Oesterhelt and Gross, 2002; Qiu et al., 2013; Curien et al., 2021). In an energy saving attempt *G. sulphuraria* will repress transporters that require more energy if a lower cost alternative is available. For example, when glucose is present the polyol and deoxy sugar transporters are repressed, this tactic ensures the most

efficient energy source is used at any moment in time (Oesterhelt et al., 1999; Oesterhelt and Gross; 2002; Barbier et al., 2005).

As the uptake systems reported are for simpler low molecular weight sugars and polyols it is proposed that for *G. sulphuraria* to display such diverse growth capacities using complex carbohydrates they likely produce extracellular enzymes to breakdown and utilise polysaccharides (Gross and Schnarrenberger, 1995; Oesterhelt et al., 1999; Gross, 2000). *G. sulphuraria* has already been shown to secrete proteins that are acid and heat resistant (Oesterhelt et al., 2008). Additionally, there has already been progress in exploiting *G. sulphuraria* for its heterotrophic metabolic abilities, using cultivations for removal of products from problematic industrial waste streams, these include food production, lactose and agricultural (e.g. unrefined biodiesel-derived glycerol and lignocellulosic biomass) waste-streams (Mulder, 2016; Rahman et al., 2020; Pleissner et al., 2021; Scherhag and Ackermann, 2021; Somers et al., 2021;). The benefits of harnessing *G. sulphuraria* in this way can lead to the development of high value products that will be highly stable (temperature and acid tolerant) including, pigment nutraceuticals (phycocyanin) (Schmidt et al., 2005; Sloth et al., 2006; Graverholt and Eriksen, 2007; Burns, 2020). This along with the collection of biochemical versatility within the species should reveal a large repertoire of metabolic enzymes, which are a rich source of thermo-stable and acid tolerant proteins for industrial biotechnology applications.

1.6.3 Phylogeny of Cyanidiophyceae

As previously discussed, the morphology differences with the Cyanidiophyceae class are slight and even more so within each genus, so much so that early studies often misclassified *G. sulphuraria* (Merola, et al., 1981). The diagnostic criteria used to distinguish between species are so similar that establishing a new species or genus is challenging (Merola et al., 1981; Sentsova 1991; Albertano et al., 2000; Pinto et al., 2003). Taking into account the evolutional history of this class (~1.3 billion years removed) it is unusual that so few distinguished species have survived and in reality, it is much more likely that the number of species has been miscalculated (Pinto et al., 2003; Ciniglia et al., 2004; Yoon et al., 2006). The huge biodiversity within this class based on rbcL plastid gene is well established, this and numerous studies on different geothermal sites across the globe has highlighted that there are more linages present than identified by morphological and physiological tools (Ciniglia et al., 2004; Toplin et al., 2008; Hsieh et al., 2015). The most recent sequenced based phylogenetic analysis performed on Cyanidiophyceae populations using the partial rbcL plastid gene revealed the genetic structure of the genus *Galdieria* in particular *G. sulphuraria* and *G. maxima*

to be much more complicated that preciously understood especially from such an ancient unicellular red alga (Iovinella et al., 2018).

1.7 Aims

The overall aim of my thesis is to evaluate the potential of *G. sulphuraria* species for biotechnical applications, namely in use for lignocellulosic biomass degradation for application in production of biofuel. Previous studies of *G. sulphuraria* highlighted a complex genetic structure and taxonomy as would be expected of an ancient microorganism. Even though the *G. sulphuraria* genome is small it contains a wide range of enzymes that allow for its success in surviving such extreme and harsh environments. The evolutionary path of individual genes will not necessarily represent the same evolution as the species as a whole. Therefore, the first aim of my thesis was to evaluate and resolve the molecular evolution of the nuclear phylogeny within the *G. sulphuraria* species. I achieved this by developing DNA extraction protocols and sequencing genomes from a range of strains found worldwide (Chapter 2). Furthering this, the aim was to identify any genes potentially involved in the degradation of lignocellulosic material that were under adaptive evolution, thus being key to *G. sulphuraria's* survival. After the identification of six lineages, I created transcriptomic and long read sequencing data that was used to complete annotations of each of the lineages, with the aim of being able to characterise and predict any CAZymes present within *G. sulphuraria* genomes (Chapter 3). Furthering the knowledge of how *G. sulphuraria* has such ability to grow on diverse substrates, experimental data was collecting using the growth of each lineage on different carbohydrates to identify potential enzymes. Growth on hemicellulose of one strain was used to identify secreted enzymes (Chapter 4). Previous chapters revealed a list of potential industrially relevant enzymes that could be involved in the degradation of lignocellulosic material. The aim for the next step in identifying and characterising these putative enzymes was to produce purified recombinant protein that could be used for functional assays (Chapter 5).

Chapter 2 - Nuclear Gene Phylogeny of *G. sulphuraria*

2.1 Introduction

Taxonomists have for many years successfully used morphological traits to determine whether a group of organisms are different populations of the same species or different species altogether for multicellular organisms. Using this species delimitation process to classify microorganisms is more challenging due to the infinitesimal differences in diagnostic characteristics (Zhao et a., 2018). Previously collected strains of *Galdieria* have been classified based on key features including shape, number of plastids, number of endospores, cell size, presence or absence of a cell wall as well as their carbohydrate growth characteristics (Merola et al., 1981; Sentsova, 1991; Albertano et al., 2000; Pinto et al., 2003; Ciniglia et al., 2014). This technique is taxing as often these crucial morphological features are not easily distinguishable from each other due to intra- and interspecific variation arising.

Recently, it was detected that the diversity observed at the molecular, biochemical and physiological level did not match the elementary shape (small round ball) and the simple ultrastructure that characterise *G. sulphuraria* cells (Ciniglia et al., 2004). *Galdieria* is known to have notable metabolic diversity. Growth has been shown to be supported by numerous carbon sources. It can be predicted that such an organism will contain a variety of potentially interesting carbohydrate acting hydrolytic enzymes used for its survival. As discussed in Yoon et al., 2004, the *Galdieria* genus is recognised to have a long evolutionary history with divergent clades. Therefore, it is strange but interesting that this lineage currently has so few recognisable species suggesting that these organisms are more genetically diverse than current estimations predict. For decades research to analyse and describe the diversity and phylogenetic relationships of the organism has previously relied on using the plastid rbcL gene (Freshwater et al., 1994; Ciniglia et al., 2004; Toplin et al., 2008; Hsieh et al., 2015; Iovinella et al., 2020). Previous phylogenetic analysis of nuclear genes has also been described, using individual nuclear genes to gain information on their origin and evolutionary history (Qiu et al., 2018; Eren et al., 2018; Del mondo et al., 2019). The collection of samples from around the world, from differing geothermal locations, has increased the molecular knowledge of this organism and in turn led to the general conclusion that more species have evolved than was previously thought.

The intricate genetic structure of *G. sulphuraria* has been highlighted in the recent phylogenetic analysis. Work based on the partial rbcL gene (ribulose-1,5-bisphosphate carboxylase/oxygenase) revealed high genetic diversity both in haplotype and nucleotide

variability along with indications that the rbcL protein-coding gene has undergone positive selection in order to adapt to its extreme environment. These works have collectively led to the hypothesis that there are diverging clades within the species (Toplin et al., 2008; Hsieh et al., 2015; Iovinella et al., 2018; Han et al., 2021). The geographical position of the populations in the subdivision of the species is reflected through the subgroups. This concludes that the isolation of populations is a result of the surrounding non-appropriate environments coupled with the difficulty of long-distance dispersion. It is likely that some populations could originate through human associated dispersal. It is expected that there would be multiple different strains, species or ecotypes to be uncovered, as a direct result from evolutionary events taking place over thousands or millions of years, by which isolated populations are evolving distantly and could eventually become distinct species (Toplin et al., 2008). Published in 2015 Hsieh et al. collected sequence data and used it to identify so-called “Operational Taxonomic Units” (OTUs), which are interpreted as presumptive species. The study also confirmed the increase in genetic diversity seen within *G. sulphuraria* could be attributed to the involvement of habitat heterogeneity (Hsieh et al., 2015). Genetic variance between subgroups of *G. sulphuraria* populations have been measured by analysing Inter-Populational Pairwise Genetic Differentiation (Iovinella et al., 2018). This analysis showed a high amount of genetic differentiation among populations indicating low levels of gene flow and thus giving rise to diverging and isolated evolution (Iovinella et al., 2018).

Next-generation sequencing (NGS) has revolutionised the analysis of diversity and evolution of microorganisms (Ronaghi et al., 1996; Zhao et al., 2018). NGS technology allows for the production of large datasets accurately and quickly, these in turn can be a used to understand the evolutionary history and infer robust phylogenomic analysis of microorganisms, such as *G. sulphuraria*. Studies of this alga focused on adaptive genomic changes, revealed that the species ability to survive in extreme environments and metabolic flexibility could be attributed to acquiring genes horizontally via various prokaryotes. Horizontal gene transfer (HGT) could be critical to the algae’s tolerance to higher temperatures, higher concentrations of heavy metals, as well as its ability to utilise urea as a nitrogen source and metabolise glycerol as a carbon source (Schönknecht et al., 2013; Qiu et al., 2013; Hsieh et al., 2015). Further sequence data from collections of *Galdieria* strains from multiple geothermal sites across the globe could provide a huge dataset to mine for interesting features, especially with regards to the alga’s metabolic flexibility in utilising numerous carbon sources.

With the production of complete genome sequences, data can be used to learn and understand more about evolution, adaptation and divergent species on a molecular scale. Evolutionary changes occurring in an organism that make it more suitable to living in its environment and thus increasing chances of survival is known as adaptive evolution. It is understood that adaptation to a favourable environment would induce more genetic changes at an amino acid level that would alter the protein sequence (Yang et al., 2000; Rocha et al., 2006; Kosiol et al., 2008; Jeffares et al., 2015; Del Amparo et al., 2021). It is also accepted that adaptive evolution affects non-coding sites in a genome. A fraction of the non-translated sequence, for example, upstream of genes is crucial in the regulation of gene expression and thus changes in these sequences can have influence an organism's fitness (Andolfatto, 2005; Eyre-Walker, 2006; Hittinger and Carroll, 2007; Dong et al., 2018).

When looking at coding sequence it is useful to look at the ratio of the rate of nucleotide changes that alter the amino acid sequence (non-synonymous substitutions; dN) to the rate of nucleotide changes that do not alter the amino-acid sequence (synonymous substitutions; dS), $\omega = dN/dS$. The dN/dS ratio measures the balance of mutations acting on a gene and the type of selection placed on the gene, this being neutral, purifying or positive selection. A ratio of $\omega=1$ indicates neutral evolution, a low ratio $\omega < 1$ signifies strong purifying or negative selection whereas a high ratio $\omega > 1$ indicates positive, adaptive or diversifying selection (Yang, 2000; Jeffares et al., 2015; Del Amparo et al., 2021). The ω ratio is widely used to statistically analyse patterns of selection on a genomic scale of protein-coding genes and summarising the evolutionary rates of genes. It is a helpful measurement in identifying how conserved genes are between species or strains as well as identifying genes that have gone under phases of adaptive evolution (Yang et al., 2000; Rocha et al., 2006; Kosiol et al., 2008; Jeffares et al., 2015; Del Amparo et al., 2021).

2.1.1 Aims

The present study aimed to improve the understanding of the phylogenetic relationship among different *Galdieria* strains, improving the analysis from a gene-level to a pan-genome one. The NGS data from 43 presumed *Galdieria* lines were used to extract the nuclear coding sequences (CDSs), which were concatenated in single alignments and used to infer the pan-genome and following on from this the species nuclear phylogeny. This analysis showed evidence of two species in the *Galdieria* genus (*G. phlegrea* and *G. sulphuraria*). To further understand the evolutionary history of *G. sulphuraria*, an overall and a lineage-level analysis of the synonymous and non-synonymous

substitutions were performed to understand if natural selection forces have been affecting the divergence and the paraphyletic evolution of the species. Then to gain information on genes that may be important in *Galdieria's* adaptation and survival, positive selection analysis was performed. Genes under positive selection were assessed for any links to secreted hydrolases, to gain insight to any important enzymes the algae is likely harbouring for aiding its growth on multiple carbon sources.

2.2 Materials and Methods

2.2.1 Strain isolation

Galdieria strains were obtained from the Algal Collection of University of Naples (www.acuf.net), the Culture Collection of Autotrophic Organisms, the Collection of Microorganisms from Extreme Environments, the Institute of Plant Physiology, Russian Academy of Sciences, the Culture Collection of Algae at Göttingen University, the Tung-Hai Algal Lab Culture Collection. All strains were isolated by streaking them across agar plates, and colonies were inoculated in Allen medium pH 2 (Allen and Stainer, 1968,Table 2.1). These were then cultivated at 37 °C under continuous fluorescent illumination of 45 µmol photons·m−2·s−1 (Supplementary Table 1)

Table 2.1: Composition of Allen medium pH 2 (Allen and Stainer, 1968).

Component	g/L	Oligoelements	mg/L
KH_2PO_4	0.3	$ZnCl_2$	0.014
K_2HPO_4	0.6	$Na_2MoO_4.2H_2O$	0.005
$CaCl_2.2H_2O$	0.02	$CuSO_4.5H_2O$	0.01
NaCl	0.1	$CoCl_2.6H_2O$	0.005
$MgSO_4.7H_2O$	0.3	$MnCl_2.4H_2O$	0.009984
$FeSO_4.7H_2O$	0.00996	H_2SO_4	
$(NH_4)_2SO_4$	1.32		

2.2.2 DNA Extraction, Sequencing and Assembly

DNA was extracted using a mixed SDS-CTAB protocol. Briefly microalgal pellets were harvested by centrifugation at 13200 rpm for 5 minutes, resuspended in 40 µl of PBS pH 7.5 and 500 µl of DNA extraction buffer 1 (Supplementary Table 2), and incubated at 55 °C for 30 minutes, mixing by inverting every 10 minutes. Then 150 µl of DNA extraction

buffer 2 (Supplementary Table 2) was added and incubated for a further 10 minutes at 65 °C. DNA was extracted by adding and gently mixing 690 µl of Phenol:Chloroform:Isoamyl Alcohol 25:24:1 to the mixture. Samples were then centrifuged at 13200 rpm for 5 minutes to collect the aqueous phase, which was then incubated with 80 % volume of isopropanol at -20 °C for 2 hours to precipitate the DNA. Next, samples were centrifuged at 15 g for 30 minutes at 4 °C and the supernatant was then discarded. Pellets were washed with 200 µl of 70% ethanol and then centrifuged to discard the supernatant. Finally, pellets were air dried, resuspended in 40 µl of TE buffer and incubated with 1 µl of RNAse A and 1 µl of Proteinase K for 2 hours at 37°C. A following clean-up step was done using Qiagen DNeasy Plant Mini Kit and DNA quality and concentration was assessed using a Nanodrop photospectrometer ND-1000 (Thermo Fisher Scientific).

Library preparations were performed using NEBNext® Ultra™ II DNA Library Prep Kit for Illumina Sequencing according to the manufacturer's instructions. Libraries were then sequenced with Illumina MiSeq (Illumina, San Diego, CA) and the resulted reads were trimmed with Trimmomatic (Bolger et al., 2014) and assembled using Spades v3.1 (Bankevich et al., 2012). The quality of the assemblies was assessed using a range of statistics shown in Supplementary Table 3. This was to insure that assembles and sequence quality was of a high enough standard and coverage to use in further analysis as well as ruling out any potential contaminants.

2.2.3 Nuclear Species Phylogeny

To obtain gene sequences from the *Galdieria* genomes first all known genes were compiled from the reference genome strain ACUF 074W (assembly ASM34128v1) obtained from GenBank (www.ncbi.nlm.nih.gov). The retrieved gene list went through a filtering process to use only suitable genes in further analysis. Firstly, all genes encoded by the mitochondria and plastid genomes were removed. Genes were then filtered by removing genes that contained less than a 40% identity match relative to each gene. This was to eliminate low matches that could affect analysis by resulting in poor alignments.

Next using an automated pipeline, full gene sequences of selected filtered genes were retrieved from the *Galdieria* assemblies. This consisted of creating BLAST databases for each of the genomes, then searching each reference gene sequence against each of these databases using BLASTN from the BLAST 2.10.0 program (NCBI; Altschul et al., 1997). The results were assessed using relative hit scores (number of bases

matched/length of the gene) where a heat map of scores were co-clustered on similarity across both genes and strains, the clustering method used was a nearest point algorithm typical for this type of analysis (Kalantari and McDonald, 1983). All genes scoring an average relative hit score of >0.4 across all strains were taken onto the next stage of analysis.

The sequences were aligned separately using MUSCLE 3.8.31 (Edgar, 2004) and the resulting alignments consisting of 5627 genes. These were uploaded to Gblocks version 0.91 b (Castresana, 2000) to remove poorly aligned regions applying the options -t = d -b5 = h (http://gensoft.pasteur.fr/docs/gblocks/0.91b/). A final check excluded all genes that were represented by <40% of the original gene alignment length. The final multigene alignment comprised of 3532 genes with 5,212,746 bp DNA positions.

Three red algal taxa belonging Bangiophyceae (*Porphyra umbilicalis*, *Pyropia haitanensis*) and *Cyanidioschyzon merolae*, strain 10D (*Cyanidiophyceae*) were chosen as outgroup taxa (Supplementary Table 4). Maximum likelihood (ML) analyses were performed with IQ-Tree v. 2.0.3 (Nguyen et al., 2015), using the best substitution model estimated under the partition scheme selected by the program (-spp, -m TEST). Phylogenetic trees were inferred applying 10000 ultrafast bootstrap replicates (UFBoot; (Minh et al., 2013) and 1000 replicates of the approximate likelihood ratio test [aLRT] and Shimodaira-Hasegawa, SH-aLRT (Anisimova et al., 2011) for the branch statistical support.

2.2.4 Consensus Network analysis

Further analysis of individual gene trees was performed using SplitsTree 5.0.0_alpha (Huson 1998; Huson and Bryant 2006). The single gene phylogenies of the 3532 genes were produced consisting of 43 taxa. These were concatenated and imported into SplitsTree to construct a consensus network, where under default options The Consensus Network method was used (Holland and Moulton, 2003).

2.2.5 Estimation of Gene Concordance Factors

Gene concordance factors (gCF) were measured to complement previous phylogenetic analysis. From the concatenation of all 3532 genes the phylogenetic species tree was used as the reference tree in the analysis. Each gene tree was also inferred for each locus alignment using IQ-TREE with a model selection. Finally using these trees gCF were calculated using in IQ-TREE with the specific option -gcf (Minh et al., 2020).

2.2.6 Estimations of non-synonymous to synonymous substitutions ratio

For each of the core species, FASTA format sequences of all protein-coding genes, and their corresponding translations were obtained. In cases where there were two or more transcript variants, the longest transcript was selected to represent the coding region. Protein sequences were aligned using MUSCLE v.3.8.31 (Edgar, 2010) and converted into codon aligned nucleotides using PAL2NAL (Suyama et al., 2006). Nonsynonymous substitutions per nonsynonymous site (dN), synonymous substitutions per synonymous site (dS), and dN/dS (ω) values were calculated for each protein-coding gene using CODEML programme in the PAML v 4.3 package (Yang, 2000). The average dN/dS (ω) were calculated using the M0 model which calculates the average ω for the whole gene, over all branches in the phylogeny.

To assess for positive selection M7 and M8 were used. M7 is defined by using the beta distribution to describe dN/dS variation among sites, where dN/dS value is in the range 0 to 1 (no positive selection is allowed). M8 is the same as M7 except it does allow for positive selection, so some dN/dS sites are >1. M7 and M8 were compared using a likelihood ratio test (LRT) to obtain LRT statistic (twice the difference of the log-likelihood between the null model and alternative model). Here this null hypothesis is that no positive selection is taking place (M7). CODEML uses this maximum likelihood approach to fit the observed sequence alignment data to the selected model of evolution, where the parameters that are the best fit along with the likelihood value are provided. To conclude if positive selection has taken place the model that allows for positive selection (M8) must fit the data better than the model that does not include positive selection (M7). Then if the model for positive selection fits best the LRT statistic was then tested for significance against a chi-squared test, where all assumptions were tested and met. The resulting list of genes was then filtered for any enzymes with predicted hydrolase function and signal peptides using The UniProt Consortium 2021 and SignalP5.0 (Armenteros et al., 2019).

2.3 Results

2.3.1 General Features of Nuclear Genomes

The *Galdieria* genomes I sequenced were 17.27 Mb long on average with a range from 13 – 30 Mb. Comparing the genomes with those of the non-*Galdieria* red algae, *Porphyra umbilicalis* (87.89 Mb) and *Pyropia haitanensis* (53.25 Mb) the *Galdieria* genomes are smaller and more conserved. The comparison of sister species *Cyanidioschyzon merolae* (16.43 Mb) is within the variation seen across the *G. sulphuraria* genomes (Supplementary Table 3).

In order to evaluate the nuclear phylogenies of the different *Galdieria* strains first the presence and absence of genes across the different genomes was assessed. A heatmap was created (Figure 2.1) using each genes the relative identity score (number of identical matches/length of gene) and clustered based on similarity across strains and across genes using an optimised algorithm based on minimum spanning tree, also known as the nearest point algorithm (Kalantari and McDonald, 1983). This showed diversity among strains and the difference genes that were present or absent varied. For example, at the top of Figure 2.1 there is a cluster of genes that appear only to be represented in the reference 074 genome. The majority of genes are represented to some degree in most of the strains.

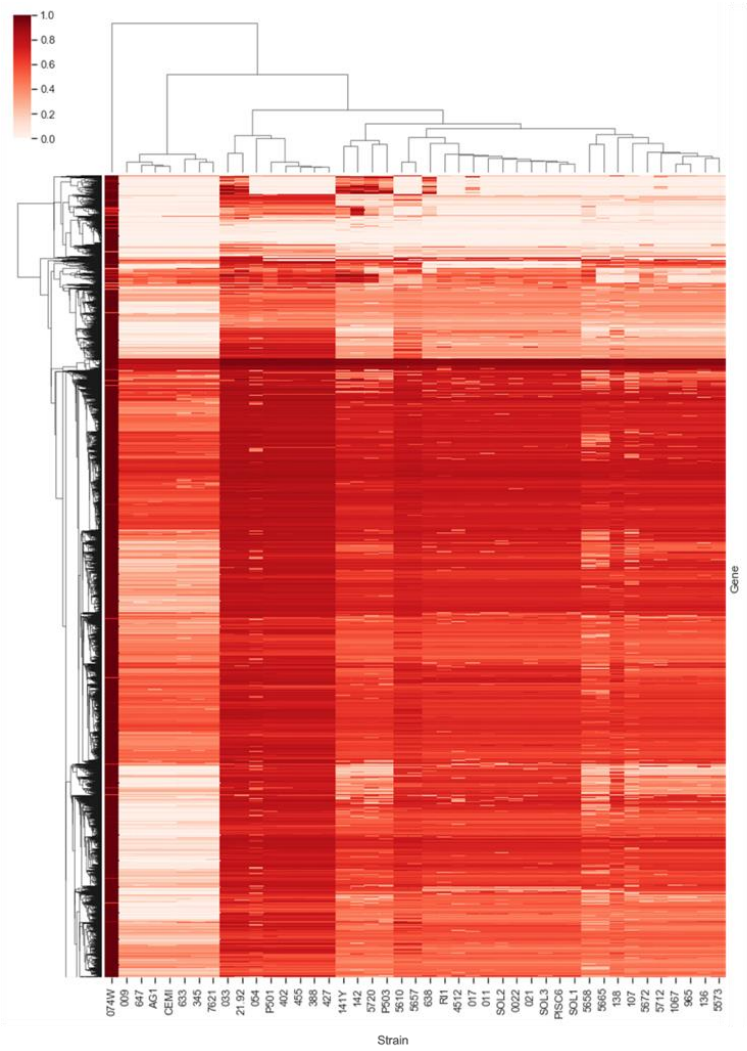


Figure 2.1: All genes relative hit score (number of bases matched/length of the gene) in 43 *Galdieria* genomes. An average relative hit score of >0.4 across all genomes were taken onto the next stage of analysis. Clustered by both Gene (y-axis) and Strain (x-axis) using the nearest point algorithm (Kalantari and McDonald, 1983).

2.3.1 Nuclear Species Phylogeny

Of the 6851 nuclear genes obtained from the 074W reference after initial analysis 3532 genes had suitable coverage across all 43 genomes to be used to determine the species phylogeny. The nuclear phylogeny strongly supported the monophyletic origin of the Cyanidiophyceae class (100% UFBoot, 100% SH-alRT). *G. sulphuraria* lineage originated from one single ancestor confirming the monophyly of the species, but it has diverged into smaller sub lineages that appear to be independent (no gene sharing after divergence). Each *G. sulphuraria* group formed a small monophyletic population

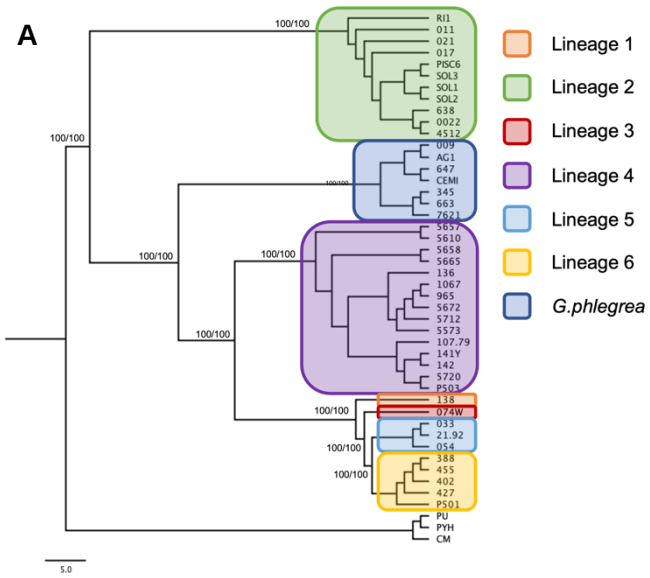
separated from the others so far and keep evolving separately and strains are clustered genetically by geography (100% UFBoot; Figure 2.2A; Figure 2.2C).

The ancestor organism that originated the *G. sulphuraria* lineage originally diverged, giving rise to the clade containing mostly Italian strains (Figure 2.2C), Lineage 2 (RI1, 011, 021, 017, PISC 6, SOL1, SOL2, SOL3 638, 0022, 4512) (100% UFBoot and SH-alRT). The strains belonging to this clade derived from a common ancestor and diverged from each other up to 1% (data not shown). The whole sub lineage, instead, is separated from the other sub lineages by around 22-29% (Figure 2.2B).

A following diverging event (100% UFBoot and SH-alRT) generated the microalga that then colonised the acido-thermal areas surrounding the Mediterranean Sea, *G. phlegrea* (Rio Tinto, Italy and Turkey). The strains within this clade, (009, AG1, 647, CEMI, 345, 663 and 7621) are more separated subgroups with all splits highly supported (100% MLB, 100% SH-alRT).

Alongside the evolution of the above-mentioned lineages, further diverging events led to the origin of more separated subgroups in *G. sulphuraria*. The biggest clade included all the strains from the Culture Collection of Microorganisms from Extreme Environments (CCMEE) and of the Culture collection of Autotrophic Organisms (CCALA), Lineage 4 (Figure 2.2C). This sub lineage is characterised by low intrapopulation genetic dissimilarity ~5% (data not shown) and a high percentage of divergence (16-29%) with the strains of the other lineages (Figure 2.2B).

The next divergent event originated the strain ACUF138, Lineage 1 (100% UFBoot and SH-alRT). The divergence of this strain, collected from the San Salvador site (Figure 2.2C), caused a high accumulation of mutations, which represent 24-31% dissimilarity of the total alignment length (Figure 2.2B). The strain ACUF 074 confirms the next separation event and single lineage (Lineage 3), collected from Indonesian island of Java (100% UFBoot and SH-alRT; Figure 2.2C). This lineage showed a 23-29% dissimilarity in nucleotide sequence compared to the other lineages. Concurrently, the remaining two lineages were separated from the latter population. This led to the well supported (100% UFBoot and SH-alRT) paraphyletic development of strains collected in Taiwan (Lineage 5) and then Iceland and Russia (Lineage 6; Figure 2.2C). These strains were characterised by an 11% sequence difference between the two lineages and a 16-28% from each of the other lineages (Figure 2.2B).



B

	Lin 1	Lin 2	Lin3	Lin4	Lin 5	Lin 6
Lin 1	-					
Lin 2	26 %	-				
Lin 3	31 %	29 %	-			
Lin 4	25 %	23 %	29 %	-		
Lin 5	24 %	22 %	28 %	16 %	-	
Lin 6	24 %	22 %	28 %	16 %	11 %	-

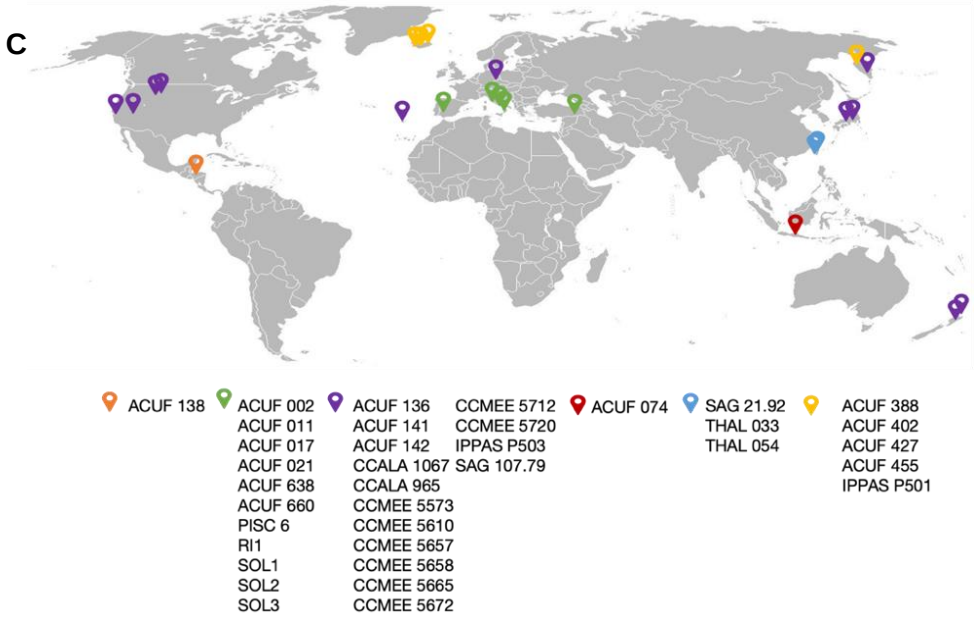


Figure 2.2: (A) Nuclear species trees of Cyanidiophyceae. The phylogeny was inferred from Maximum Likelihood (ML) analysis using the concatenated DNA sequence from 3532 nuclear genes, and the partition scheme for the best substitution model. Ultrafast bootstrap (UFBoot) and the Approximate Likelihood Ratio Test [aLRT] and Shimodaira-Hasegawa (SH-aLRT) support values are indicated near nodes. (B) The table shows the percentage of sequences dissimilarity between lineages. (C) Worldwide distribution of *G. sulphuraria* strains used in this study, coloured according to lineage. Details of the collection sites, along with the sample source and corresponding reference are listed in Supplementary Table 1.

2.3.2 Consensus Network analysis and Estimation of Gene Concordance Factors

The consensus network of the individual phylogenetic trees of the 3532 genes was obtained by 67 splits, resulting in a splits network with 69 nodes and 69 edges (Figure 2.3; majority of nodes and edges are inside the six lineages). This shows clear distinct separation of groups of strains, supporting the formation of the six lineages.

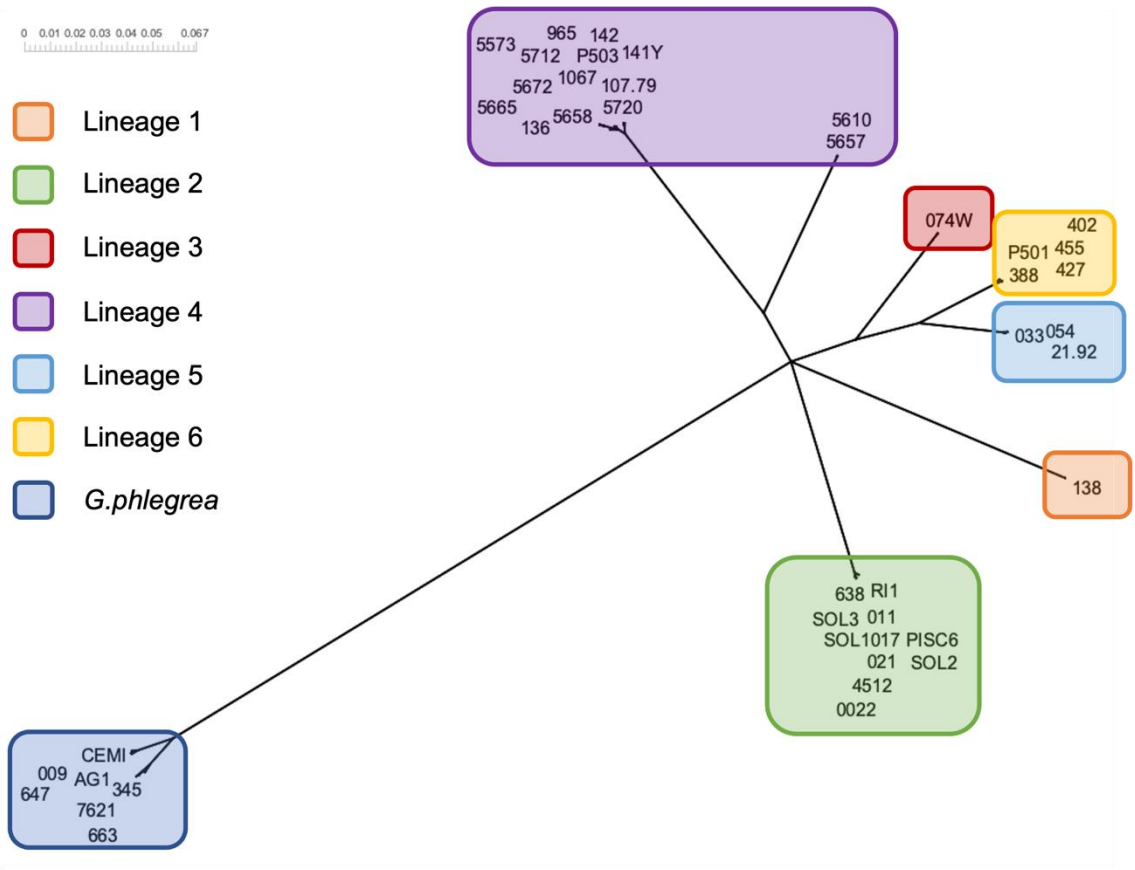


Figure 2.3: Consensus network of 3532 single gene phylogenies using 43 *Galdieria* strains. The Consensus Network method (Holland and Moulton 2003) was used (default options) so as to obtain 67 splits and the Splits Network Algorithm method (Dress and Huson, 2004) was used (default options) giving splits network with 69 nodes and 69 edges.

Concordance factors for each node on the resolved species tree were measured and compared with discordance factors, which relate to the proportion of genes that support a different resolution of the node (gDF). The number of gene trees that supported each branching event (gCF) is shown in Figure 2.4 along with a simplified nuclear species tree. This analysis shows all the final lineages are (for the exception of Lineage 1) highly

supported by the gene trees. The branching events leading to the lineages do however show less support from the individual gene trees.

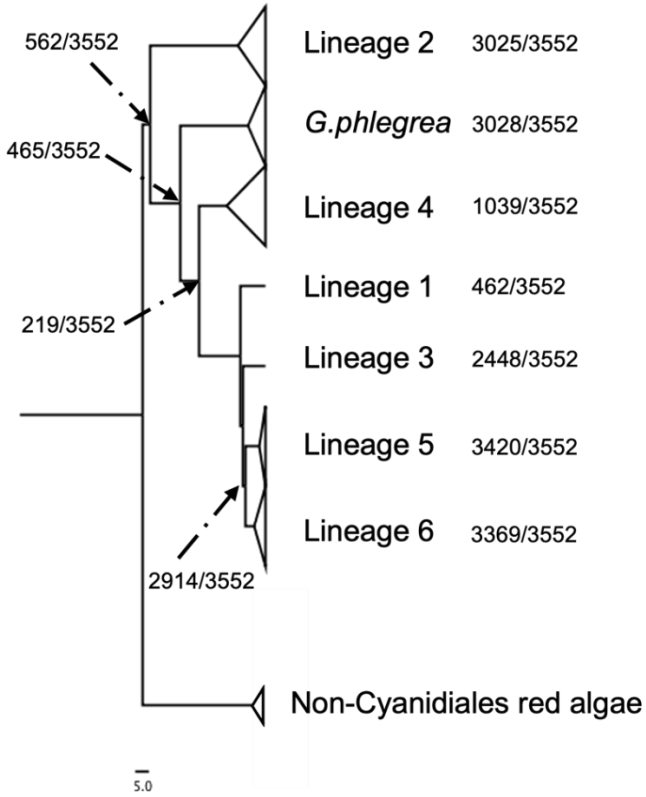


Figure 2.4: Simplified nuclear species tree of Cyanidiophyceae. Number of nuclear gene trees supporting the species tree topology are indicated near the lineages and by the arrows near the nodes, collected form gene concordant (gCF) analysis.

2.3.3 Estimation of non-synonymous to synonymous substitutions ratio

As the previous analysis has highlighted the majority of the divergence observed is between the six lineages and not within them. Therefore one *G. sulphuraria* strain from each of the six lineages identified were selected to represent each lineage in further analysis. These are as follows; 017 (Lin 2), 033 (Lin 5), 074W (Lin 3), 107.79 (Lin 4), 138 (Lin 1) and 427 (Lin 6). These strains are referred to as the core six or 017, 033, 074, 107, 138 and 427 retrospectively.

Substitution rates were measured in nuclear genomes between the six *G. sulphuraria* lineages using the core six strains (Table 2.2). Our analyses included 1947 nuclear encoded genes that were present in all six genomes. The average number of substitutions per synonymous site between the genomes was 2.74 ± 2.32, showing high

variability across the genes. The rates of substitution at nonsynonymous sites (dN) were lower at 0.36 ±0.25 per site. The dN/dS ratio, which can be used to gauge the intensity and directionality of selection, was 0.13 ± 0.07 which is consistent with purifying selection acting on the majority of nonsynonymous sites. Figure 2.5A shows the plots of the dN vs dS of the nuclear genes. The majority of genes have dN <1 and dS <10. Additionally, Figure 2.5B shows the distribution of dN/dS = ω to be normal and most genes have ω ≤ 0.2 as is expected. For all genes the ratio of substitutions is never above 0.43.

Table 2.2: Nucleotide substitution rates in the nuclear genomes between the six lineages of *G. sulphuraria* (strains: 017, 033, 074W, 107, 138 and 427). CV=SD/average.

	Substitutions per gene
	Nuclear
Synonymous sites	
Average (SD)	2.74 (2.32)
CV	0.847
Nonsynonymous sites	
Average (SD)	0.36 (0.25)
CV	0.694
dN/dS	
Average (SD)	0.13 (0.07)

Analysis of the dN and dS values under different models was used to assess any genes under positive selection. This resulted in 288 nuclear genes (Supplementary Table 5) showing positive selection that would require further investigation. For the purpose of this work the resulting genes were assessed for putative secreted hydrolases, this revealed one gene (Gasu_27500) a beta-galactosidase that could be relevant in the degradation of lignocellulosic material with a test statistic of 23.43 compared to a chi squared value of 13.82 at $p<0.001$.

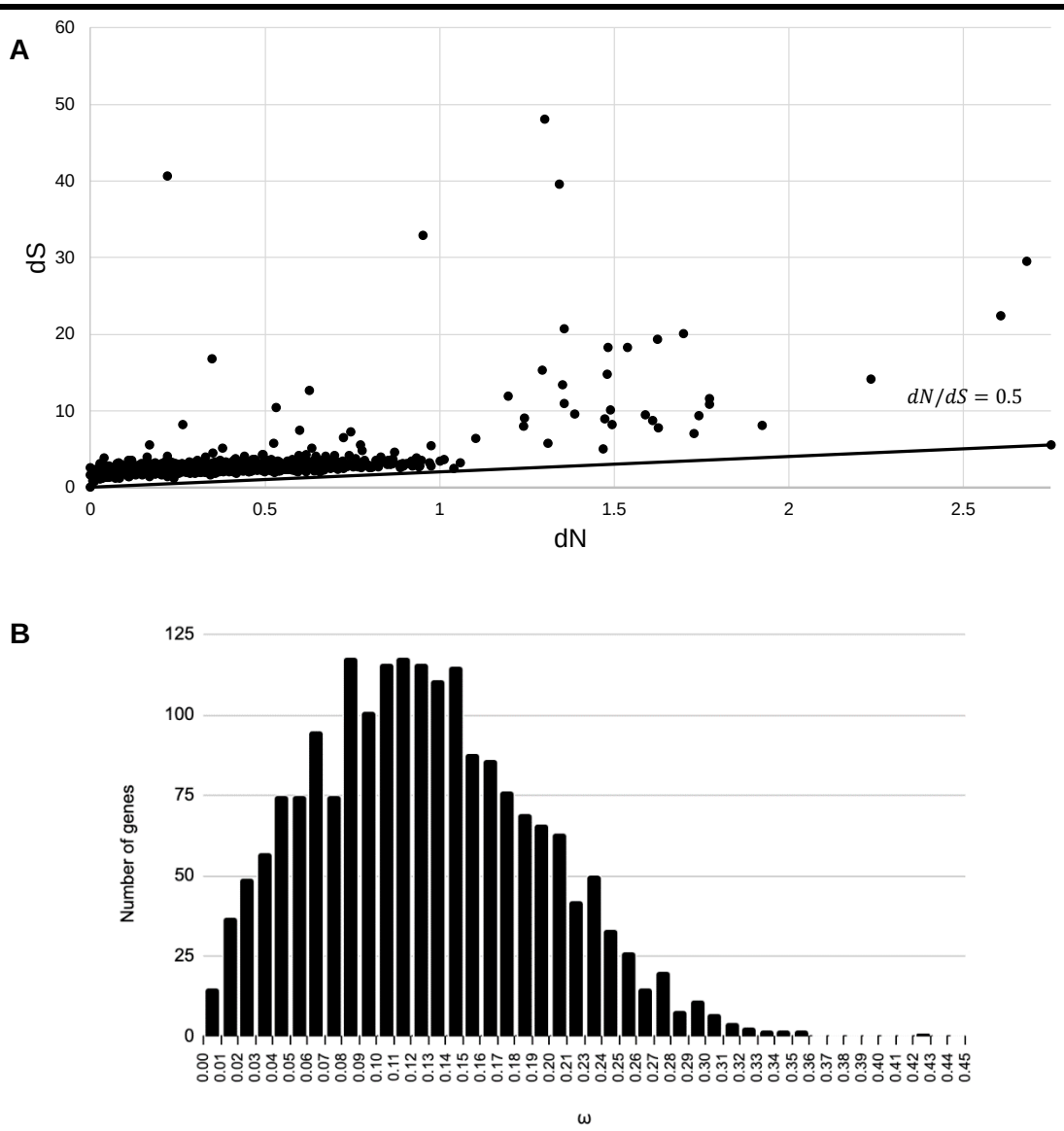


Figure 2.5: (A). Pairwise omega (dN/dS) values. This graph shows pairwise dN vs dS values for *Galdieria* nuclear genes. The line is dN/dS = 0.5. (B) Histogram showing the distribution of ω.

2.4 Discussion

2.4.1 General Features of Nuclear Genomes

Often due to the small number of unambiguous morphological features to distinguish between them, identification of different species and genera within unicellular microorganisms has previously been a challenge. This is the case for the classification of *Galdieria*, hence the use of techniques like NGS are indispensable when trying to understand the taxonomy and fundamental biology of the organism. Initially the

sequencing of the 43 genomes discussed in this chapter revealed a range in size of the genomes (Supplementary Table 3), along with initial sequence coverage of genes obtained from reference strain ACUF 074W (ASM34128v1), (Figure 2.1) showing extremely varied coverage across the different genomes. Even when using the not statistically supported nearest point clustering algorithm, clusters of strains are obvious, emphasising the diversity across the genomes along with the likely diverging evolutionary paths (Kalantari and McDonald, 1983).

Eukaryotic genomes vary dramatically in size and gene counts, typically these factors reveal little about the complexity of the organism, however, genome size does matter (Pray, 2008; Maloy et al., 2013). Genome size is influenced by many things including the rate at which changes in the base DNA occur (deletions and insertions) along with how effectively an organism reacts to these changes and whether they are selected for or against (Yampolsky, 2016). Genome size is also closely linked to morphology, namely cell size, i.e typically larger cells will have larger genomes, it is suggested that cell size can explain a high proportion of variation in genome size (Shuter et al., 1983; Gregory, 2005; Beaulieu et al., 2008; Jiang et al., 2010; Malerba et al., 2020). It has been shown that typically species with larger genomes will have lower metabolic rates, as well as developing and growing at a decreased rate compared to species with smaller genomes (Gregory, 2005; Vinogradov and Anatskaya, 2006; Lane and Martin, 2010; Malerba et al., 2020). The correlation between cell size and genome size is based on that an accumulation of redundant DNA (transposons, introns, junk DNA) will have a fitness cost. Meaning that producing excessive or large amounts of DNA is an energetic burden to the cell, it then follows that larger cells will better tolerate larger genomes. A recent study by Malerba et al., 2020 found this to be true for a eukaryotic green Alga *Dunaliella tertiolecta*, there was direct evidence that reduction in relative genome size showed associated fitness benefits. In terms of total biovolume and maximum growth rate a higher fitness was observed in lineages that contained relatively smaller genomes (Malerba et al., 2020). Previous research on sister species *Cyanidioschyzon* that has a similar sized genome revealed that had condensed its genome size by a reduction in the number of genes and had lost nearly all introns (Matsuzaki et al., 2004; Keeling and Slamovits, 2005). Lynch and Conery, 2003 argue that population size is the main driving factor effecting genome size, that an increase in population size is followed by a decrease in cell size thus causing a decrease in genome size. In this scenario low population size leads to an accumulation of slightly deleterious material (such as transposons, more and larger introns), which leads to an increase in cell size. This is as the relative efficacy of purifying selection vs genetic drift is lower when population sizes

are lower. This provides an explanation for *Galdieria's* relatively small genome and suggests the range in genome size seen across the different lineages could be due to separate populations experiencing different circumstances influencing genome reduction, cell size and population size.

A recent study by Xu et al., 2020 showed that plants exposed to high selective pressure i.e., extreme environments caused the independent appearance of the same trait in different lineages (genomic convergence). Multiple types of conversion events were found, examples included changes in gene copy number, amino acid usage, gene expression, and even GC content (Xu et al., 2020). Thus, as would be expected for *Galdieria*, an organism under extreme environments the genomes, broadly speaking over all eukaryotes are small and sit towards the lower end of the smallest reported eukaryotic genome ~ 10 Mb (Blommaert, 2020). However, across the 43 genomes there was a 17 Mb range in genome size highlighting once again the diversity shown across the species and the effect evolution of isolated populations can have. Often it is non-coding regions that expand or contact like intergenic regions, introns and transposons and could well be what is happening within these genomes. Sequencing of genomes undergoing adaptive evolution under exposure to extreme environments is an effective approach for identifying potential genes related to survival and functional environmental adaptation such as utilisation of carbon sources.

2.4.2 Nuclear Species Phylogeny

The phylogeny work from this study is congruent with published data and confirms the monophyly of *Cyanidiophyceae* species (Ciniglia et al., 2004; Yoon et al., 2006; Iovinella et al., 2018). Within gene sequences there was a significant variability between nucleotides, thus leading to the well-supported (100% UFBoot and SH-alRT) phylogenetic divergence of *G. sulphuraria* into six lineages (Figure 2.2). Analysis of the phylogenetic relationship within the mitochondrial and plastid species trees in *Galdieria* reveals an incongruent evolution between the three genomes (Iovinella and Lock et al., unpublished). Though this is not unexpected as for photosynthetic eukaryotes the relative mutation rates among mitochondrial, plastid, and nuclear genomes have been shown to be different between, plants, green and red algae (Lynch and Walsh, 2007; Drouin et al., 2008; Leliaert et al., 2012; Smith et al., 2012). Thus, resulting in different evolution between the respective genomes. The difference in evolution across the three genomes in each case still resulted in the same six lineages being identified (Iovinella and Lock, unpublished). Discordant phylogenies within a species are completely expected consequence of meiotic or some other type of recombination. Analysis of single

nucleotide polymorphisms (SNPs) within the *G. sulphuraria* lineages showed linkage disequilibrium which is also indicative of recombination (Jessica Downing).

The branches in the nuclear phylogenies leading to the each of the different lineages are always very long in comparison to the terminal branches leading to the single strains which are always very short. This indicates a high divergence between each of the lineages but suggests a low genetic diversity within them. All the lineages diverged from each other upwards of 11% when assessing nucleotide sequence dissimilarity (average 23.6 %; Figure 2.2B), these percentages easily fit into the 8-11% range identified as the threshold level for genus assignment in Rhodophyta (Cassano et al., 2012; Liu et al., 2020). Though it should be noted that this threshold level was typically used for multicellular algae and in conjunction with morphological criteria and is not to suggest each linage is a different genus but rather highlight the extensive diversity and divergence. This work gives strong supporting evidence for the six identified lineages to be separate species, however, this cannot be said with certainty without further research into characterising the lineages based on other traits and assessing whether lineages populations can interbreed.

2.4.3 Consensus Network Analysis and Estimation of Gene Concordance Factors

Phylogenetic relationships between the six lineages were assessed in every nuclear gene to understand their contribution to the divergence of *G. sulphuraria.* It should be noted the previously presented concatenation analysis returned a return fully resolved and well-supported species tree (Figure 2.2A). Consensus network analysis complimented the divergence of *G. sulphuraria* and the formation of the six lineages (Figure 2.3). It also indicates that there is no evidence for recombination between the various lineages, meaning that the populations are isolated and thus, they are evolving separately. This analysis is further evidence that these lineages are interbreeding sets of strains, which exist within a location that in fact represent individual species. Additionally, all the genes in the concordance analysis supported the divergence of the species into the six lineages, however not all of them supported the phylogenetic relationships among them (Figure 2.4).

Not unsurprisingly the majority of genes presented a distinct phylogeny, this is as each individual gene is likely to have a unique evolutionary path. It is known that different genes will evolve at different rates and be under different selection pressures. Recombination will also result in differing tree topologies and make one consistent

phylogeny across the genome extremely unlikely; this is standard when looking at eukaryotic genetic data. Ultimately the divergence into the six lineages is generally well supported but the relationship between the lineages in the branching events leading to the lineages is variable. This is evidence in support of these lineages being isolated populations evolving differently, and that within these lineages there are interbreeding populations. Analysis of single nucleotide polymorphisms (SNPs) within the *G. sulphuraria* lineages showed linkage disequilibrium, which is also indicative of recombination (Jessica Downing). Taking all of this into consideration along with the overall well supported species tree (100% UFBoot and SH-alRT), there is high confidence that the *G. sulphuraria* species tree is resolved.

2.4.4 Estimations of non-synonymous to synonymous substitutions ratio

The evolution of a protein coding gene is influenced by many factors, with correlations to intron number, gene expression and the essentiality of the gene to name a few (Wall et al., 2005; Drummond et al., 2005; Larracuente et al., 2008; Jo and Choi, 2015). The types of substitutions acting on the sequence are important, synonymous substitutions within a protein are random and will likely be tolerated across generations. Non-synonymous substitutions are due to neutral evolution and more often removed by purifying selection, however, a proportion are fixed as a result of positive selection and thus increasing the rate of protein evolution.

Analysing the rate of the substitutions occurring in a protein can identify information about which selective pressures are happening (Del Amparo et al., 2021). The calculation of dN/dS can therefore help to identify genes that are under particular biochemical or ecological constraint, or conversely putative proteins involved in survival adaptation. Analysis of the synonymous and non-synonymous substitutions of the nuclear genes present in all core six strains confirmed different evolutionary pressures across genes (Figure 2.5). High substitution rates along with events such as horizontal gene transfer (HGT), could be the main evolutionary forces shaping the divergence of the *G. sulphuraria* species*,* these biological mechanisms have been linked to inducing phylogenetic incongruence (Hill et al., 2010; Som, 2015; Paquola et al., 2018*).* HGT events have been widely confirmed in *G. sulphuraria* (Schönknecht et al., 2013; Jain et al., 2014; Rossoni et al., 2018, 2019), this may strongly influence the phylogenetic relationship among strains. Genes acquired horizontally are likely to be involved in the adaption of *Galdieria* to its harsh environment, this will include osmotic resistance, salt

tolerance, carbon and amino acid metabolism, metal and xenobiotic resistance/detoxification non-metabolic and uncertain functions (Rossoni et al., 2019).

High variability of the synonymous substitutions across nuclear genes is generally not considered deleterious as these mutations are typically regarded as neutral or at least have a much smaller effect on fitness, compared to non-synonymous substitutions (Eyrree-Walker and Keightley, 2007; Parmley and Hurst, 2007). Calculation of the coefficient of variation (SD/mean; Table 2.2) gives a standardised measure of the dispersion of the distribution of dN (0.694) to dS (0.847), and in this case dN is much higher than originally calculated and so relatively more variable (Yang and Nielsen, 2000; Spielman and Wilke, 2015; Moutinho et al., 2020). This could be due to different selection pressures on genes, for example typically proteins such as histones are among the most conserved proteins (Isenberg, 1979; Peterson and Laniel, 2004) whereas membrane and exported proteins tend not to be (Drouault et al., 2002; Nuhse et al., 2007; Liu and Zhang, 2018). It is not unusual in these types of analysis to see dN/dS rates < 0.5 as it is expected that most genes will be under purifying or neutral selection (Yang and Nielsen, 2000). This is as nonsynonymous changes are more likely to have a functional consequence and therefore will generally be deleterious. This means they are removed from populations more rapidly and thus their rate is typically slower than the rate of synonymous changes.

However, a dN/dS <1 does not mean that all genes are under purifying or neutral selection. Genes under adaptive evolution are favoured for and in the case of *G. sulphuraria* can contribute to its adaptation and survival in extreme environments. The analysis for identifying nuclear genes under positive selection revealed 288 genes. These genes are a valuable resource in investigating the adaptations of *G. sulphuraria* and how it not only survives but thrives in low pH and high temperatures. However, the focus of this thesis is to look for genes involved in the degradation of lignocellulosic material. Analysis of this list showed only one gene that had both predicted hydrolases activity and the presence of a signal peptide. Gasu_27500 a beta-galactosidase, part of the family of glycoside hydrolase enzymes catalysing the hydrolysis of beta-galactosides through breaking the glycosidic bond to form monosaccharides (Lombard et al., 2014; Saqib et al., 2017). The presence of the signal peptide means the protein is on the secretory pathway and likely excreted extracellularly meaning the protein will be tolerant to acidic conditions, a useful feature for industrial applications.

2.4.5 Conclusions

This chapter aimed to understand the phylogenetic relationship among different *G. sulphuraria* strains and give an insight into the evolutionary history of the species using NGS sequencing data. It was found that even if morphological traits slightly vary between *Galdieria,* molecular tools allowed the identification of huge variability between them. The resulting phylogenetic analysis identified the divergence of the species into six clear linages that have been evolving separately. Analysis of the synonymous and non-synonymous substitutions confirmed the differential evolutionary pressure between the strains and gave rise to multiple genes under positive selection. These genes present good candidates for exploration into *G. sulphuraria's* adaptation to its environment. One gene Gasu_27500 was identified as having potential involvement in the degradation of lignocellulosic material with predicted hydrolase activity and presence of a signal peptide. The *G. sulphuraria* genomes presented in this chapter are extremely diverse and hosting a large number of potentially noteworthy enzymes. It would be of interest to further examine each of the core six lineages for novel industrially relevant enzymes involved in the degradation of lignocellulosic material.

Chapter 3 - CAZyme repertoire of *G. sulphuraria*

3.1 Introduction

The demand for sustainable and renewable energy is at an all-time high, it is vital in the goal to mitigate global climate change and stop the use of finite fossil fuels before it is too late. There are multiple alternatives to fossil fuels derived from natural sources, such as hydro and wind energy as well as the use of lignocellulosic material. Lignocellulosic biomass is the most abundant organic raw material worldwide, comprised of mostly of hemicellulose (20–35%), cellulose (35–50%) and lignin (10–25%) (Harrison et al., 2011; Bharathiraja et al., 2017; Woiciechowski et al., 2020). It is a promising source for renewable energy and useful biproducts, thus considered one of the most competitive alternatives to fossil fuels (Bhatia et al., 2020; Strazzulli et al., 2020).

The pool of currently untapped lignocellulose biomass is a potentially rich source of fermentable sugars for the production of bioethanol (Vohra et al., 2014). However, the conversion of lignocellulose into biofuels is a challenging process (Talebnia et al., 2010; Hassan et al., 2018). It requires saccharification of the hydrolytically resistant polymers before the useful fermentable sugars can be released (Himmel et al., 2007; Singh et al., 2010). An established route to release these sugars is through pre-treatment and enzymatic hydrolysis (Kumar et al., 2009). During these processes hemicellulose is converted into pentoses (arabinose and xylose) and hexoses (glucose, galactose, and mannose) while cellulose is converted into glucose (Lynd et al., 2002). Xylan is the most abundant hemicellulose thus making it a majority component of plant biomass. Due to the structure of lignocellulosic materials hemicellulose such as xylan present the most accessible polysaccharides ready for degradation into its fermentable sugars (Bastawde, 1992; Saha, 2003; Rennie and Scheller, 2014).

Carbohydrate-active enzymes (CAZymes) are classified as any enzyme involved in the synthesis, modification, metabolism and degradation of carbohydrates (Lombard et al., 2014). CAZymes are separated into multiple classes based on catalytic activity and assigned to further subfamilies based on amino acid sequence and structure similarity. The five families are glycosyltransferases (GTs), glycoside hydrolases (GHs), carbohydrate esterases (CEs), polysaccharide lyases (PLs), and auxiliary activities (AA) (http://www.cazy.org/). Additionally, there is a category for carbohydrate-binding modules (CBMs), these enzymes have carbohydrate-binding activity. Presently, CAZymes that are thermostable and even acid tolerant offer an advantage in the production of biofuels and hence are key in current attempts of increasing productivity,

efficiency and yield in second-generation biorefineries (Mukhtar and Aslam, 2020; Chettri et al., 2021).

Enzymes from extremophiles (extremozymes) have gained huge interest in biotechnology due to their ability to function in extreme environments where their mesophilic counterparts would quickly denature. There is a long list of desirable features that these extremozymes can demonstrate, including resistance to extreme temperatures and pH, as well as high concentrations of salt, detergents and organic solvents (Espliego et al 2019; Merino et al., 2019; Strazzulli et al., 2020; Mukhtar and Aslam, 2020). These attributes thus make them ideal tools for applications in an industrial setting, for instance the paper and textile industry. Another useful application is the conversion of lignocellulosic material into biofuels, where the conditions used are often harsh both chemically and physically (Raddadi et al., 2015; Jin et al., 2019).

Galdieria sulphuraria is a eukaryotic unicellular red alga from the family Cyanidiophyceae, an ancient class of rhodophytes (Inovella et al., 2019*)*. This extremophilic microalgal species is found thriving in geothermal sites all over the world, where they have adapted to extreme growth conditions. They are subjected to a wide range of temperatures (up to 56 °C) and areas of low pH (0-4) (Gross and Schnarrenberger 1995; Ciniglia et al., 2004; Yoon et al., 2006; Pinto 2007). G. sulphuraria demonstrates the distinguished and unique ability to grow autotrophically, mixotrophically and heterotrophically (Gross et al., 1998; Gross and Oesterhelt, 1999). In comparison to most other microorganisms *G. sulphuraria* utilises a larger number of carbohydrates, it has been documented that over 50 different carbons sources support growth (Gross and Schnarrenberger, 1995; Schönknecht et al., 2013; Sloth et al., 2017; Náhlík et al., 2021; Curien et al., 2021). *G. sulphiraria* demonstrates incredible tolerance to high salinity (up to 2–3 M), elevated pressures, high concentrations of heavy metals and sugar concentrations greater than 400 g/L (Gross and Oesterhelt 1999; Weber et al., 2004; Schmidt et al., 2005; Schönknecht et al., 2013; Sloth et al., 2017).

The adaptations of *G. sulphuraria* to its harsh environment along with its unique growth characteristics and metabolic versatility make it a promising candidate for investigation for potential biotechnological development. The enzymes produced are expected to be highly thermostable, and any secreted enzymes are likely to display high acid tolerance. As such, any extremozymes discovered involved in the breakdown of polysaccharides would make ideal candidates for utilisation during the pre-treatment steps of lignocellulosic degradation during biofuel production. They could aid by increasing

efficiency, and yields whilst reducing by-products, and even improving the cost effectiveness of the entire process (Strazzulli et al., 2020; Chettri et al., 2021).

3.1.1 Aims

The *G. sulphuraria* genomes in the previous chapter were shown to be extremely diverse, hosting a large number of potentially interesting enzymes. Therefore, it is of interest to examine the diverse core 6 lineages identified in Chapter 2 to explore the species for novel CAZymes. This chapter will provide an investigation into the CAZyme profile of the core six *G. sulphuraria* genomes along with another extremophilic red algae *Cyanidioschyzon merolae* and a mesophilic red alga *Porphyridium purpureum*. This will lead to information on how *G. sulphuraria* may achieve its growth on numerous carbon sources and should provide an excellent collection of targets for industrially relevant lignocellulose degradation enzymes that can be explored.

3.2 Materials and Methods

One *G. sulphuraria* strain from each of the six lineages identified in Chapter 2 were selected to represent each lineage in further analysis. These are as follows; 017 (Lin 2), 033 (Lin 5), 074 (Lin 3), 107 (Lin 4), 138 (Lin 1) and 427 (Lin 6).

3.2.1 DNA preparation, extraction and sequencing

Cultures for each of the six strains were grown in Allen medium mixotrophically with 10 g/L sucrose at pH 2, under a 12h/12h light/dark cycle (42 µmol m^{-2} s^{-1} at 37 °C). Samples were collected from stock solutions by centrifugation (5 m at 13.2 rpm) and supernatant discarded. Tubes were placed in a dry ice ethanol bath for 30 seconds then transferred to a 30°C water bath, this was repeated 4 times. Next 1 µl of Protienase K and 100 µl of Viscozyme™ were added and the tube incubated for an hour at 37°C. Then 40 µl of PBS pH 7.5 was added and vortexed to mix. 500 µl of DNA extraction buffer 1.1 was added and incubated at 55 °C for 30 minutes, mixing by inverting every 10 minutes. Then 150 µl of DNA extraction buffer 2 (2.1 for strains 017, 033, 074 and 2.2 for strains 107, 138 and 427; Table 3.1) was added and incubated for a further 10 minutes at 65°C. Next 690 µl of Phenol:Chloroform:Isoamyl Alcohol 25:24:1 was added and mixed gently though inversion for 5 minutes. Centrifuged at 13.2 rpm for 5 minutes and 600 µl of the top layer of the supernatant was then taken and placed into a fresh tube. Here 480 µl of isopropanol was added and samples stored at -20 °C for 2 hours. Following this, samples were centrifuged at 15 g for 30 minutes at 4 °C, supernatant was then discarded. 200 µl

of 70 % ethanol was added then tubes centrifuged at 13.2 rpm for 5 minutes and supernatant discarded. Finally, tubes were air-dried, and DNA re-suspended in 40 µl of TE buffer. Clean-up of DNA samples were completed using Zymo DNA Clean & Concentrator™-25 kit (Zymo Research, D4033) according to manufacturer's instructions. Prior to elution DNA Elution Buffer was heated to 65°C.

3.2.1.1 Library preparation

Long read sequencing libraries were prepared for sequencing on an Oxford Nanopore Technologies MinION sequencer, using the most recently available ligation sequencing kit and flow cell at the time of sequencing. For a summary of the sequencing kit number and flow cell used, see Supplementary Table 6. In all cases, genomic DNA was subject to an additional clean up step using a 0.6:1 ratio of AMPure XP beads:sample prior to long read sequencing using the Oxford Nanopore Technologies' (ONT) MinION system. The ligation sequencing protocols were performed as per the manufacturer's guidelines with modifications as follows: Incubation times for end repair steps were increased from 5 minutes to 30 minutes; ligation reactions were performed at room temperature for 1 hour, and elution steps were performed at 37°C for 15 minutes. The resulting DNA libraries were sequenced on MinION flow cells with a 48-hour run time.

3.2.2 Genome assembly

For all strains oxford nanopore technologies basecaller Guppy v4.0.11 was used to call raw reads. Then strains 017, 033, 074 and 427, were assembled with SMARTdenovo (Liu et al., 2020), haplotypes were cleaned up by removing contigs with less than 10% unique material, raw reads were polished with medaka (https://github.com/nanoporetech/medaka), then polished three times with pilon (Walker et al., 2014) using Illumina reads (Chapter 2), which were mapped to the draft assembly using the burrows wheel aligner (Li and Durbin, 2010). For 138 the initial assembler used was Canu2.1 (Koren et al., 2017) and the same polishing method was used. For 107 assemblies were made using multiple assemblers (Canu, Miniasm (Li, 2016), Raven (Vaser and Sikic, 2021) and SMARTdenovo) then these were aligned against each other using minimap2 (Li, 2018). Finally, these contigs were manually checked and resolved to give final assembly this was then polished according to method described above (Dr John Davey).

Table 3.1: Composition of different DNA extraction buffers.

Buffer 1.1	Buffer 2.1	Buffer 2.2
200 mM Tris-HCl pH 8	200 mM Tris-HCl pH 8	100 mM Tris-HCl pH 8
200 mM NaCl	200 mM NaCl	700 mM NaCl
100 mM LiCl	100 mM LiCl	20 mM EDTA pH 8
25 mM EDTA pH 8	25 mM EDTA pH 8	2 % CTAB
1 M Urea	1 M Urea	0.0125 mM PVP-40
1 % SDS	1 % CTAP	
1 % NP-40	100 mM Lithium acetate	

3.2.3 RNA preparation, extraction and sequencing

G. sulphuraria cultures (017, 033, 074, 107, 138 and 427) were grown under a 12h/12h light/dark cycle under 42 µmol m^{-2} s^{-1} at 37 °C on an orbital shaker (130rpm). The experimental design followed different growth conditions to obtain a great variety of mRNAs. Samples were grown in Allen medium mixotrophically supplemented with either, 10 g/L Sucrose (Sigma Aldrich), 0.5 % (w/v) Carboxymethylcellulose sodium salt (CMC Cellulose, CAS Number:9004-32-4, Sigma-Aldrich), 0.5 % (w/v) Xylan from Corn Core (CAS Number:9014-63-4, Tokyo Chemical Industry UK Ltd), or 0.5 % (w/v) Laminarin (CAS Number:9008-22-4, Sigma Aldrich), all titrated to pH 2. Samples were collected by centrifugation at 1h, 12h, 96h, 192h and 336h. For all the treatments described above, pellets were washed three times in a PBS buffer pH 7.4 and then stored at -80 °C until RNA extractions were carried out.

Total RNA was isolated using RNeasy Kit (Qiagen) after the frozen biomass was mechanically disrupted with a pestle to form a fine powder. All RNAs were treated with DNaseI (Qiagen) and then pooled by strain relative to the concentration of each sample, measured by Nanodrop photospectrometer ND-1000 (Thermo Fisher Scientific). RNA integrity was further determined using an Agilent BioAnalyzer 2100 (Agilent Technologies). RNA library preparation and sequencing were performed at Novogene (UK) Company Limited (Cambridge). Library preparation was performed using NEB Next® UltraTM RNA Library Prep Kit (NEB, San Diego, CA, USA), employing AMPure XP Beads to purify the products of the reactions during the library prep. Poly-a mRNA was isolated using poly-T oligo-attached magnetic beads, then fragmented through

sonication and enriched into 250-300 bp fragments. The purified mRNA was converted to cDNA and subjected to the adaptor ligation. The barcoded fragments were finally multiplexed and ran on the Illumina Novaseq 6000 (s4 flow cell) to acquire 20 million read pairs per sample, using the 150 bp PE sequencing mode.

3.2.4 Genome annotation

The first step in genome annotation is to find all genes in a given genomic sequence. Gene prediction was performed on the genome sequences by De novo gene prediction using the programme AUGUSTUS (Stanke et al., 2004) with parameters trained from *G. sulphuraria* strain 074W, obtained from GenBank (www.ncbi.nlm.nih.gov) (this gene prediction work was carried out by Jessica Downing, unpublished). The prediction consists of the protein coding parts of the genes as well as the amino-acid sequences of the predicted genes. Untrimmed RNA sequencing reads were aligned to their respective Illumina assemblies using the STAR aligner v. 2.7.3 (Dobin, 2012). Alignments were filtered with AUGUSTUS v. 3.3.3 filterBAM and converted to AUGUSTUS hints with bam2hints using the defaults (Stanke et al., 2006). The annotation was performed with AUGUSTUS using both the generated hints and de novo gene prediction. Coding sequences with over 94% identity were removed with CD-HIT v. 4.8.1 (Fu, 2006; Li, 2012). Predicted amino acid sequences were generated with EMBOSS transeq v. 6.6.0 (Rice, 2000) with a minimum ORF length of 40. Sequences without start codons and with stop codons contained in the sequence were removed.

3.2.5 Ortholog Identification and Clustering

The predicted amino-acid sequences of *G. sulphuraria* proteins were clustered into orthologous groups using OrthoFinder (version 2.3.11) software (Emms and Kelly, 2015). Orthologs were identified and clustered by an all-versus-all protein comparison with predicted proteins of the core 6 strains along with extremophile *Cyanidioscshyzon merolae* and mesophilic *Porphyridium purpureum*.

3.2.6 CAZyme Gene Identification and Signal Peptide Prediction

CAZymes in the *G. sulphuraria* core six genomes, were identified and annotated using the HMMER v3.3.2 package (http://hmmer.org/) with the dbCAN CAZyme database (https://bcb.unl.edu/dbCAN2/blast.php) (Yin et al., 2012). CAZymes were filtered for genes that contained a predicted signal peptide and showed hydrolase, peroxidase or uncharacterised function. Prediction of signal peptides was conducted using the SignalP5.0 (Armenteros at al., 2019)

3.3 Results

3.3.1 Genome sequencing Assembly, Gene modelling and Genome Comparisons

The general features of the sequencing and assembly of the *G. sulphuraria* genomes are presented in Table 3.2. Long length DNA sequencing of the core six *G. sulphuraria* genomes was achieved using oxford nanopore MinION technology and revealed a range in genome size between the strains (as expected). The lowest number of contigs a strain was resolved into was 74 for strain 107 and the highest was 190 contigs from strain 138. All strains sequenced has a coverage >100, notably strain 138 having 1472 times coverage.

Table 3.2: *G. sulphuraria* core six genomes (017, 033, 074, 107, 138 and 427) sequencing and assembly statistics. Genomes were sequenced using MinION technology and assembled according to Section 3.2.2. *Assuming 13 Mb genome

Strain	Coverage*	Assembly size bp	Largest Contig bp	Average Contig bp	Num Contigs	Contig N50 bp
017	140	13,702,654	335,351	141,264	97	184,865
033	949	15,041,697	345,493	115,705	130	191,209
074	824	14,540,173	411,047	120,167	121	186,075
107	334	14,206,823	499,985	191,984	74	202,925
138	1472	19,750,973	379,478	103,952	190	171,979
427	106	12,933,762	328,005	148,664	87	190,919

3.3.2 Orthologue analysis

The total number of genes in each of the *G. sulphuraria* genomes varied along with genome size (Table 3.3). Strain 017 was the largest genomes contained the highest number of genes, while strain 033 contained the lowest number of genes, it was in fact the third largest genome. *C. merolae* has the lowest number of genes but a genome larger than 5/6 of the *G. sulphuraria* strains. The *P. purpureum* genome was of similar size to strain 017 but contained over 3000 more genes. On average across the *G. sulphuraria* genomes GC content was 38.6 % this is lower than the values given from the *C. merolae* (54.9 %) and *P. purpureum* (55.8 %). Cluster analysis on the three

species and eight genomes (*G. sulphuraria*, *C. merolae* and *P. purpureum*), identified 6298 orthologous clusters (Supplementary Table 8). Analysis of these clusters suggested that on average across the *G. sulphuraria* genomes 75.55 % of orthogroups contain proteins from the species, compared to 52.6 % from *C. merolae* and 64.3 % *P. purpureum*. Among the set of homologous genes, there were in total 313 single copy orthologs in *G. sulphuraria*. This ranged across the strains with strain 033 containing the lowest species-specific orthogroups (7 genes) to strain 107 with the highest (145 genes). *C. merolae* contained 142 species-specific gene and *P. purpureum* contained 2473 (Table 3.3).

Table 3.3: *G. sulphuraria* core six genomes (017, 033, 074, 107, 138 and 427) assembly and orthologs analysis statistics. Genomes were annotated using AUGUSTUS and orthogroups assessed using OrthoFinder.

	017	***033***	***074***	***107***	***138***	***427***	***C. merolae***	***P. purpureum***
Genome Assembly (Mb)	19.75	14.21	15.04	14.54	13.7	12.93	16.42	19.67
Number of Protein-Coding Genes	6,445	5,092	5,876	6,147	6,046	5776	4,803	9,898
GC Content (%)	38.72	40.25	37.85	37.67	39.34	37.91	54.94	55.8
Number of orthogroups containing species	4865	4482	4869	4722	4783	4818	3309	4051
Percentage of orthogroups containing species	77.3	71.2	77.3	75	76	76.5	52.6	64.3
Number of genes in species-specific orthogroups	24	7	36	145	86	15	142	2473

3.3.3 CAZyme analysis

Given the potential of *G. sulphuraria* to the biotechnology community, a detailed examination of the CAZyme repertoire of the core six genomes was performed. For comparison two other Rhodophyta species were analysed, the extremophile *C. merolae* and mesophilic *P. purpureum.* Figure 3.1 shows the distribution of CAZymes in the *G. sulphuraria* genomes and the other red algae genomes. In total, 58 putative CAZy families were identified with an average of 55 families *per G. sulphuraria* strain (Supplementary Table 7). While *C. merolae* showed 41 and *P. purpureum* 46 different families. Strain 074 had in total the most identified CAZymes (135) followed by strains 138, 033, 017, 427 and 107 (134, 128, 127, 125 and 121 CAZymes retrospectively). *C. merolae* had 92 identified CAZymes whereas *P. purpureum* contained 114 different CAZymes.

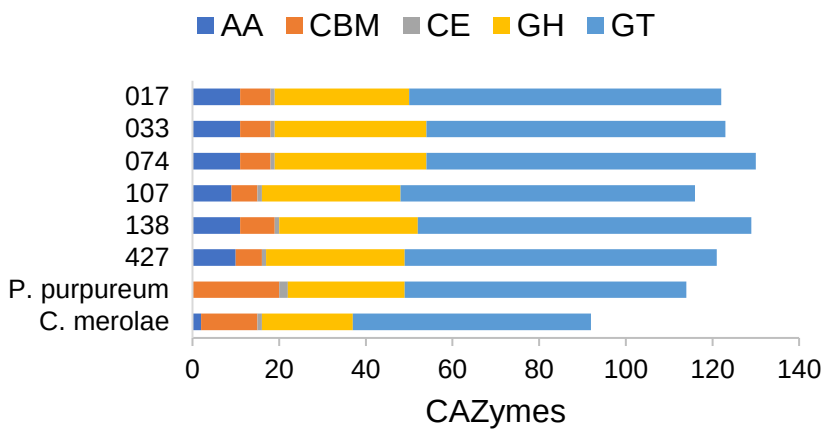


Figure 3.1: Carbohydrate-active enzymes in six *G. sulphuraria* genomes. AA, auxiliary activities; GH, glycoside hydrolase; GT, glycosyltransferase; CBM, carbohydrates-binding module; PL, polysaccharide lyase.

Glycosyl Transferases (GTs), Glycoside Hydrolases (GHs), Auxiliary Activities (AAs), Carbohydrate-Binding Modules (CBMs) and Carbohydrate Esterases (CEs) were present in all *G. sulphuraria* genomes and in *C. merolae* (Figure 3.1). *P. purpureum* did not contain any AAs. The GT family was the most abundant (34 modules on average per *G. sulphuraria* strain), followed by the GHs, AAs, CBMs and CEs modules (on average 14.5, 3, 2 and 1 per *G. sulphuraria* strain, respectively).

3.3.3.1 Glycosyltransferases (GTs)

GTs accounted for the highest proportion of identified CAZymes in *G. sulphuraria* showing on average 75.3 proteins (~59 %). CAZyme analysis revealed that across the *G. sulphuraria* genomes there were a total of 37 GTs families. Of these families over 50 % (19 families) contain on average one gene or less per genome. The majority of GT CAZymes identified in the *G. sulphuraria* genomes belong to GT4 or GT31 (Figure 3.2A). The total number of GTs in *G. sulphuraria* genomes were higher than that of the other red agal genomes (*C. merolae* with 55 GTs and *P. purpureum* with 65 GTs). *G. sulphuraria* showed higher amounts of GT4 and much lower amounts of GT39 when compared to the other genomes (Figure 3.2A).

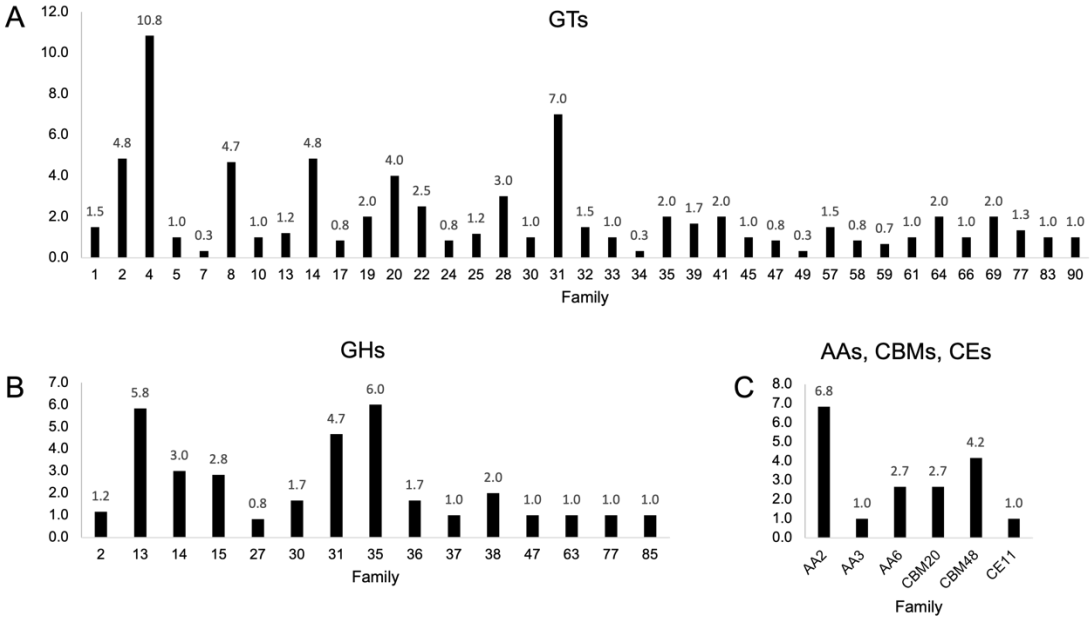


Figure 3.2: Number of CAZymes in *G. sulphuraria*. Number of (A) GT families; (B) GH families; (C) AA, CBM and CE families.

3.3.3.2 Glycoside Hydrolases (GHs)

In this study a total of 35 GHs into 15 families were predicted on average across the *G. sulphuraria* genomes (Figure 3.2B). This classification also revealed that six of these families contained on average one gene or less. Out of all the genomes analysed the total number of GHs was highest in *G. sulphuraria* then *P. purpureum* and then *C. merolae* (Figure 3.3B). GH35 family was the most prominent in *G. sulphuraria* followed by GH13 and GH31, whereas in both *C. merolae* and *P. purpureum* GH13 family had the highest predicted number of proteins. There were four GH families, GH15, GH27, GH30 and GH38 that were present in the *G. sulphuraria* genomes and not in either of

the other species analysed. Alternatively, there were two families (GH5 and GH20) that were only represented by a single gene from *P. purpureum*. GH77 family members only comprised of genes from *C. merolae* and *P. purpureum* with 2 and 3 genes retrospectively (Figure 3.3A)

3.3.3.3 Carbohydrate-Binding Modules (CBM)

Across the *G. sulphuraria* genomes in this study a total average of 6.8 CBMs were classified into two CBM families, these were CBM48 followed by CBM20 (Figure 3.2C). The comparison to the other red algal genomes showed different results, both contained more CMB genes than *G. sulphuraria* (Figure 3.3C). *C. merolae* contained a total of 13 CBMs, with a higher number of proteins in the CBM20 family as well as a single gene in CBM41. *P. purpureum* had a total of 20 CBMs in six families, there were single genes in CBM9, CBM25, CBM33 and CBM57. Along with 10 genes in CBM48 (Figure 3.3C).

3.3.3.4 Carbohydrate Esterases (CEs)

Results revealed just one CEs family represented in each of the *G. sulphuraria* genomes and *C. merolae*, CE11*. P. purpureum* also showed a single gene in family CE15 (Figure 3.2C and Figure 3.3C).

3.3.3.5 Auxiliary Activities (AAs)

The CAZyme analysis in this study also revealed that *G. sulphuraria* genomes contained 3 AA families with an average total of 10.5 AAs per genome (Figure 3.2C). AAs family classification revealed that the majority of AAs were AA2 family members followed by AA6 and AA3 (Figure 3.2C). The comparison to other red algal genomes revealed less genes in this family, *C. merolae* showed 2 AAs families (AA3 and AA6) both with a single gene in (Figure 3.3C). The *P. purpureum* genome showed no predicted AAs family CAZymes.

3.3.1 Selection of putative CAZymes for further study

The results were assessed and analysed for suitable enzymes for further investigation that are relevant for use in industrial biotechnology. Selection criteria included the presence of a predicted signal peptide and predicted hydrolase, peroxidase or unknown function. This resulted in the identification of 14 putative CAZymes of which there were six predicted to have hydrolase function, two with peroxidase function and a further six with unknow function (Table 3.4)

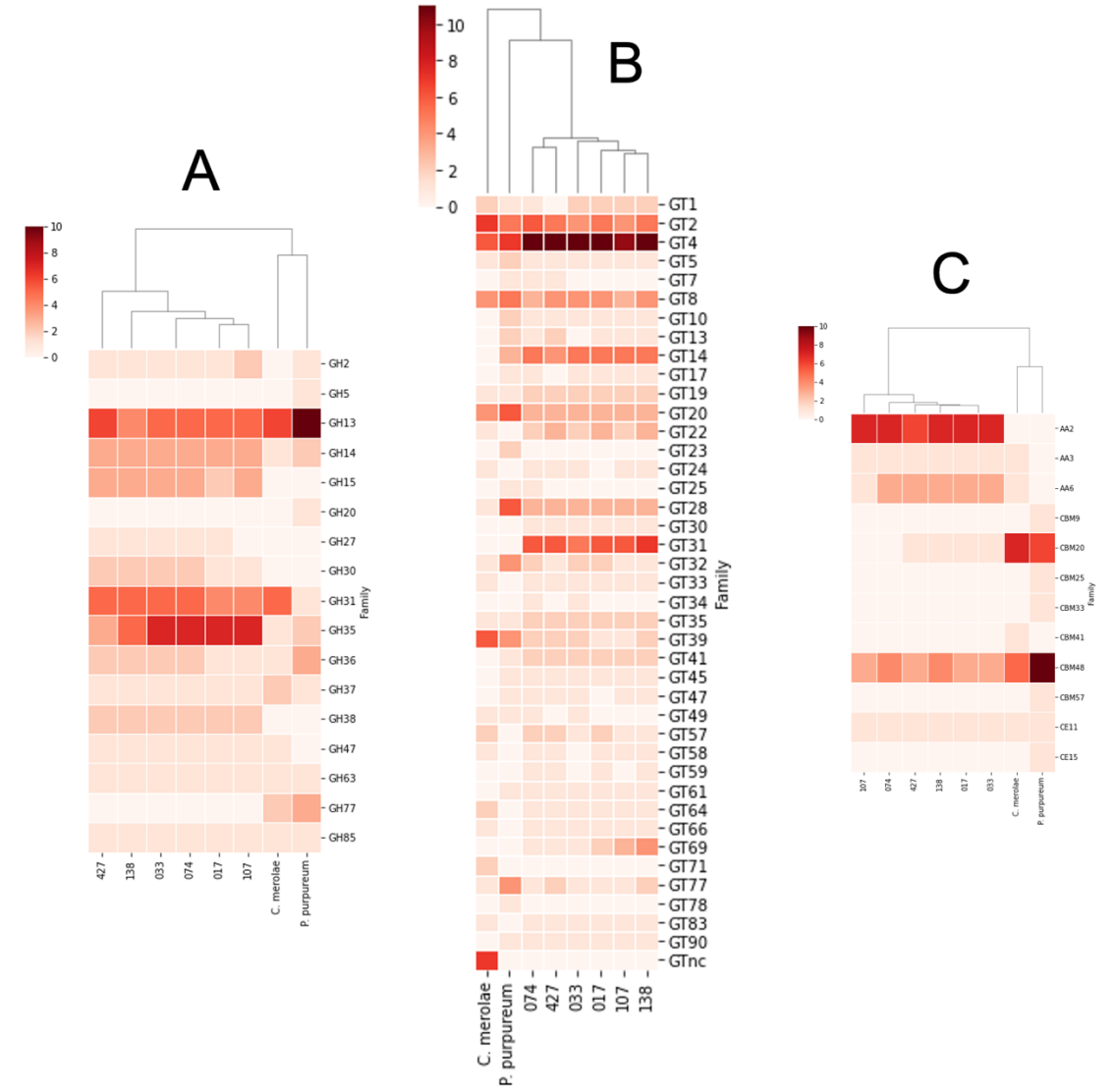


Figure 3.3: Distribution of CAZymes in *G. sulphuraria* strains 017, 033, 074, 107, 138, 427 and other red algal species, *C. merolae* and *P. purpureum*. (A) GH families; (B) GT families; (C) AA, CBM and CE families

Table 3.4: Table of final 14 putative CAZymes obtained from analysis of *G. sulphuraria* genomes. The table shows the gene ID, predicted gene ontology functions, CAZyme family identification and predicted EC number.

Gene Name	Gene Ontology	CAZyme Family	EC number
Gasu_01530	UDP-glucose:glycoprotein glucosyltransferase activity [GO:0003980]; protein glycosylation [GO:0006486]	GT24	NA
Gasu_05550	carbohydrate binding [GO:0030246]; glucan 1,3-alpha-glucosidase activity [GO:0033919]; carbohydrate metabolic process [GO:0005975]	GH31	3.2.1.84
Gasu_06640	glucan 1,4-alpha-glucosidase activity [GO:0004339]; carbohydrate metabolic process [GO:0005975]	GH15	3.2.1.3
Gasu_12000	integral component of membrane [GO:0016021]; transferase activity [GO:0016740]	GT8	NA
Gasu_17790	heme binding [GO:0020037]; metal ion binding [GO:0046872]; peroxidase activity [GO:0004601]; response to oxidative stress [GO:0006979]	AA2	1.11.1.7
Gasu_17800	heme binding [GO:0020037]; peroxidase activity [GO:0004601]; response to oxidative stress [GO:0006979]	AA2	1.11.1.7
Gasu_25530	glucan 1,4-alpha-glucosidase activity [GO:0004339]; polysaccharide metabolic process [GO:0005976]	GH15	3.2.1.3
Gasu_26360	integral component of membrane [GO:0016021]; transferase activity [GO:0016740]	GT8	NA
Gasu_27490	beta-galactosidase activity [GO:0004565]; carbohydrate metabolic process [GO:0005975]	GH35	3.2.1.23
Gasu_27500	beta-galactosidase activity [GO:0004565]; carbohydrate metabolic process [GO:0005975]	GH35	3.2.1.23
Gasu_47280	hydrolase activity, hydrolyzing O-glycosyl compounds [GO:0004553]	GH30_5	NA
Gasu_48600	catalytic activity [GO:0003824]; carbohydrate metabolic process [GO:0005975]	GH13	NA
Gasu_52030	endoplasmic reticulum quality control compartment [GO:0044322]; membrane [GO:0016020]; calcium ion binding [GO:0005509]; mannosyl-oligosaccharide 1,2-alpha-mannosidase activity [GO:0004571]; carbohydrate metabolic process [GO:0005975]; endoplasmic reticulum mannose trimming [GO:1904380]; mannose trimming involved in glycoprotein ERAD pathway [GO:1904382]	GH47	3.2.1.-
Gasu_64540	integral component of membrane [GO:0016021]; transferase activity, transferring glycosyl groups [GO:0016757]	GT34	NA

3.4 Discussion

3.4.1 Genome sequencing

G. sulphuraria is a unique organism that can grow both heterotrophically and autotrophically whilst in extremely acidic conditions under high temperature (for a eukaryote). Based on this it holds enormous promise as an industrial biotechnological resource. In order to delve into the multitude of enzymes present across the *G. sulphuraria* species it was a must to achieve high quality sequence data. To generate this data a non-trivial approach to extracting long sequence DNA was developed, this successfully allowed for the generation of long reads using Oxford Nanopore Technologies (ONT) (Goodwin et al., 2015). Typically, a reasonable read coverage for a de novo genome-sequencing project would be consider in the 50-60X range. This coverage allows for sufficient reads that uniquely anchor the longest repeat regions in the genome assembly (Lu et al., 2016). The coverage produced in this study far exceeded this and lead to the data producing accurate and essentially complete genomes.

The number of chromosomes in *G. sulphuraria* had previously been determined by pulse-field gel electrophoresis to be 40, ranging in size from 100 – 420 kb (Moreira et al., 1994). Later Contour-Clamped Homogenous Electric Field (CHEF) analysis supported this and it was generally accepted that characteristically *Galdieria* contained a large number of small chromosomes (Moreira et al., 1994; Takahara et al., 1999). The results of assemblies in Table 3.3 shows evidence consistent with *Galdieria* having a large number of small chromosomes with the average contig lengths showing less than 150 kb. However, the assemblies shown here indicate a chromosome number much higher, 107 shows evidence for this number to be closer to 74. It is likely that many of these chromosomes are similar sizes so would have been indistinguishable when using pulse-field gel electrophoresis. Alongside estimating the number of chromosomes in 1994 Moreira et al also estimated genome size. This was long before the sensitivity of sequencing available today and this resulted in an estimation of *G. sulphuraria* genome being 9.8 Mb which is vastly different to the 13.7-19.7 Mb genomes shown in this study (Table 3.3). These accurate and near complete genomes mean they can be used for further investigations into *G. sulphuraria* not only in this these but for any future research.

3.4.2 Orthologues

Ortholog classification can highlight evolutionary relationships from diverging speciation events. Orthologous genes known commonly as ‘same gene different species’, is referring to genes that originated from a common ancestor which then underwent a diverging specification event (Setubal and Stadler, 2018). These genes are then usually syntenic between species that are closely related. If high sequence similarity is shown between orthologous genes in multiple species, there is a high likelihood that those gene will continue to perform similar biological functions (Emms and Kelly, 2015; 2019). To identify unique and/or shared gene families between the *G. sulphuraria* genomes orthologous clustering of each of their predicted proteomes was performed. By comparing the orthologous proteins, we could link gene families and infer potential differences between strains. As expected in the majority of clusters there was an orthologous relationship between the six clades, with upward of 70 % of clusters containing each *G. sulphuraria* species. The function of these were mostly assigned to the cellular metabolic process implying a conserved role in fundamental biological processes, which is expected. However, it is shown that across all the lineages there are differences with every strain containing a number of singleton protein sequences, meaning that these could not be found in any other genome. These singletons show the diversity across the species and how different populations evolve differently and uniquely. Determining orthologs is a crucial step in comparatively looking at the CAZymes present in *G. sulphuraria* species.

3.4.3 Glycosyltransferases (GTs)

GTs catalyse the formation of glycosidic linkages to form glycosides, which are involved in the creation of a diverse range of polysaccharides, oligosaccharides, and glycoconjugates (Breton et al., 2006; Lairson et al., 2008). The reaction involves the transfer of activated forms of monosaccharides to a saccharide, protein, lipid, DNA or small molecule acceptor to form the glycosidic bonds (Breton et al., 2006; Lairson et al., 2008). GTs constitute one of the largest family of CAZymes they currently have been classified into 114 different subfamilies, the *G. sulphuraria* genomes contain 37 of these different GT families. Among the GTs families represented seven families; GT4, GT31, GT2, GT14, GT8, GT20 and GT28, account for over half the total number of GTs predicted on average. One of the largest GT families is GT4 containing not only CAZymes that utilise nucleotide sugar donors but also simple phospho and lipid-phospho sugar doners, this diversity is reflected in not just these enzymes sequences but also their potential functions (Martinez-Fleites et al., 2006). Many of the GT families shown in

Galdieria have links to sugar and cellulose synthesis (GT2, GT4, GT8, GT14, GT28, GT31) however characterisation of these enzymes is difficult (Aspeborg et al., 2005; Stone et al., 2018). Typically, as in plants these CAZymes are membrane bound thus making the isolation and characterisation difficult. Confirmed though this genome-wide comparison, GT31 family was prominent across the *G. sulphuraria* genomes, but not present in the other red algae, this suggests that the GT31 family is a not major component of GT families in red algae, but rather unique to *G. sulphuraria*. It is difficult to say with certainty the function of a particular GT gene, this is the result of GTs being classified into families based on amino acid sequence. With many different GTs having many different functions means that putative function is hard to predict using sequence similarity alone (Breton et al., 2012).

GTs found in these genomes are involved many processes including cell wall biosynthesis. It is known the number of GTs found in Rhodophyta are generally much lower than in land plants which possess complex ridged cell wall structures (Ulvskov et al., 2013). Notably, though not surprising the number of GTs is highest in *Galdieria* then *P. purpureum* and lastly *C. merolae*, this is likely a reflection of the complexity of their respective cell walls. In red microalga the cell walls lack the cellulose microfibrillar component seen in their multicellular counterparts and instead are often encapsulated within a gel sulphated polysaccharide. These cell wall polysaccharides equip the cells with environmental protection to withstand such factors as desiccation, temperature stability, pH and salinity (Arad, 1988; Arad and Levy-Ontman; 2010). *C. merolae* lacks a cell wall, this could explain the lower number of GTs observed in its genome. However, *Galdieria* has a ridged cell wall that is able to withstand the proton gradient attached to the internal pH 7 against external ~pH 2 (Oesterhelt et al., 2007). Sealing the cell wall against an intrusion of H^+ is the only way to accomplish maintaining the inward acting H^+ gradient of $1{:}1 \times 10^5$ (Enami et al., 1986), though this is still poorly understood. Though this selection of GT CAZymes likely are not involved in the degradation of lignocellulose they present an interesting question as to whether their function hold part of the key to their ability to withstand such harsh environments.

3.4.4 Glycoside Hydrolases (GHs)

GHs are the enzymes responsible for catalysing the hydrolysis of glycosidic bonds of complex carbohydrates (Bourne and Henrissat, 2001; Naumoff, 2011). GHs are essentially found in all domains of life and represent an important collection of enzymes involved in the degradation of carbohydrates, namely they assist in the breakdown of lignocellulosic biomass (cellulose, hemicellulose and starch) (Cragg et al., 2015;

Berlemont and Martiny, 2016; Ezeilo et al., 2017). GHs form the largest enzyme class in the CAZyme database comprising at current of 172 families of which this analysis revealed *G. sulphuraria* is represented by just 15 families. The variation of hydrolytic activities from this class is large. Activity on polysaccharides can be either endo- or exo-acting, this refers to the way in which the enzyme cleaves the polysaccharide chain. This is either at a random mid chain point (endo acting) or from the end of the chain (exo-acting) (Bourne and Henrissat, 2001; Andlar et al., 2018). Additionally, GHs are sometimes assisted by polysaccharide esterases that will remove methyl, acetyl and phenolic esters making way for the GHs to be able to function on the rest of the polysaccharide chain (Andlar et al., 2018; Suleiman et al., 2020).

Analysis showed numerous putative alpha-glucosidases, alpha-galactosidases and alpha-L-arabinofuranosidase B among the families represented in *G. sulphuraria.* The family containing the most genes, GH35 consists almost exclusively of beta-galactosidases, these enzymes specifically catalyse the hydrolysis of the glycosidic bond in beta-galactoside into its monosaccharides (Asraf and Gunasekaran, 2010). The GHs present in the *G. sulphuraria* genomes shown an abundance of enzymes that cleave nonreducing carbohydrates in oligosaccharides and the side chains of hemicelluloses and pectins acting on starch and glycogen (GH2, GH13, GH31, GH35) (Nguyen et al., 2018; Yang et al., 2021). Analysis also revealed a number of enzymes with unknow function but containing highly conserved single GH domains. These results highlight the lack of any predicted xylanases or cellulases that are essential in degradation of lignocellulosic material (Ezeilo et al., 2017). It could be that these predicted GHs are multifunctional or contain overlapping multiple domains and therefore traditional methods of classification via sequence homology are underestimating the ability of these enzymes. It is possible that the enzymes with unknow function could be new types of xylanases or cellulases that are previously unseen. Given the uniqueness and growth diversity shown by *G. sulphuraria* it is surprising that there are so few GHs present in its genome and suggests that the enzymes it harbours may be unlike anything seen before and that there is still much to discover.

3.4.5 Carbohydrate-Binding Modules (CBM)

CBMs represent a large group of protein domains where the amino acid sequence has carbohydrate binding activity (Lombard et al., 2014). They themselves have no catalytic activity but bind to carbohydrate ligands, they are most commonly found attached to GH enzymes (Lombard et al., 2014; Sidar et al., 2020). It has been acknowledged that the binding of a CBM enhances the catalytic efficiency of a CAZyme, this is achieved through

aiding in targeting the CAZyme to the substrate as well as disrupting the crystallinity of any insoluble portion of the substrate (Reyes-Ortiz et al., 2013; Bernardes et al., 2019; Sidar et al., 2020). Of the 88 families of CBMs classified in the CAZy database *G. sulphuraria* genomes only identified CBM20 and CBM48. CBM20 has strong links to starch-binding and in this case connected with catalytic domains in GH77s, the β-amylase family. CBM20 enzymes have been identified in bacterial β-amylases, this could indicate the horizontal gene transfer of such enzymes into *G. sulphuraria* (Christiansen et al., 2009; Janeček et al., 2019). CBM48 is often appended to GH13 modules, pullulanse subfamily proteins and the beta-subunit of AMP activated protein kinase. It has been established this module contain putative starch binding domains and is predicted to facilitate cytosolic starch-binding interactions in red algae, hence explaining its presence in all genomes (Janeček et al., 2011; Janeček et al., 2019). The presence of CBMs families identified in conjunction with other CAZymes suggests that those CAZymes require the CMBs in order to efficiently degrade substrates.

3.4.6 Carbohydrate Esterases (CEs)

Esterases, are hydrolytic enzymes that act on ester bonds, they are widely used in industrial process and biotechnology as biocatalysts (Nakamura et al., 2017; Armendáriz-Ruiz et al., 2018). CEs are a class of esterases, they typically catalyse the *O*-de- or *N*-deacylation to remove esters of substituted saccharides (Lombard et al., 2014; Armendáriz-Ruiz et al., 2018). CEs are currently classified into 19 families that show a large diversity in substrate specificity, such as xylan, acetic ester, chitin, peptidoglycan, feruloyl-polysaccharide and pectin (Biely, 2012; Nakamura et al., 2017). Given the growth capacity of *G. sulphuraria* is it surprising that only one CE was uncovered (CE11). CE11 is a zinc protein that is involved in the biosynthesis of Lipid A, which is a component of endotoxin. Endotoxin is a component of the exterior cell wall of gram negative bacteria and is responsible for the bacteria's toxicity (McClure et al., 2003). This only furthers the question of how exactly is this organism growing on such an array of carbohydrates.

3.4.7 Auxiliary Activities (AAs)

Lignin possesses great potential as a high value compound however, due to its recalcitrant nature it is difficult to breakdown. Therefore, arguably the most sought after CAZyme families AAs contain lignin degradation enzymes. This family was launched after it was highlighted that lignin is invariably found in the plant cell walls together with polysaccharides. Thus, is expected that lignin fragments are likely to act together with

lytic polysaccharide mono-oxygenases (LPMO). Creation of the AA families allowed for the accommodation of the full range of enzyme mechanisms and substrates that was otherwise difficult to fit into the previously defined families (Levasseu et al., 2013). Currently at 17 classified families, they contain members that are predominantly associated with the depolymerization of lignin (non-carbohydrate structural components). With nine families of ligninolytic enzymes along with seven families of LPMO (Rytioja et al., 2014; Park et al., 2018).

This analysis revealed that family AA2 was the most abundant of the AAs, also this family notably was not represented in either *C. merolae* or *P. purpureum* making interesting targets for investigation as they could be relevant in *G. sulphuraria* metabolic abilities in harsh environments. The AA2 family interestingly includes the plant peroxidase superfamily which contains known enzymes such as manganese peroxidase, lignin peroxidase and versatile peroxidase (Fawal et al., 2012; Levasseu et al., 2013). Upon further inspection the genes present in the *G. sulphuraria* genomes are one cytochrome c peroxidase, two class I ascorbate peroxidases and a small family of class III peroxidases. These class III peroxidases are suspected to be involved in cell wall modification and one purified enzyme from *G. sulphuraria* (Pxr04) was shown to be heat and acid stable, though no function was revealed (Oesterhelt et al., 2008).

3.4.8 Conclusion

This chapter aimed to advance the knowledge of the CAZyme repertoire within the *G. sulphuraria* species and aid in understanding its versatile metabolic abilities. Additionally, the identification of any novel CAZymes produced by *G. sulphuraria* that could be involved in lignocellulosic degradation. In this chapter sequencing of the core six genomes was used to identify any enzymes potentially involved in lignocellulosic biomass degradation. As described above, many CAZymes were identified including 75.3 GTs, 34.6 GHs, 1 CE, 6.8 CMBs and 10.5 AAs on average across the six genomes. Though many presented were involved in putative degradation of polysaccharides they appeared to be acting on side chains, there were no enzymes acting solely on lignin, cellulose or hemicellulose chains themselves.

The interest of this study is focused on CAZymes that are involved in lignocellulosic degradation. Although analysis did not uncover any lignin modifying peroxidases, xylanases or cellulases there is still great potential in GHs and AAs for their use in biotechnological and industrial applications due to the expected heat and acid tolerant nature of any enzymes. For example, any products produced by these enzymes could

be used to in various industries involving the generation of bio-based products, these include paper, food and textile industries additionally any other chemicals potentially having a use in the production of biofuel. Though the number of CAZyme are overall unexpected given the capability of *G. sulphuraria* to grown on numerous carbon sources. Further detailed investigation of the novel CAZymes is required along with experimental work to identify genes involved in the growth of *G. sulphuraria* on substrates that perhaps have not registered as CAZymes. These unknow genes could hold huge potential in the discovery of unseen unique enzymes involved in lignocellulosic degradation. It is important to look further into these putative enzymes that are of high interest for similarity to know activities from other enzymes. Also identifying specific secretomes of *G. sulphuraria* grown under individual substrates will provide an even more relevant set of enzymes to investigate.

Chapter 4 - Growth Experiments and Secreted Protein Identification Studies

4.1 Introduction

4.1.1 Lignocellulosic biomass

Representing one of the most promising carbon-neutral alternatives to fossil fuels, nonedible lignocellulosic biomass is the most abundant and underexploited renewable organic carbon source across the globe (Limayem and Ricke, 2012; Strazzulli et al., 2020). This material provides an ideal raw material for the production of chemicals and second generation biofuels (Li et al., 2014; Strazzulli et al., 2020). First generation biofuels that are made mainly from edible feedstocks such as sugar, starch, and vegetable oil are limited in their scope, as there is a threshold at which they cannot produce a high enough yield without threatening biodiversity and food security. This development from first generation biofuels, provides an environmentally beneficial way of making biofuels, without the potential negative impact on food security (Dahman et al., 2019).

Lignocellulosic material is primarily composed of three polymers: lignin, cellulose and hemicellulose (Figure 4.1). Additionally, there are lesser amounts of other components such as proteins, pectin and water (Baruah et al., 2018). The specific composition of these major components varies depending on the plant species, but typically the lignin, cellulose and hemicellulose contents fall within 15-40 %, 40-60 % and 20-40 % respectively (Dahadha et al., 2017; Wang et al., 2017). These polymers are rigidly bound making up an elaborate composite structure. The lignocellulosic matrix is formed through a series of non-covalent bonds as well as covalent cross-linkages between the polymer groups (Baruah et al., 2018; Zoghlami and Paës, 2019).

Straw from farming of cereal crops such as rice, wheat and maize represent a large proportion of the non-edible lignocellulosic biomass produced by farming, yet are not well utilised (Glithero et al., 2013). Wheat is one of the major crops in the UK, with an annual estimated wheat straw yield of 8-10 million tonnes. Wheat straw typically contains 13-15 % lignin, 37-41 % cellulose and 27-32 % hemicellulose. Hence, wheat straw biomass presents as an appealing feedstock to produce second-generation biofuel (Wang et al., 2013, Tian et al., 2018, Raud et al., 2019).

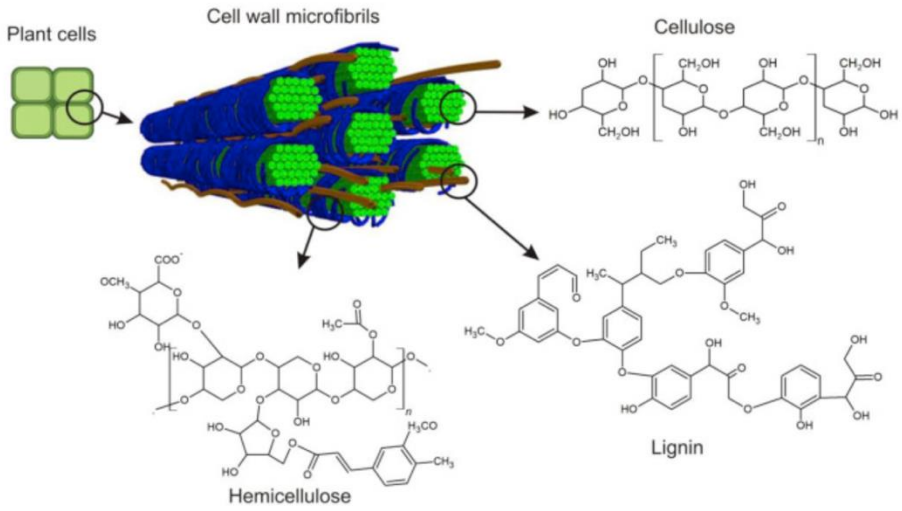


Figure 4.1: Structural components of lignocellulosic biomass. Showing the composition and interaction of cellulose, hemicellulose and lignin in the plant cell wall. Adapted from Raud et al., 2019.

4.1.2 Biofuel production

Usually, to produce biofuel from lignocellulosic biomass, the polysaccharides in the biomass must first be converted into sugars. These sugars are then ready to be fermented into fuel (Raud et al., 2019). The recalcitrant nature of lignocellulosic biomass and the tight packaging of hemicellulose covering the cellulose and lignin this leads to difficulties in accessing material ready for conversion into fermentable sugars, thus creating a bottleneck in the use of this feedstock (Akhtar et al., 2016). To overcome this, it is necessary for the lignin, cellulose and hemicellulose structure to be broken-down/weakened to make polysaccharides easily accessible for further processing. Currently, the pre-treatment techniques used to achieve this can be categorised as either chemical, physical, biological or physicochemical processes (Baruah et al., 2018; Zoghlami and Paës, 2019; Raud et al., 2019).

It is standard practice to use physical pre-treatment alongside any of the other pre-treatment methods. The aim is to increase the surface area of the material, while simultaneously decreasing the degree of polymerisation and crystallinity. This is achieved by reducing the particle size through methods such as milling and ultrasonication (Rajendran et al., 2017; Baruah et al., 2018). Chemical pre- treatment methods, use different chemicals to break down the lignocellulosic biomass structure such as organic solvents or acid and alkaline reagents (Baruah et al., 2018). In

physiochemical pre-treatment, changing conditions such as temperature and pressure can disturb the lignocellulosic structure, allowing the lignin and hemicellulose to be separated (Shirkavand et al., 2016; Raud et al., 2019). For example, in steam explosion, water molecules at high pressure and temperature (0.69–4.83 MPa, 160–260 °C) are able to penetrate the biomass structure. The water molecules are then allowed to escape in an explosive way by suddenly reducing the pressure, thus, disturbing the lignocellulosic matrix (Baruah et al., 2018; Raud et al., 2019). Lastly, biological methods rely on different microorganisms such as fungi and bacteria to degrade and weaken the components of lignocellulosic biomass. For example, different types of fungi (white, brown and soft rot) are known to use enzymes to degrade lignocellulosic biomass (Sindhu et al., 2016; Raud et al., 2019). Hydrolytic enzymes are responsible for the breakdown of both cellulose and hemicelluloses, whilst ligninolytic enzymes depolymerises lignin. The key principle of each of these methods is that they separate the distinctive components pertaining to lignocellulosic material efficiently and selectively (Raud et al., 2019).

After pre-treatment resulting material can be used for further processing to increase conversion rates. Often a combination of pre-treatment methods are used. The next steps in production can include further thermochemical processing which involve heating the biomass up to 800° C, either via gasification (with oxygen) or pyrolysis (absence of oxygen). Additionally, a biochemical route can be taken where the now available polysaccharides are converted into fermentable sugars using hydrolytic enzymes (Dos Santos et al., 2019; Raud et al., 2019).

4.1.3 Enzymatic breakdown of lignocellulosic biomass

Enzymatic hydrolysis of lignocellulosic material requires a variety of specific enzyme functions. Enzymatic degradation of lignin is attributed to two families of ligninolytic enzymes, these are phenol oxidase or laccase and peroxidases (Li et al., 2019; Kinnunen et al., 2019; Raud et al., 2019). Alternatively cellulose degradation requires at least the coordinated activity of no less than three enzymes: endo-β-glucanase, exo-β-glucanase and β-glucosidase. In order to decrease the length of the cellulose chain Endoglucanases act to randomly hydrolyse internal β-1,4-glycosidic bonds, to then split off the cellobiose from the shortened cellulose chain exo-β-glucanases act, then to hydrolyse the cellobiose into glucose requires β-glucosidases (Wang et al., 2011; Houfani et al., 2020). However, other activities are usually employed by cellulolytic organisms, including LPMOs that oxidatively attack the crystalline regions of cellulose

(Dutta and Wu, 2014; Eibinger et al., 2017) and expansin-like proteins such as swollenin that help disrupt cellulose structure (Arantes and Saddler, 2013; Gourlay et al., 2013).

Several classes of enzymes are essential for the effective breakdown of hemicellulose (Houfani et al., 2020). Such examples of these enzymes are CAZymes glycoside hydrolases (GHs), carbohydrate esterases (CEs), polysaccharide lyases (PLs) and endo-hemicellulases. These enzymes contribute to the breakdown of hemicellulose by their collective actions in which glycosidic bonds, ester bonds are hydrolysed, and side chains removed (Piccinni et al., 2019, Houfani et al., 2020). These include α-L-arabinofuranosidase, acetylxylan esterase, α-glucuronidase, β-mannosidase, β-mannanase, β-xylosidase and endo-1,4-β-xylanase (Zoghlami and Paës, 2019; Piccinni et al., 2019; Houfani et al., 2020).

4.1.4 Extremozymes

Enzymes from extremophilic microorganisms, coined extremozymes, have generated a lot of interest in their application to industrial biotechnology due to their ability to function in inhospitable environments where their mesophilic counterparts would quickly denature (Strazzulli et al., 2020). Extremozymes produced by these extremophiles show remarkable tolerance to extreme temperatures, extreme pH, high salt and detergents (Dumorne et al., 2017; Chettri et al., 2021). Thus, they are an ideal tool for various industrial applications, in particular those involving harsh physical and chemical conditions. An excellent example of this is the conversion of lignocellulosic material for use in the biofuels market or the pulp and paper industry (Dumorne et al., 2017; Chettri et al., 2021).

4.1.4.1 *Galdieria*: A source of Extremozymes

Galdieria sulphuraria is an extremophile eukaryotic unicellular red alga living in geothermal sites where the ecological conditions are very extreme, such as areas of low pH (0-4) and high temperatures (50-56°C) (Gross and Schnarrenberger 1995; Ciniglia et al., 2004; Yoon et al., 2006; Pinto 2007). It is this ability to live in these extremes that make *G. sulphuraria* an interesting target for potential uses in industrial biotechnology. Alongside this *G. sulphuraria* has the unique capability of both photoautotrophic, heterotrophic and mixotrophic growth (Barbier et al., 2005) suggesting that it harbours an extensive array of enzymes. In addition, *G. sulphuraria* has a vast array of metabolic properties that allow it to grow vigorously on a wide range of carbon sources, as well as displaying high tolerance to heavy metals (Gross and Oesterhelt, 1999; Jain et al., 2014). Over 50 carbon sources have currently been identified as supporting its growth including

several sugars, sugar alcohols, amino acids, and organic acids (Schönknecht *et al*., 2013; Qiu et al., 2013). The biochemical versatility within this alga reveals a large repertoire of metabolic enzymes which are a rich source of thermostable proteins for biotechnology (Chae et al., 2014).

These unique characteristics and metabolic flexibility of *G. sulphuraria* make it an attractive candidate to discover novel lignocellulose degrading enzymes that would be thermo and acid tolerant. Understanding its capacity to degrade lignocellulosic biomass could facilitate the development of more effective strategies in breaking down and utilising this feedstock, these enzymes should display characteristics that would potentially aid in the viable and efficient production of bioethanol from lignocellulosic material.

4.1.5 Aims

In the previous chapter *Galdieria* genomes were searched informatically for CAZymes that could be involved with the degradation of lignocellulosic biomass. Though these results were informative, the lower than expected number of CAZymes, showed gaps in knowledge where experimental data could add information and understanding. This chapter will explain the process of identifying and selecting putative lignocellulosic acting enzymes produced from *G. sulphuraria* using both informatic and proteomics techniques. This chapter will focus on preparation of optimal substrate to produce supernatant samples containing proteins for proteomics analysis. Using mass spectrometry and informatic analysis for discovery of interesting target genes.

4.2 Materials and Methods

4.2.1 Algal growth under different substrates

Each culture was inoculated from the *G. sulphuraria* stocks as described in Section 2.2.1. Subcultures were centrifuged, any media discarded and washed three times in PBS pH 7.4 before resuspending in the media relevant to each experiment and condition. Cultures were grown heterotrophically, in the dark at 37°C and on an orbital shaker at 130 rpm.

4.2.1.1 Experiment 1

For each strain (107 and 427), nine vessels were prepared with Allen medium titrated to pH 2 (Allen and Stanier, 1968; Table 2.1): 2 x 10g/L of Sucrose (Sigma-Aldrich), 2 x 10

g/L 5 mm milled Wheat straw, 2 x 10 g/L Flour, 2 x 2.5 g/L freeze dried *Chlamydomonas reinhardtii* and 1 x medium only. These were then then autoclaved for sterilisation. For each substrate type one vessel contained medium with no cells to allow for comparison of substrate degradation. All vessels were incubated together.

4.2.1.2 Experiment 2

For *G. sulphuraria* 074W seven separate cultures were prepared with Allen medium titrated to pH 2 (Allen and Stanier, 1968; Table 2.1) with following substances as a growth substrate then autoclaved for sterilisation: 10 g/L of Sucrose (Sigma-Aldrich), 10 g/L Lignin, alkali (CAS Number:8068-05-1, Sigma-Aldrich), 10 g/L Carboxymethylcellulose sodium salt (CMC Cellulose, CAS Number:9004-32-4, Sigma-Aldrich) and 3 x 20 g/L Xylan from Corn Core (CAS Number:9014-63-4, Tokyo Chemical Industry UK Ltd). A culture of no additional substrate was also prepared as a negative control.

4.2.1.3 Experiment 3

Each *G. sulphuraria* strain, (017, 033, 074, 107, 138 and 427) was prepared in Allen medium titrated to pH 2 (Allen and Stanier, 1968, Table 2.1) then supplemented with either 2 % (w/v) Xylan from Corn Core (CAS Number:9014-63-4, Tokyo Chemical Industry UK Ltd), D-(+)-Xylose (CAS Number:58-86-6, Sigma-Aldrich), or Sucrose (CAS Number:57-50-1, Sigma-Aldrich). A culture of no additional was also prepared as a control for each strain.

4.2.1.4 Experiment 4

G. sulphuraria strain 074 was prepared in Allen medium titrated to pH 2 (Allen and Stanier, 1968, Table 1) then supplemented with either 2 % (w/v) xylan from Corn Core (CAS Number:9014-63-4, Tokyo Chemical Industry UK Ltd) or 0.5 % (w/v) Sucrose (CAS Number:57-50-1, Sigma-Aldrich). A culture of Allen medium titrated to pH 2 was also prepared as a control. Each condition had three replicates.

4.2.2 Quantification of Biomass

4.2.2.1 Experiment 1

Cell growth was monitored over the experiment for three weeks through a daily cell count using a haemocytometer and light microscope in order to assess growth of the cultures on each of the given substrates.

4.2.2.2 Experiment 2, 3 & 4

Optical density (OD) was measured on the day of inoculation and every ~1-3 days for 21 days then every 10 days until 41 days using a UV/Vis Spectrophotometer (Biochrom Ltd. Libra S12) at 800 nm (OD_{800}), advised by the wavelength scan and in accordance with previous investigations using 074W (Gross and Schnarrenberger 1995; Oesterhelt and Gross 2002).

4.2.3 Scanning Electron microscopy (SEM)

A 5ml sample was taken from the wheat straw culture, cell-free Wheat straw culture from Experiment 1. Samples were centrifuged at 3900 rpm for 15 minutes to pellet substrate and cells, and the supernatant discarded. Pelleted substrate and cells were then prepared and mounted for scanning electron microscopy (SEM). Images of cells and substrates were then taken at between 250x and 3000x magnifications to allow assessment of cell morphology and substrate degradation.

4.2.4 Sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE)

4.2.4.1 Preparation of supernatant samples - TCA

Supernatant was harvested from the cultures by centrifugation at 3900 rpm for 10 minutes. 1 mL of supernatant was added to 8 mL of ice-cold acetone and 1 mL of 0.2 % Dithiothreitol (DTT) in 20% Trichloroacetic Acid (TCA). Solutions were incubated at -20°C for 1 hour to promote precipitation. Precipitated solutions were centrifuged at 18,000 g for 15 minutes at 4°C and pellets subjected to three cycles of washing in ice cold acetone with 0.2 % DTT. Concentrated supernatant protein pellets were dried at room temperature. SDS PAGE gels were used to analyse proteins present in the supernatant of culture grown on different substrates as described in Table 4.1.

TCA precipitated protein pellets were dissolved in 50 µl of 2X Laemmli sample buffer with β-mercaptoethanol and then heated at 95°C for 5 minutes (Laemmli, 1970). Samples were centrifuged at 18,000 g for 10 minutes and 20 µl of supernatant loaded onto an SDS-PAGE gel. In addition, 5 µl of PageRuler™ Prestained Protein Ladder (Catalog number: 26617) was used.

Table 4.1: Composition of SDS-PAGE used.

Component	Stacking gel (4%)	Resolving gel (10%)
30 % bis-Acrylamide	1.32 mL	5 mL
0.5 M Tris-HCl pH6.8	2.52 mL	----
1.5 M Tris-HCl pH 8.8	----	3.75 mL
10 % SDS	100 µL	150 µL
diH_2O	6 mL	6 mL
TEMED	10 µL	75 µL
10 % APS	50 µL	75 µL

SDS PAGE gels were run using running buffer (25 mM Tris, 193 mM Glycine and 0.1 % SDS). Gels were stained using staining solution (40% (v/v) ethanol, 10% (v/v) acetic acid and 0.1% (w/v) Coomassie® Brilliant Blue R-250 in deionised water) prepared according to Thermo Fisher Scientific (2019). Gels were de-stained at room temperature overnight in 10% (v/v) ethanol and 7.5% (v/v) acetic acid in deionised water.

4.2.5 Mass Spectrometry analysis

The aim was to identify which proteins were likely play a key role in degradation along with suitability for use in industrial setting, to move forward with to cloning. The Liquid chromatography–mass spectrometry (LC MS/MS) of the 074W xylan grown secretome described in this section was performed by Dr. Jagroop Pandhal, University of Sheffield, who provided the following description. The gel in Figure 4.9 was given to Dr Jagroop Pandhal where a trypsin digest was performed on the total protein precipitates.

LC MS/MS was performed and analysed by nano-flow liquid chromatography (U3000 RSLCnano, Thermo Scientific) coupled to a hybrid quadrupole-orbitrap mass spectrometer (Q Exactive HF, Thermo Scientific). Peptides were separated on an Easy-Spray C18 column (75 µm x 50 cm) using a 2-step gradient from 3 % solvent A (0.1 % formic acid in water) to 50 % solvent B (0.1 % formic acid in 80% acetronitrile) over 30 at 300 nL min-1. The mass spectrometer was programmed for data dependent acquisition with 10 product ion scans (resolution 30,000, automatic gain control 1e5, maximum injection time 60 ms, isolation window 1.2 Th, normalised collision energy 27, intensity threshold 3.3e4) per full MS scan (resolution 120,000, automatic gain control 1e6, maximum injection time 60ms) with a 20 second exclusion time.

4.2.5.1 Database searching

MaxQuant (version 1.5.2.8) software was used for database searching with the *.raw MS data file using standard settings. The data for searched against the *Galdieria sulphuraria* Uniprot proteome database using the following settings: Digestion type: trypsin; Variable modifications: Acetyl (Protein N-term); Oxidation (M); fixed modifications: carbamidomethyl (C); MS scan type: MS2; PSM FDR 0.01; Protein FDR 0.01; Site FDR 0.01; MS tolerance 0.2 Da; MS/MS tolerance 0.2 Da; min peptide length 7; max peptide length 4600; max mis-cleavages 2; min number of peptides 1.

Proteins identified in the secretome sample by LC-MS were analysed for the presence of a signal peptide, using SignalP5.0 (Armenteros et al., 2019). For each protein the EC number, GC %, exon count, molecular weight, isoelectric point (pI), gene ontology (GO), conserved domains and sequence homology identified using the UniProt Knowledge Base and BLASTp (Altschul et al., 1990, UniProt Consortium 2019). Additionally, where mentioned protein sequences were search against the protein data bank (https://www.rcsb.org/) as well predicted protein structures. I-TASSER (Iterative Threading ASSEmbly Refinement) was used for protein structure prediction (Yang et al., 2015; Yang et al., 2015; Zheng et al., 2021) alongside AlphaFold v2.1.0 Colab (https://alphafold.ebi.ac.uk/).

4.2.6 Cell lysate LC-MS

Cells from triplicate 074W cultures grown mixotrophically under light with 0.5 % (w/v) sucrose for 10 days were harvested by centrifugation. Cell pellets were washed three times in PBS pH 7.4 and were resuspended in 500 µl of PBS pH 7.4 and 80 µl of NuPageTM sample buffer (InvitrogenTM) with protease inhibitor (cOmpleteTM, Roche). To lyse cells single replicates were either frozen with liquid nitrogen and grinded by drilling (IKA® RW16), subjected to bead beating for 7 minutes at full power with 100 1 mm silica beads (Qiagen MM300 TissueLyser) or sonicated at 100% power for 30 seconds (Bandelin SonoPuls). Solutions of lysed cells were centrifuged at 13.2 rpm for 1 minute and 40 µl of supernatant was loaded onto an SDS-PAGE gel and ran until protein samples had moved into the resolving gel. The gel was washed in distilled water and submerged in SimplyBlueTM Safe Stain (Thermo Fisher Scientific) overnight. The band containing all protein was excised from the gel and sent for trypsinolysis and LC-MS/MS analysis, this was carried out at the Bioscience Technology Facility, University of York. With these results peptides were mapped back to the *Galdieria sulphuraria* Uniprot proteome database and genes identified.

4.3 Results

4.3.1 Experiment 1

Growth Experiment 1 was carried out initially to investigate *Galdieria's* ability to grow on different substrates. During Experiment 1, *G. sulphuraria* strains 107 and 427 both successfully grew on the four substrates tested. The two strains showed differences in growth relative to each other and across substrates. With all substrates 107 showed a higher cell density when compared to 427. In both strains Flour, Sucrose, *C. reinhardtii,* then wheat straw best supported growth (Figure 4.2 and Figure 4.3).

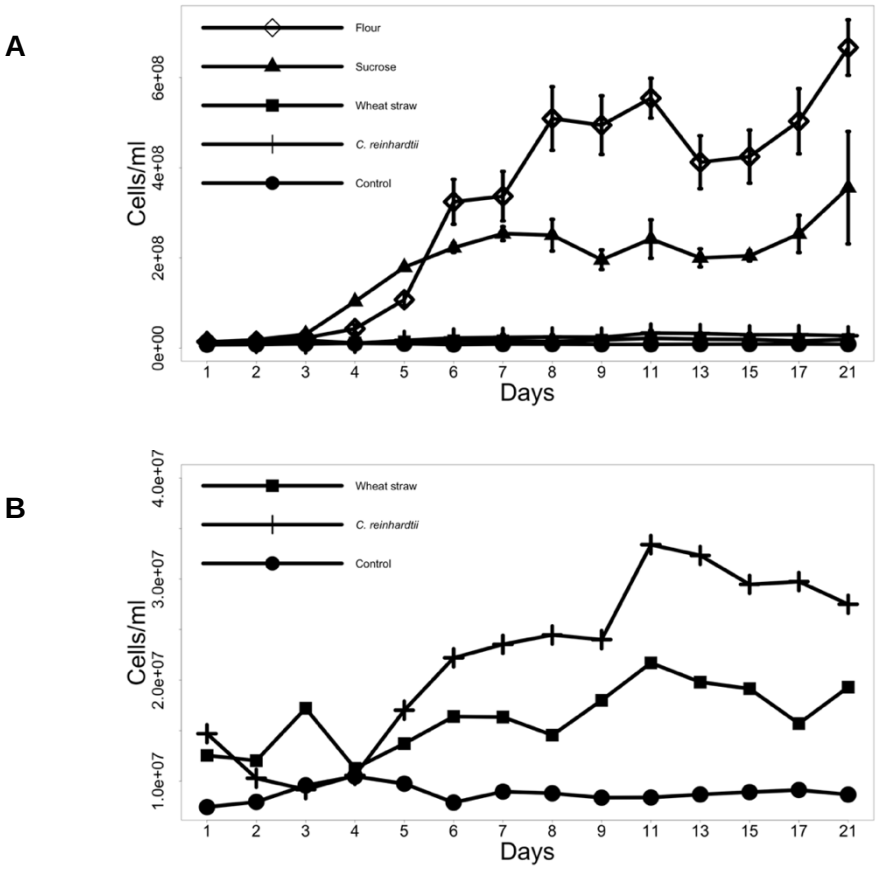


Figure 4.2: (A) Growth curves of *G. sulphuraria* 107 grown on four different substrates and a control over 21 Days. (B) Excluding Sucrose and Flour. Cultures were grown in heterotrophic conditions at 37°C with constant orbital shaking. Error bars are shown as the standard error of three replicates.

The relative growth rate is calculated as a proportion of change in cell density from Day 1 to Day 21 to provide an approximation of cell growth, Table 4.2 summarises these results. The increase in cells per mL in wheat straw is the lowest for both strains, for

strain 107 this is 1.686 % and for strain 427 2.488 %, it should be noted that this equates to more than 10 million cells in both cases.

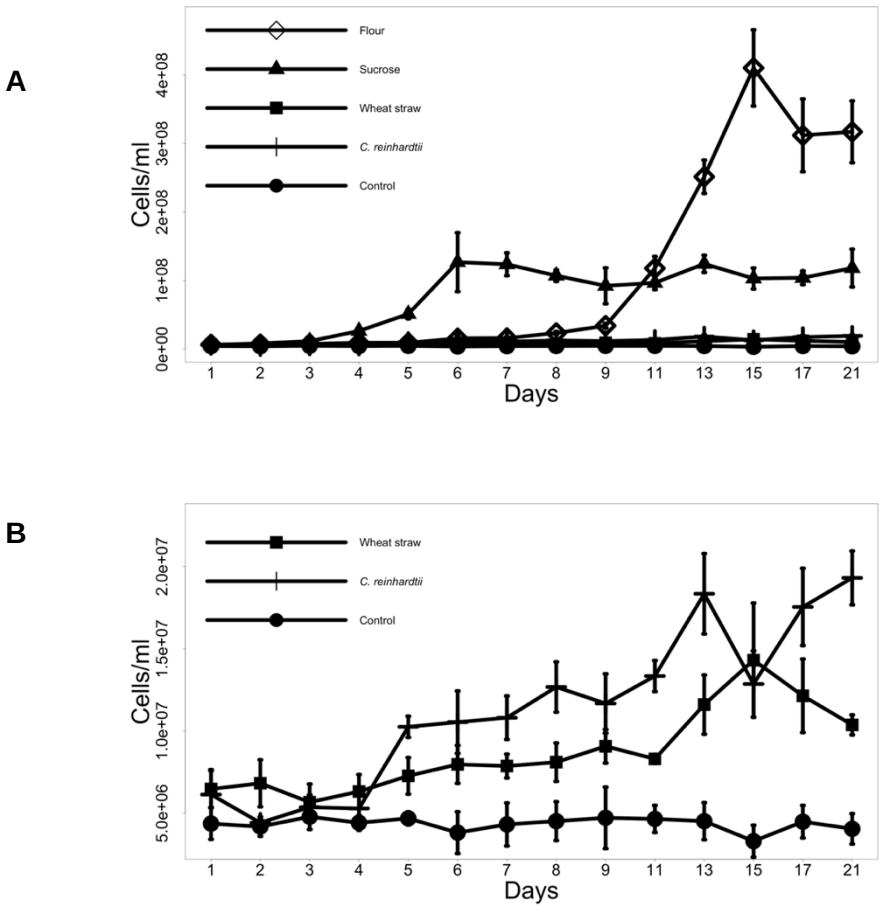


Figure 4.3: (A) Growth curves of *G. sulphuraria* 427 grown on four different substrates and a control over 21 Days. (B) Excluding Sucrose and Flour. Cultures were grown in heterotrophic conditions at 37°C with constant orbital shaking. Error bars are shown as the standard error of three replicates

Table 4.2: The percentage increase in Cells/mL of *G. sulphuraria* 107 and 427 cultures grown heterotrophically in the presence of different substrates after 21 days.

% Increase in Cells/mL	Substrate			
	Flour	Sucrose	Wheat straw	*C. reinhardtii*
107	58.218	31.077	1.686	2.402
427	76.064	28.368	2.488	4.632

4.3.1.1 SEM imaging

Cell morphology and substrate degradation were assessed at the same time as the biomass growth during Experiment 1. Figure 4.4 shows SEM images of Wheat straw grown samples from Experiment 1 both with and without *G. sulphuraria* 427 cells present. Notably, with the sample containing cells the substrate showed disruption to the substrate surface and clear visible holes in the surface when compared to the culture containing no cells (Figure 4.4A and Figure 4.4C). This along with the growth analysis (Figure 4.3, Table 4.2) supports the hypothesis that *G. sulphuraria* is secreting hydrolytic enzymes to degrade lignocellulosic material.

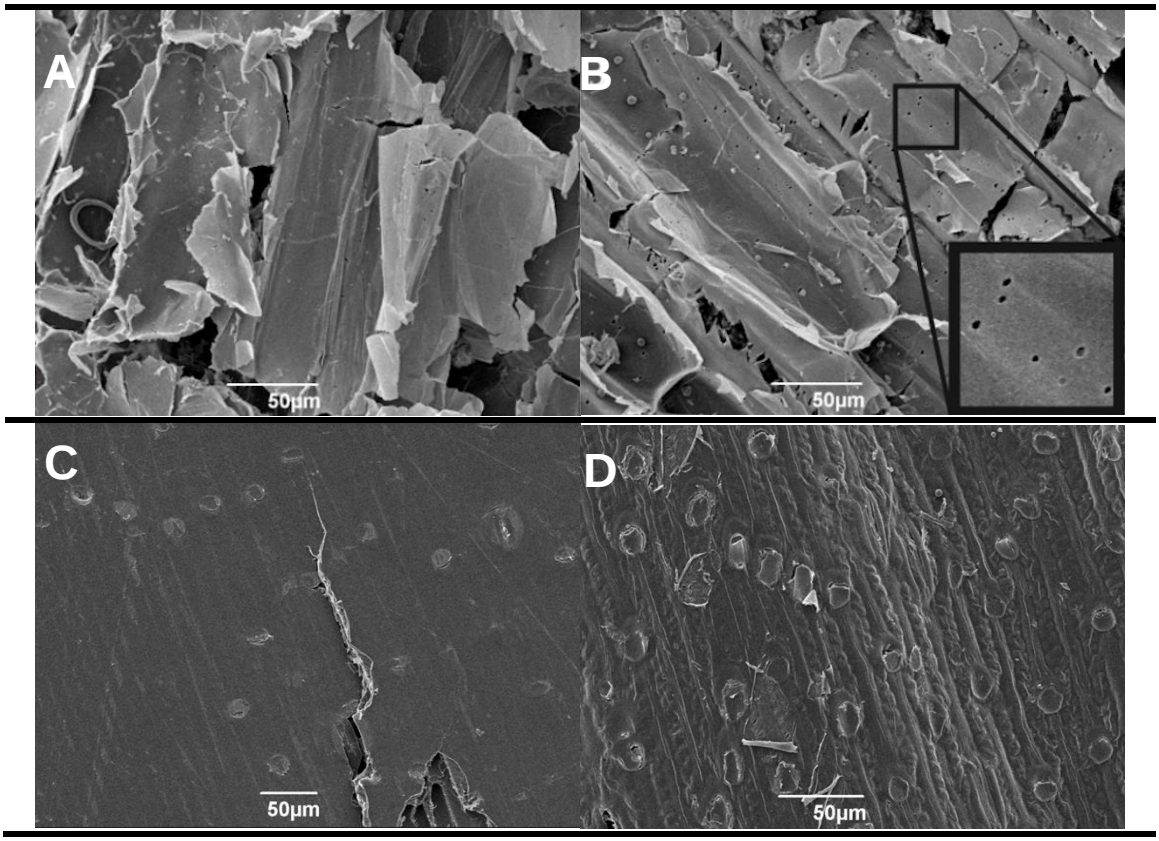


Figure 4.4: SEM images of wheat straw grown in Allen medium at 37 °C for 21 days at 2 % (w/v) without (A and C) and with (B and D) *G. sulphuraria* strain 427. Cultures were grown in heterotrophic conditions with constant orbital shaking. The black square highlights the pores formed in culture containing *G. sulphuraria* cells (B).

4.3.2 Experiment 2

To further explore this hypothesis Experiment 2 was carried out, whereby *G. sulphuraria* 074's growth was assessed on components that make up wheat straw (Figure 4.5). The hemi-cellulose xylan was used alongside lignin, cellulose and sucrose as well as a non-carbon control. During this experiment, *G. sulphuraria* 074 successfully grew on sucrose

and xylan with an increase in OD_{800} of 12.24 % and 10.48 % retrospectively. Lignin showed a smaller increase in growth with relative OD_{800} reaching 4.08 % and lastly the cellulose saw no significant increase in growth when compared to the control.

The high growth rate seen by 074 on xylan (Figure 4) is promising for discovering lignocellulosic degrading enzymes. Due to the varied growth across *G. sulphuraria* strains shown previously it is expected that strains will produce slightly differing repertoires of enzymes.

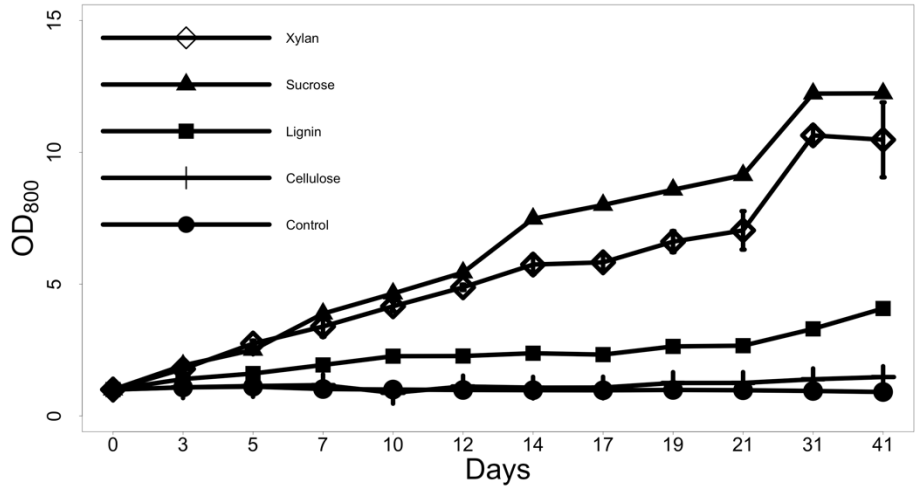


Figure 4.5: Growth curves of *G. sulphuraria* 074 grown on four different substrates and a control over 41 Days. OD was measured at 800 nm for liquid cultures of 074 in Allen Medium supplemented with 2 % (w/v) xylan, 2 % (w/v) sucrose, no added carbon source, 2 % lignin or 2% cellulose. Cultures were grown in heterotrophic conditions at 37°C with constant orbital shaking. Error bars are shown as the standard error of three replicates.

4.3.3 Experiment 3

As a species *G. sulphuraria* has already been shown to be extremely diverse and varied. As a first step toward identifying any potential genes in xylan degradation through growth experiments, we performed both growth assays and TCA precipitations of supernatants of the core 6 strains (017, 033, 074, 107, 138, 427) to allow for comparison of the secreted proteins present. Additionally, to confirm that the growth previously seen on xylan (Figure 4.5) was not due to the acidic media already having degraded the substrate during Experiment 3, strains were also grown with the corresponding monosaccharide xylose.

During Experiment 3, *G. sulphuraria* strains 017, 033, 074, 107, 138 and 427 all successfully grew on the three substrates tested (Figure 4.7). All strains showed variety in growth relative to each other and between the substrates. At the end point of the experiment (31 days) for all strains: ((A) 017, (B) 033, (C) 074, D) 107, (E) 138 and (F) 427) xylan was the substrate with the lowest growth. Growth on xylan showed between 7.141 % and 15.04 % increase in OD_{800} across all strains (Table 4.3). Half of the strains showed sucrose to have the highest growth (017,074,138) and half xylose (033, 107, 427). Table 4 shows the relative increase in OD_{800} across all strains at the endpoint of the experiment. This again highlights the differences in growth capacities between the strains. Strains 138 and 074 show the highest increase in OD_{800} when grown on xylan (15. 04 % and 13.737 %).

Table 4.3: The percentage increase OD measured at 800 nm for liquid cultures of Galdieria strains (017, 033, 074, 107, 138, 427) grown heterotrophically in the presence of three different substrates and control after 31 Days. NC: no carbon control.

	Xylan	**Xylose**	**Sucrose**	**NC**
017	12.016	17.811	17.854	1.321
033	7.141	24.660	11.012	1.007
074	13.737	18.503	19.313	1.615
107	9.814	22.950	21.421	1.199
138	15.040	17.478	21.216	1.103
427	13.600	29.695	27.883	1.320

In addition to biomass growth samples of the supernatant on Day 20 were collected and a TCA precipitation carried out. The supernatant proteins present are important in aiding the discovery and identification of any substrate degrading enzymes present. An SDS-PAGE visualising this is shown in Figure 4.6.

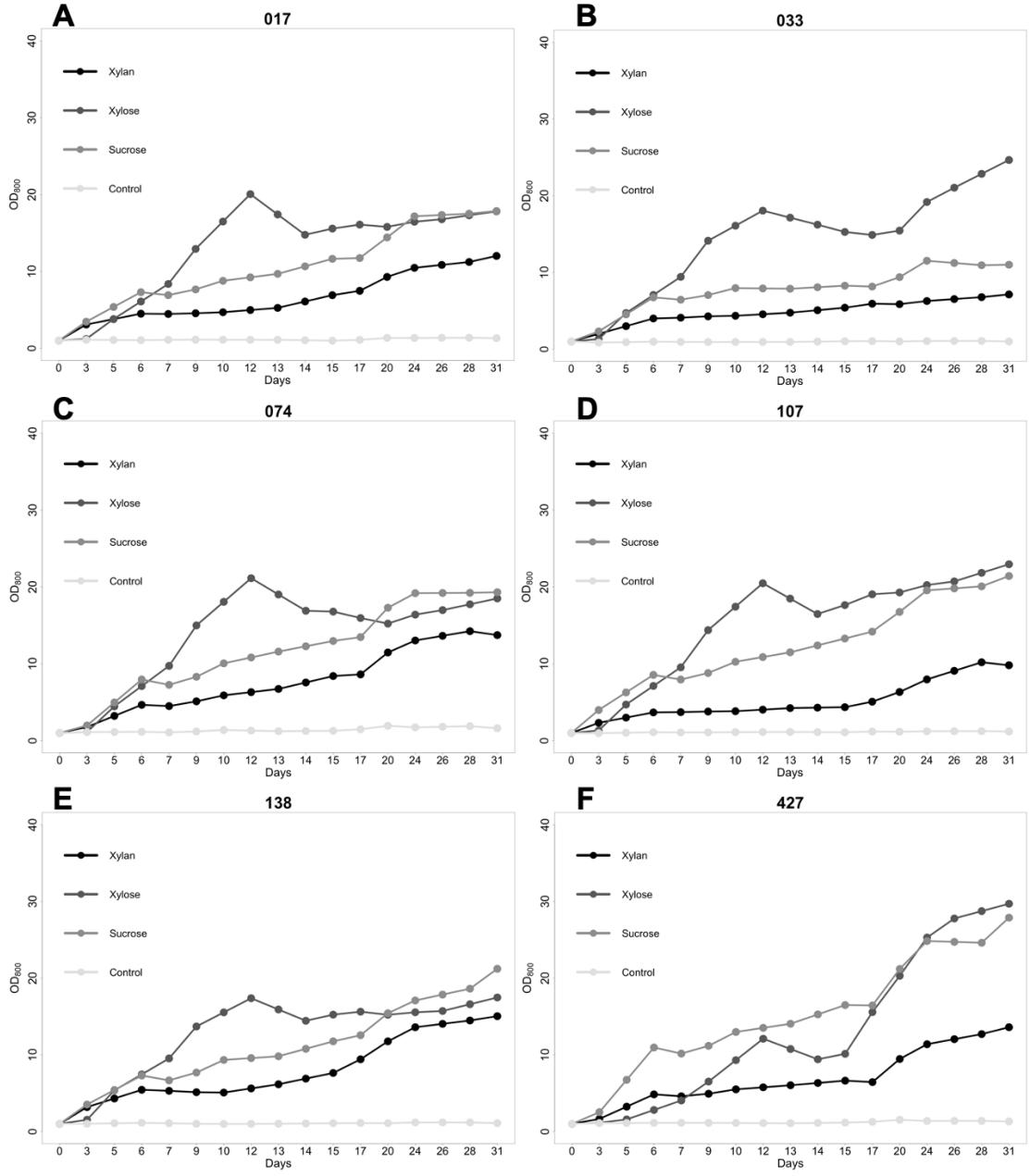


Figure 4.6: Growth curves of *G. sulphuraria* strains grown on three different substrates and a control over 31 Days. (A) Strain 017, (B) strain 033, (C) strain 074, (D) strain 107, (E) strain 138 and (F) strain 427. OD was measured at 800 nm for liquid cultures in Allen Medium supplemented with 2 % (w/v) xylan, 2 % (w/v) xylose, 2 % (w/v) sucrose and no added carbon source as the control. Cultures were grown in heterotrophic conditions at 37°C with constant orbital shaking. Each curve had been normalised relative to its inoculation on Day 0.

Firstly, for the xylan grown cultures (Figure 4.7A) there is a clear difference in banding pattern between the six strains. Strain 074 shows the largest set of bands, interestingly 138 which had the highest growth increase (Table 4.3) has one of the faintest bands. Strain 427 shows a very clear band just below the ~ 70 kDa mark, this band could also be in the 074 sample but due to smearing is undetectable. The SDS-PAGE showing the xylose samples (Figure 4.7B) although still different between all six strains shows more banding in common with each other. Notably a set of bands between ~55 – 75 kDa and just below ~55 kDa. Lastly the visualisation of the Sucrose grown samples (Figure 4.7C) show a much more conserved banding pattern between the six strains, the differences are less obvious. Most notably is that there are different banding patterns between substrates.

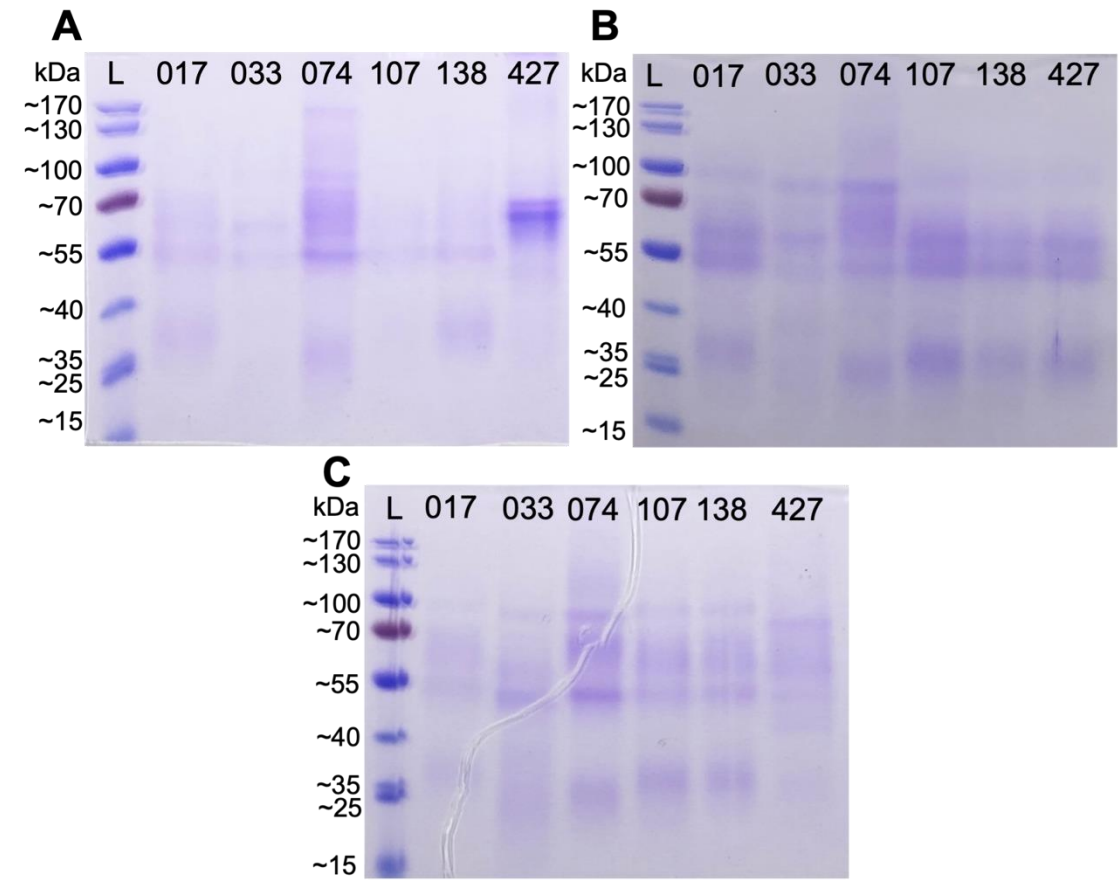


Figure 4.7: Coomassie stained SDS-PAGE gels of TCA precipitated supernatant from *G. Sulphuraria* strains 017, 033, 074, 107, 138 and 427 grown heterotrophically at 37 °C on (A) 2 % xylan, (B) 2 % xylose and (C) 2 % sucrose.

4.3.4 Experiment 4

To explore the *G. sulphuraria* proteome, Experiment 4 was carried out. Strain 074W was grown on 2 % (w/v) xylan, 0.5 % (w/v) sucrose as a positive control and no carbon as a negative control. The growth curves of this are shown in Figure 4.8 for each condition there were three replicates. Again, xylan supported *Galdieria* growth well, in this case the growth on sucrose plateaus around Day 7. A Trichloroacetic Acid (TCA) precipitation was carried out samples collected on day 10 to concentrate protein and change the pH of the buffer to be detectable on an SDS-PAGE. Figure 4.9 shows the visualisation of the three replicates of supernatant from xylan grown cultures on a 10% SDS PAGE stained with Coomassie Brilliant Blue r-250 where proteins can be clearly identified. It should be noted a no carbon and 0.5 % sucrose sample showed no visible bands so are not displayed.

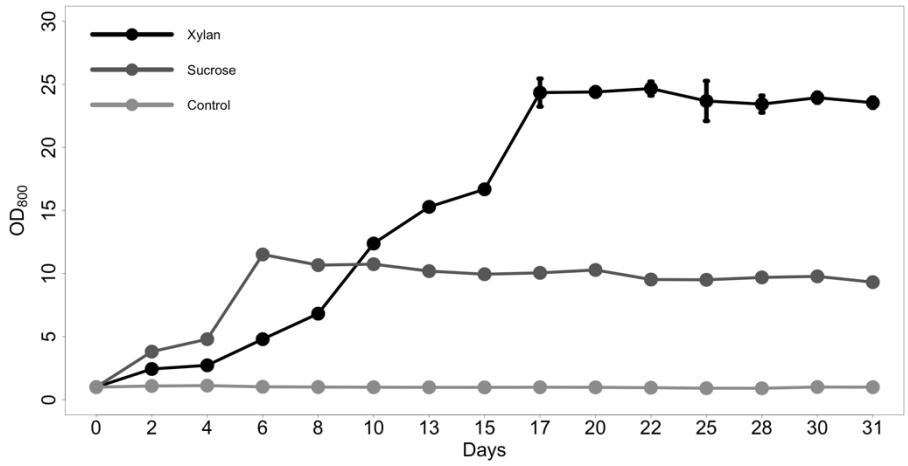


Figure 4.8: Growth curves of *G. sulphuraria* 074W grown on two different substrates and a control over 31 Days. OD was measured at 800 nm for liquid cultures of 074W in Allen Medium supplemented with 2 % (w/v) xylan, 0.5 % (w/v) sucrose, or no added carbon source. Cultures were grown in heterotrophic conditions at 37°C with constant orbital shaking. Error bars are shown as the standard error of three replicates.

Proteomic analysis provides an important means for identifying proteins present in mixtures such as a xylanase cocktail. By growing *G. sulphuraria* on xylan, concentrating the supernatant and visualising the protein content on an SDS-PAGE, it was possible to carry out an in-depth analysis of the secreted proteins produced by the algae. In order to begin to build a full secretome profile of *G. sulphuraria*, the protein bands shown in Figure 4.9 were sent for Liquid Chromatography Mass spectrometry (LC-MS) analysis. This was completed by the University of Sheffield as described in Section 4.2.5.

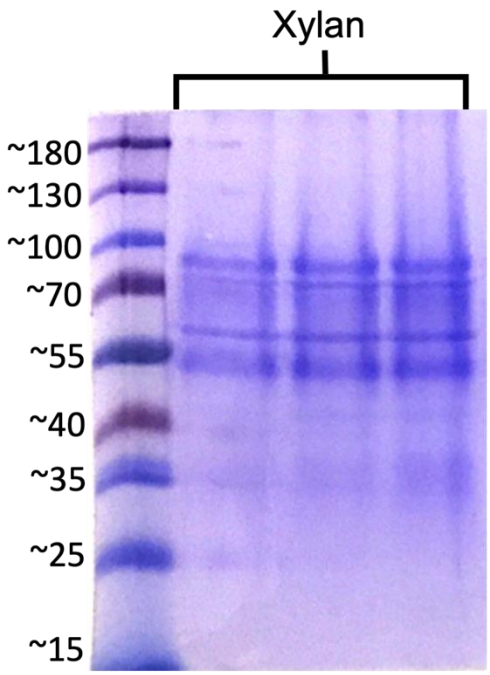


Figure 4.9: Coomassie stained SDS-PAGE gels of TCA precipitated supernatant from *G. Sulphuraria* strain 074 grown heterotrophically on 2 % xylan at 37 °C. The no carbon and 0.5 % sucrose control samples showed no visible bands so are not displayed.

4.3.5 Secretome analysis

The LC-MS of xylan grown supernatant produced peptide sequences via trypsin digest, these were then mapped back to the published genome (The UniProt Consortium, 2018). This connects information at the protein level to information at the genome level (Alves et al., 2007). The output contained 26 proteins, of these identified proteins, 23.1% are classified as uncharacterised proteins, 15.4% Peroxidases, 3.8% Ubiquitin, 7.7% Beta-Ig-H3/fasciclin, 7.7% Beta-galactosidase, 3.8% Alpha-galactosidase, 7.7% Alpha-glucosidase, 3.8% Purple acid phosphatase, 3.8% Aspartyl protease, 3.8% Molecular chaperone DnaK 2-dehydrogenase, 3.8% Serine/threonine protein kinase, 3.8% Actin, 3.8% Elongation factor 1-alpha, 3.8% VanW family protein, 3.8% Enolase, 3.8% Aldo/keto reductase, 3.8% Elongation factor 1-alpha (Figure 4.10).

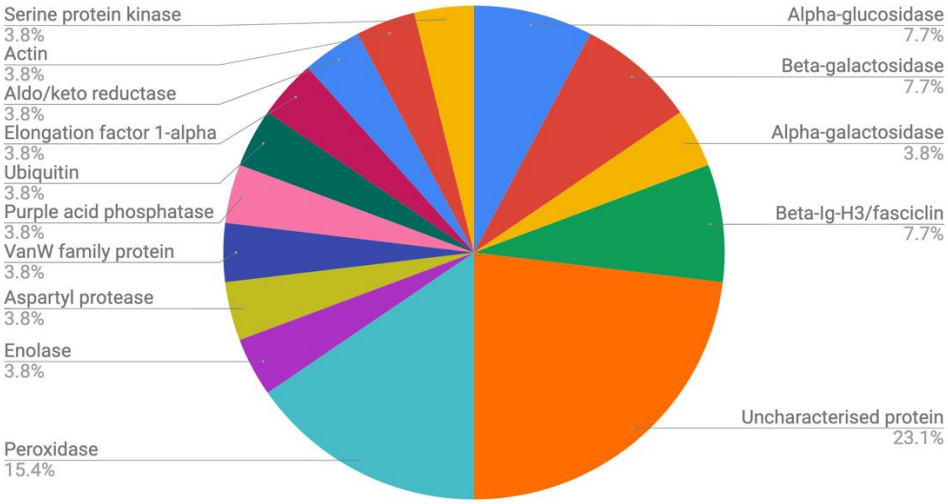


Figure 4.10: Proportion of protein groups found from *G. sulphuraria* 074W secretome (26 genes) when grown in Allen's medium with 2% (w/v) xylan.

A full list of these proteins and their corresponding peptide sequences are listed in Supplementary Table 9. From these 26 encoded proteins, 11 were identified as being potentially involved in *G. sulphuraria* ability to grow on xylan. This was determined by placing restrictions on LC-MS/MS secretome data, focusing on proteins with a unique peptide hit >1 then filtered for only peroxidases, hydrolases and uncharacterised enzymes that had a predicted signal peptide (Figure 4.11). It is worth remarking the significant proportion of uncharacterised proteins found.

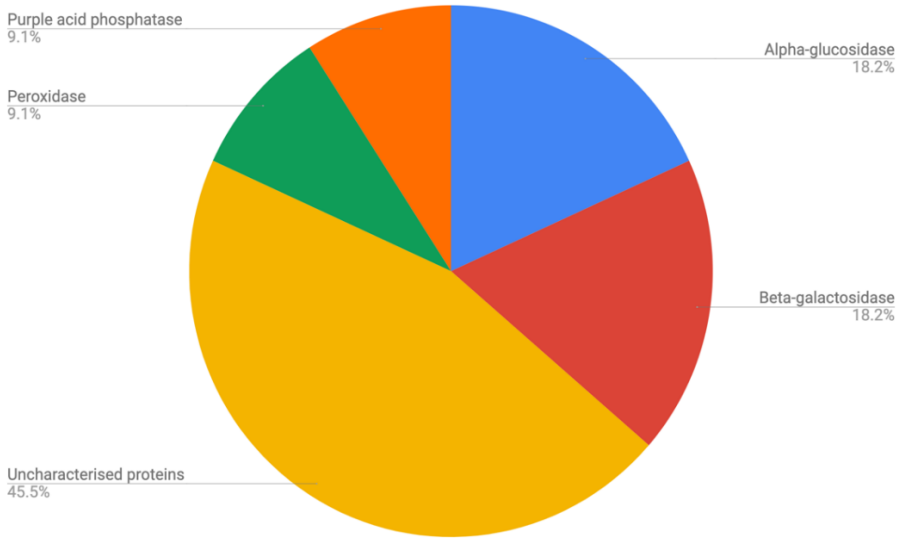


Figure 4.11: Distribution of proteins identified from *G. sulphuraria* 074W reduced secretome (11 genes) when grown in Allen's medium with 2% (w/v) xylan. The reduced secretome was achieved by including only proteins with a unique peptide hit >1 and filtered for only peroxidases, uncharacterised proteins and hydrolytic enzymes with proteins showing predicted signal peptides using SignalP (Petersen et al., 2011).

To gather more information on the type and function of protein groups seen in Figure 4.11 and to try and identify their role in potential xylan breakdown, further analysis was required. The protein identifiers retrieved from the LC-MS data were cross referenced with The UniProt Consortium (2021) to obtain gene name, protein ID, EC number, protein name and gene ontology (GO). Identified proteins were functionally classified according to their biological role**.** Table 4.4 summarises the proteins including the number of unique peptide hits, GC%, exon count and isoelectric point (pI). Next, in order to attain which enzymes are likely to make the best candidates for recombinant protein expression conserved domains were assessed. Table 4.5 displays the conserved domains for 11 candidate genes and the presence/absence of these genes as identified by LC-MS. This may assist in identifying the most likely candidates for secreted enzymes involved in the degradation of xylan. It is seen in Table 4.5, six of the 11 genes were present only in the xylan secretome analysis (Gasu_17800, Gasu_24700, Gasu_27490, Gasu_27500, Gasu_29970 and Gasu_31410). Additionally, Supplementary Table 10 displays the top 10 BLASTp hits for each gene.

Table 4.4: Detailed information on 11 genes from *G. sulphuraria* 074W, identified by mass spectrometry when grown in Allen's medium with 2% (w/v) xylan. MW; molecular weight, pI; isoelectric point.

Gene Name	Protein ID	Unique peptide hits	EC number	Protein name	Length (aa)	GC %	Exon	MW (kDa)	pI	Gene ontology (GO)
Gasu_17800	A5JW32	6	1.11.1.7	Peroxidase	323	43	3	31.73	4.53	heme binding [GO:0020037]; peroxidase activity [GO:0004601]; response to oxidative stress [GO:0006979]
Gasu_21790	M2W430	2	NA	Uncharacterised	401	40	5	41.10	4.71	NA
Gasu_21980	M2W452	5	NA	Uncharacterised	393	40	1	40.81	4.85	NA
Gasu_24700	M2W312	2	NA	Uncharacterised	244	38	1	24.23	4.66	NA
Gasu_25520	M2XJ88	29	3.2.1.3	Glucan 1,4-alpha-glucosidase	508	40	6	55.40	5.48	glucan 1,4-alpha-glucosidase activity [GO:0004339]; polysaccharide metabolic process [GO:0005976]
Gasu_25530	M2Y2J8	26	3.2.1.3	Glucan 1,4-alpha-glucosidase	529	40	5	56.93	6.33	glucan 1,4-alpha-glucosidase activity [GO:0004339]; polysaccharide metabolic process [GO:0005976]
Gasu_27490	M2W2P6	14	3.2.1.23	Beta-galactosidase	952	41	2	105.96	5.2	beta-galactosidase activity [GO:0004565]; carbohydrate metabolic process [GO:0005975]
Gasu_27500	M2XIP7	17	3.2.1.23	Beta-galactosidase	1038	41	1	115.70	5.42	beta-galactosidase activity [GO:0004565]; carbohydrate metabolic process [GO:0005975]
Gasu_29970	M2XHJ6	3	3.1.3.2	Purple acid phosphatase	538	41	1	58.58	5.29	acid phosphatase activity [GO:0003993]; metal ion binding [GO:0046872]
Gasu_31410	M2XHE4	6	NA	Uncharacterised	378	43	1	38.31	5.9	NA
Gasu_41530	M2WWC0	4	NA	Uncharacterised	228	39	1	23.27	5.77	NA

Table 4.5: Information on conserved domains and the presence/absence in cell lysate LC-MS for 11 genes from *G. sulphuraria* 074W, identified by mass spectrometry when grown in Allen's medium with 2% (w/v) xylan.

Gene Name	Conserved Domains	Present in cell lysate
Gasu_17800	Heme-dependent peroxidase, L-ascorbate peroxidase, Catalase (peroxidase I)	N
Gasu_21790	NA	Y
Gasu_21980	SEST_like domain, SGNH	Y
Gasu_24700	NA	N
Gasu_25520	GH15 domain, oligosaccharide amylase domain, Glucoamylase (glucan-1,4-alpha-glucosidase) - GH15 family	Y
Gasu_25530	GH15 domain, oligosaccharide amylase domain, Glucoamylase (glucan-1,4-alpha-glucosidase) - GH15 family	Y
Gasu_27490	GH35 domain, Beta-galactosidase - domain 2, 3, 4 & 5, Beta-galactosidase jelly roll domain	N
Gasu_27500	GH35 domain, Beta-galactosidase - domain 2, 3, 4 & 5, Beta-galactosidase jelly roll domain	N
Gasu_29970	metallophosphatase superfamily, Calcineurin-like phosphoesterase, Purple acid Phosphatase, N-terminal domain, 3',5'-cyclic AMP phosphodiesterase CpdA	N
Gasu_31410	NA	N
Gasu_41530	NA	Y

4.3.5.1 Peroxidases

Gene Gasu_17800 displayed conserved domains in three peroxidase families and thus is a predicted peroxidase. Gasu_17800 measured a high GC % content and the lowest pI value (Table 4.4 and Table 4.5). BLASTp was used to identify the top homology hits based on the E-value for the peroxidase Gasu_17800. Supplementary Table 10 fully details the top 10 matches of which there were four further *G. sulphuraria* peroxidase identified. The four identified *G. sulphuraria* genes along with their percentage identity to gene Gasu_17800 are as follows; Gasu_50490 (85.5% identity, E-value 7e-46), Gasu_17790 (70.29% identity, E-value 5e-124), Gasu_50480 (65.04% identity, E-value

4e-116) and Gasu_64070 (85.89% identity, E-value 9e-99). The sequences after this show matches from E-vales at 2e-30 are all eudicots and monocots.

4.3.5.2 Alpha-glucosidase

Two alpha-glucosidases were identified by LC-MS in both the cell lysate and the xylan grown secretome, specifically conserved domains in glycoside hydrolase family 15 (GH15) (Table 4.5). When search via BLASTp the two alpha-glucosidase Gasu_25520 and Gasu_25530 identify each other as their top hit (49.8% identity, E-value 9e-177 and 50.4% identity, E-value 1e-176 respectively), followed by matches from E-values at 7e-83 to various bacteria, fungi and yeast. For both genes the sequence homology hits are mostly for glucan 1,4-alpha-glucosidases (EC 3.2.1.3) or the broader (GH15) and some hypothetical proteins. Both alpha-glucosidases revealed no matches to any well characterised proteins in the PBD database.

4.3.5.3 Beta-galactosidase

Two of the identified genes (Gasu_27490 and Gasu_27500) had predicted beta-galactosidase function, both had conserved beta-galactosidase 2, 3, 4 & 5 domains and were only present in the xylan grown secretome (Table 4.4 and Table 4.5). Of the two Beta-galactosidases Gasu_27490 top hit is with another *Galdieria* beta-galactosidase Gasu_09330 (47.14% identity, E-value >1e-250). All other matches are with mostly bacteria and few fungi with all E-values >1e-250. Gasu_27500 top match is also with the *Galdieria* beta-galactosidase Gasu_09330 (48.22% identity, E-value >1e-250). It's second match is with Gasu_27490 (46.58% identity, E-value >1e-250), and all other matches are with bacteria with all E-values >1e-250. Both genes revealed no matches to any well characterised proteins in the PDB database.

4.3.5.4 Purple acid phosphatase

Gasu_29970 was the only purple acid phosphate present and was only present in xylan grown samples (Table 4.4). The top four BLASTp hits to this protein were all red algae. The first is a hypothetical protein in sister species *Cyanidioscshyzon merolae* (63.23% identity, E-value >1e-250), next is G. *sulphuraria* gene Gasu_43490 another metallo-dependent acid phosphatase (45.81% identity, E-value 2e-143). Next are two hypothetical proteins from *Cyanidiococcus yangmingshanensis* (73.11% and 72.57% identity, E-value 1e-56 and 6e-55). There were no matches to any proteins in the PDB database.

4.3.5.5 Uncharacterised

The uncharacterised proteins are the largest group of proteins identified. Sequence homology is a good way to identify potential functions within the uncharacterised proteins and search for any conserved domains. Firstly Gasu_21790 gave no results, then both Gasu_24700 and Gasu_41530 only hits were each other (27.75% identity, E-value 3e-10 and 27.27% identity, E-value 3e-11 respectively). Next Gasu_21980 had eight hits in total, each were types of bacteria; the top hit was a hypothetical protein from *Pseudomonas syringae* with (85.07% identity, E-value 2e-75). The next is a SGNH/GDSL hydrolase family protein from *Actinobacteria bacterium* (26.33% identity, E-value 4e-19), this gene contains a conserved region from SEST_like domain - SGNH. Lastly Gasu_31410, which had four hits in total, the first was to a structural protein in *Myoviridae sp*. a virus from a family of bacteriophage (26.6 % identity, E-value 6e-13). The next hit was from *Acetobacter senegalensis,* a species of thermophilic acetic acid bacteria with a hypothetical protein (23.88% identity, E-value 6e-07). The last two hits were for hypothetical proteins from a species of archaea, *Thermoplasmata archaeon (*27.41% identity, E-value 7e-06 and 27.41% identity, E-value 8e-06). Predicted protein structures aquired from i-TASSER and AlphaFold when used to search for structural similarities revealed no matched and so not included.

4.4 Discussion

4.4.1 Experiment 1

This study shows that both *G. sulphuraria* strains 107 and 427 grown on media containing flour, sucrose, wheat straw and *C. reinhardtii* had higher levels of cells per mL than cultures grown without any carbon source (Figure 4.2 and Figure 4.3). For each of the strains the highest % increase in cell density occurred in a culture medium containing flour with a 58.2 % increase in 107 and 76.1 % increase in 427. This was followed by sucrose with a 31.1 % increase in 107 and 28.4 % increase in 427, although initially sucrose growth was faster with flour not overtaking until day 6 (strain 107) and day 11 (strain 427). The reason for initial fast growth on sucrose could be due to the availability of simple sugars. The acidic nature of the media and the autoclaving process will have caused hydrolysis of the glycosidic bond in the disaccharide, therefore leaving the glucose and fructose molecules free for uptake (Eggleston and Vercellotti, 2000; Les, 2001; Bower, 2008). In comparison, flour contains a high proportion of more complex polysaccharides such as starch, consequently there was a delay before the growth rate increased dramatically. For *G. sulphuraria* to use flour as a carbon source, some

glycosidic bonds must be broken. Though it is likely that acidic media when autoclaved would hydrolyse some of the proteins and starch in flour due to the more complex nature compared to sucrose, it is improbable that this would have resulted in the same availability of simple monosaccharides as sucrose. For this breakdown of polysaccharides, it is probable that *G. sulphuraria* is using secreted glycoside hydrolases to break the glycosidic bonds.

The two remaining carbon sources investigated still supported *G. sulphuraria* growth, however, to a much lower rate. *C. Reinhardtii* showed a higher percentage increase in cells per mL in strain 427 compared to 107, 4.6 % and 2.4 % retrospectively. Of these carbon sources *Chlamydomonas*, a green alga and wheat straw represent feedstocks that could be encountered by *G. sulphuraria* in its natural habitat. In ecological sites *G. sulphuraria* is most probably hydrolysing green algal cell debris, as well as lignocellulosic plant material. *C. reinhardtii*, cell wall composition includes hydroxyproline-rich glycoproteins (Adair and Snell, 1990), alongside this wheat straw, is mostly made up of cellulose, hemicellulose and lignin (Wang et al., 2013; Tian et al., 2018; Raud et al., 2019). These more complex structures are a possible explanation for the delayed cell density increase in both substrates. It can be hypothesised that *G. sulphuraria* will have low levels of hydrolases secreted, then once these act on a substrate *G. sulphuraria* will sense the monosaccharides this can lead to hydrolase induction of such complex materials, and that will prompt further production of the relevant enzymes for degradation of substrates.

4.4.1.1 SEM Imaging

To investigate the degradation of substrates samples were taken from Experiment 1 during growth for SEM imaging. The SEM image of the control wheat straw sample (Figure 4.4A) shows the pocket like structure of the straw with a generally smooth surface of the control sample. The SEM image of the cell grown wheat straw (Figure 4.4B) displays the same pocket-like structure of the wheat straw along with clear signs of structural degradation. There are examples of irregular disruption and cracking along the surface but most prominently pores penetrating the substrate surface. These pores are evidence of disruption and dissolution to the lignocellulosic matrix in wheat straw which indicates the degradation of lignin, cellulose and hemicelluloses has occurred. The SEM image in Figure 4.4C showed the outer surface of the control sample of wheat straw and appears almost homogeneous and smooth with few discrepancies along the surface. However, similar to the differences between Figure 4.4A and Figure 4.4B the SEM image in Figure 4.4D showing outer surface of the cell grown wheat straw displays some

irregular disruption, shallow grooves as well as a general looseness, and partial removal of the outer surface layer. These changes in the microstructure of the wheat straw outer and interior surface and then results obtained from the assessment of growth support the hypothesis that *G. sulphuraria* is firstly capable of growing on lignocellulosic material and secondly that there is some form of enzymatic process in assisting the degrading of this material.

4.4.2 Experiment 2

Furthering the discovery for lignocellulosic degrading enzymes from *G. sulphuraria* Experiment 2 was set up to identify the growth capacity of *G. sulphuraria* specifically on the individual components, lignin, cellulose and hemicellulose. Xylan is a type of hemicellulose and represents after cellulose the most abundant renewable polysaccharide, it also contributes > 30 % of the dry weight of land plant cell walls (Ahmed et al., 2009).

This experiment showed that *G. sulphuraria* did not grow on cellulose, meaning it is unlikely that there are any cellulose acting enzymes encoded in the genome. However, the growth on lignin, which increased the relative OD by four times (Figure 4.5) suggesting some potential enzymatic activity to account for the growth. *G. sulphuraria* has no predicted phenol oxidases (laccases) which are typically one of the two families of ligninolytic enzymes. It does however contain several predicted peroxidases, namely a family of Class III secreted plant peroxidases (Sano et al., 2001; Oesterhelt et al., 2008). This class of plant peroxidase is involved in several cellular process, that include auxin metabolism, wound healing, defence against pathogens, cell elongation and lignification.

Xylan basic chain structure is a polymer backbone of β-1,4 linked xylose units, with side branching units of α-arabinofuranose and/or α-glucuronic acids (Moreira, 2016, Lopes et al., 2018). This complex structure is crosslinked by both covalent and ionic bonds, providing a physical barrier to cellulose and limiting penetration from mechanical or microbial attacks. The breakdown of xylan requires a conglomerate of enzymes both endo- an exo- acting, that will hydrolyse internal and terminal glycosidic linkages (Lopes et al., 2018). The growth curve in Figure 4.5 clearly shows xylan to be the substrate with the highest supported growth of the three components tested. Notably its growth curve shows a very similar pattern to that seen from the sucrose sample, as previously discussed the autoclaving process along with the acidity of the media will have saccharified the sucrose into monomers glucose and fructose. Therefore, the effect of

autoclaving and acid hydrolysis on the structure and decomposition of xylan should be explored.

The degree of polymerisation of xylan is much lower than cellulose, this along with its amorphous structure and length of the polysaccharide chain mean its thermal and chemical stability is lower (Hilpman et al., 2016). Literature shows there is a strong impact of the acid concentration along with temperature. For full conversion of xylan to xylose, the pH was required to be less than 2 and temperatures upward of 90 °C for 16 hours (Hilpman et al., 2016; Jiang et al., 2016; Mittal et al., 2019). The cultures were grown in Allen medium adjusted to pH 2 using H_2SO_4 and then autoclaved (121° C, 15 psi for 30 minutes). It is unlikely that the xylan in the media is being completely broken down into its monosaccharide xylose. However, the growth of *G. sulphuraria* over both xylan and xylose should be observed whilst also checking for any difference in secreted proteins whilst in the presence of these substrates. This was explored in Experiment 3.

4.4.3 Experiment 3

Experiment 3 was designed to determine the growth capacities of the core six *G. sulphuraria* strains on xylan, xylose and sucrose, and to visualise and compare the proteins present in the supernatant as a basis to move forward with mass spectrometry. During the experiment all six strains (017, 033, 074, 107, 138 and 427) successfully grew on all three substrates tested (Figure 4.6). Strain 017, 074 and 138 showed sucrose to be the substrate at best supporting growth while for strains 033, 107 and 427 it was xylose (Figure 4.6, Table 4.3). *Galdieria* has been shown to metabolise xylose, glucose and fructose and the differences between the strains are likely a result of the evolution and diversity between them, as showed in Chapter 2 (Oesterhelt and Gross, 2002; Schönknecht *et al*., 2013; Qiu, Price *et al.,* 2013). The *G. sulphuraria* genomes also contain three predicted xylose degrading enzymes (Gasu_29170, Gasu_44520 and Gasu_32580) that are likely involved in the growth seen on xylose (Schönknecht *et al*., 2013). The lowest increase in growth for all strains was xylan. This was expected and supports the hypothesis that the breakdown of xylan into its monosaccharide (xylose) is due to *G. sulphuraria* secreting some enzymes to hydrolyse the substrate, as opposed to the media preparation. The slower response in growth is explained by the lag phase shown by cells when transferred from autotrophic to heterotrophic conditions (Gross and Schnarrenberger, 1995). The lag phase observed can be explained by the requirement of the sugar and polyol uptake system needing to be induced and ‘switched on’ (Oesterhelt et al., 1999; Oesterhelt and Gross, 2002; Barbier et al., 2005). This uptake system is responsible for the cells uptake of substrates and consists of over 14 sugar

and polyol transporters. This system is induced under heterotrophic conditions not by darkness alone but equally the presence of a metabolizable substrate. Depending on the type of substrate the induction pattern of transporters changes (Oesterhelt et al., 1999; Oesterhelt and Gross, 2002; Barbier et al., 2005).

To visualise these differences in growth and the suspected proteins responsible, samples were taken from the experiment at day 20 and the supernatants analysed for proteins. Figure 4.7 shows the proteins present in the supernatant for the six strains when grown on xylan, xylose and sucrose. Firstly, once again this highlights the diversity between these 6 strains and lineages as the protein banding pattern present across all the substrates between the strains is notably different, reflecting the results of the growth assay. There are clear differences in the banding patterns between the different substrates. This is evidence that the enzymes secreted are substrate dependent supporting the hypothesis that there must be some sort of feedback sugar sensing mechanism. *G. sulphuraria* had already been shown to harbour secreted enzymes that are acid tolerant, so it is likely that the secretomes visualised here harbour acid tolerant proteins (Oesterhelt et al., 2008).

Genome analysis of *G. sulphuraria* has revealed sequences that encode a variety of genes, have archaeal and bacterial origins and were acquired through horizontal gene transfer (HGT) (for example, archaeal ATPases, bacterial pumps and antiporters) (Schonknecht et al., 2014; Lee et al., 2017). HGT events from extremophile bacteria and archaea may be responsible for facilitating the alga's ability to become the dominant species in its extreme environments (Oesterhelt et al., 2008; Schönknecht et al., 2013; Lee et al., 2017; Rossoni et al., 2019). The proteins visualised in Figure 4.7 are likely to function under elevated temperatures and acidic conditions, they are also involved in the degradation of xylan (Figure 4.7A). To further explore this, Experiment 4 was carried out. The analysis was repeated on xylan using strain 074. At the time of analysis, this strain was the only fully annotated genome available.

4.4.4 Experiment 4

Strain 074 was grown in triplicate along with a sucrose positive control and a no carbon negative control. As Figure 4.8 shows all replicates were very similar to each other as expected. TCA precipitation was performed on the three xylan samples and visualised in an SDS-PAGE (Figure 4.9). This gel sample was sent for Liquid Chromatography Mass spectrometry (LC-MS) at The University of Sheffield to assist in building a secretome profile for *Galdieria*, of the total proteins encoded by a genome, secreted

proteins account for around 10% (Mukherjee and Mani, 2013). LC-MS of the cell pellets alongside the secretome allowed differences in peptide sequences from the secreted samples to be identified. In the supernatant, proteins from cell death that are not the target secreted proteins are often present. For example, cell wall components and other polysaccharides.

4.4.5 LC/MS

Proteomic analysis is a vital resource for identifying proteins present in mixtures such as, substrate specific supernatant samples. In particular mass spectrometry-based proteomics is a well-utilised tool used in the identification and quantification of an organisms secretome (Karpievitch et al., 2010). By growing *G. sulphuraria* on xylan, concentrating the supernatant and visualising the protein content on an SDS-PAGE, it was possible to carry out a comprehensive analysis of any secreted proteins produced by the algae. Over a quarter of genes identified were sugar hydrolases which is as expected due to the ability of *G. sulphuraria* to utilise multiple carbon sources. Many of the other enzymes are typical of eukaryotic cells, for example ubiquitin and aspartyl protease, which is optimally active under acidic conditions, all of these types of enzymes are used to maintain cellular function. Interestingly nearly a quarter of the genes identified were uncharacterised, it can be hypothesised that these particular genes hold the potential to understanding how *G. sulphuraria* can grow on complex substrates under such harsh conditions. In order to investigate this further and begin to build a the secretome profile of *G. sulphuraria*, LC-MS results were filtered (Figure 4.11).

4.4.5.1 Peroxidases

As previously discussed, peroxidases are known to play a significant role in the degradation of lignin (Raud et al., 2019). Gene Gasu_17800 is the only predicted peroxidase present in the secretome list (Table 4.4), with three conserved domains from the peroxidase family (Heme-dependent peroxidase, L-ascorbate peroxidase, Catalase (peroxidase I; Table 4.5). The top sequence homology hits were as expected with the other *Galdieria* peroxidases with sequence identity matches from 85 – 65 % (Oesterhelt et al., 2008). The remaining hits were with numerous plant species, this is consistent with it being a class III plant peroxidase (Supplementary Table 10). Literature has shown all land plants contain members of the gene family that encodes class III peroxidases making it extremely conserved (Duroux and Welinder, 2003).

Class III peroxidases are secretory enzymes that catalyse the oxidation of a variety of substrates by hydrogen peroxide (H_2O_2) and also work on a multifunctional level. They

take electrons from phenolic substances, lignin precursors and other secondary metabolites, and are involved in cell wall cross-linking, loosening and lignification (Hiraga et al., 2001; Oesterhelt et al., 2008). The *Galdieria* cell wall can withstand a proton gradient from the intracellular pH of 6.8-7.0 and an external pH 2, meaning the cell wall has to be especially ridged (Merola et al., 1981; Enami et al., 1986). This *Galdieria* peroxidase is associated with cell wall synthesis/modification and hence may have properties useful in degrading lignocellulosic material (Oesterhelt et al., 2008). The interest in looking at this gene in more detail and characterising it comes from *Galdieria*'s unique ability to grow in extremely unfavourable conditions. This protein was also only present in the secretome sample and not in the cell lysate sample suggesting it may be an extracellular secreted enzyme and thus must function under acidic conditions. Although this peroxidase does not appear to share any predicted function with the typical class II lignin degrading peroxidases (manganese peroxidase (MnP), lignin peroxidase (LiP) and versatile peroxidase (VP)), it displays interesting and industrially relevant features. Exploration of this novel peroxidase may yield more exciting discovery (Abdel-Hamid et al., 2013).

4.4.5.2 Alpha-glucosidase

Both alpha-glucosidases that were present in the secretome sample were also present in cell lysate LC-MS samples (Table 4.4). They contain highly conserved regions linking to glycoside hydrolase family 15 (GH15). This type of enzyme is involved in the breakdown of starch and disaccharides to glucose. More specifically they catalyse the hydrolytic release of α-glucose molecules from the terminal of non-reducing (1- 4) linked α-glucose (Sinnott M. L, 1990). Meaning this group of enzymes typically are exo acting on the hydrolysis of 1-4 α-glucosidic linkages (ExPASy, 2019). They play a critical role enabling the growth of microorganisms that use the released sugars as an energy source (Bandick 1999; Kato et al., 2002). They can also be used in biotechnology with both medical and food industry applications or to conjugate sugars with biologically useful materials (Crittenden and Playne 1996; Eggleston and Cote 2003; Seeberger and Werz 2007). The appearance of these enzymes in both the cell lysate and secretome sample could be an indication of constant low-level secretion, allowing for the release of sugars which aids the induction of uptake transporters (Oesterhelt et al.,1999). Lastly the homology matches suggest it is likely that these two proteins are horizontally acquired, this is not shocking as *Galdieria's* genome has been shown to harbour up to 5 % of genes that are horizontally acquired. It is hypothesised that the use of these acquired

genes is contributing to *Galdieria's* adaptation to living in the extreme environments where it is found (Rossoni et al., 2019).

4.4.5.3 Beta-galactosidase

Two predicted Beta-galactosidases, EC 3.2.1.23 (Gasu_27490 and Gasu_27500) have highly conserved domains of Beta-galactosidases activity (Table 4.5). Once again, the homology searches revealed the high likelihood that these genes have been horizontally acquired. These genes only appear in the secretome sample, heavily suggesting these are extracellular proteins with high acid tolerance. They act on substrates containing galactose specifically β-galactosides, where they hydrolyse the terminal non-reducing β-galactose residues. Thus, they create monosaccharides through the breaking of a glycosidic bond. Beta-galactosidases are commonly used in the food industry, mainly for the removal of lactose from dairy products, however, they are also widely used to improve sweetness, solubility and flavour of various food types (Husain, 2010). More relevantly they are responsible for catalysing the hydrolysis of o-glycosyl bonds in hemicellulose (Blumer-Schuette et al., 2014; Fernández-Bayo et al., 2019). The relevance and value of such enzymes as these in this field has been noted, with patents for enzyme cocktails containing beta-galactosidases existing for lignocellulosic deconstruction and saccharification (Zavrel et al., 2018; Fernández-Bayo et al., 2019).

4.4.5.4 Purple acid phosphatase

Purple acid phosphates (PAPs) are a large family of metalloenzymes found both in plants and animals. They specifically hydrolyse phosphate esters and anhydrides under acidic conditions. The homology searches showed that for Gasu_29970, the highest sequence homology scores came from other red algae. Second to this were numerous hits from the animal kingdom (Table 4.4 and Table 4.5). PAPs are linked to various biological functions, for example mammalian PAPs are associated with roles such as iron transport and generation of reactive oxygen species (Sticklen, 2006; Schenk et al., 2013). There is little literature on the role of PAPs specifically in red algae, or algae in general. It is unlikely that it this gene is involved in the degradation of lignocellulosic material.

4.4.5.5 Uncharacterised

Uncharacterised proteins were included in the secretome profile and for further investigation. With unknown function, they may contain the enzymes that are breaking down the xylan but just be novel and unrecognised. Of the five uncharacterised proteins listed in Table 4.4 and Table 4.5, sequence homology searches showed no information for three of the genes (Gasu_21790, Gasu_24700 and Gasu_41530).

Gasu_21980 showed one conserved SGNH domain (Table 4.5); this was a shared region from all eight hits in the homology searches, all of which were bacteria. This suggests that this gene could be horizontally acquired. SGNH/GDSL esterases and lipases are hydrolases that have been shown to have broad substrate specificity and other multifunctional properties (Akoh et al., 2004; Lai et al., 2017). They have potential uses in an industrial setting for food, chemicals and pharmaceuticals where they are involved in hydrolysing and synthesising important ester compounds (Akoh et al., 2004). It has been found that there are some esterases belonging to the GDSL-family that involved in the degradation of complex polysaccharides (Dalrymple et al., 1997). These enzymes have been shown to exhibit acetyl xylan esterase activity, where O-acetyl groups are removed from the xylose residues in xylan and xylo-oligomers. It has been suggested that the acetyl xylan esterases may contribute to the degradation of plant cell walls acting in connection with cellulases, mannanases and xylanases (Dalrymple et al., 1997). Gasu_21980 is a good candidate gene for further analysis. However, it may not be an extracellular secreted protein as it also appeared in the cell lysate sample.

Lastly *G. sulphuraria* gene Gasu_31410 despite showing no predicted conserved domains, the homology searches revealed some interesting results. This protein had similarity with proteins from a virus, bacteria and archaea. This is an encouraging suggestion it may be a horizontally acquired protein. There was a similarity match with a pectate lyase superfamily protein from *Lechevalieria xinjiangensis*. These pectate lyases cleave heteropolysaccharides based on galacturonic acid (Tamaru and Doi, 2001). These enzymes have been found in multienzyme complexes that contain multiple carbohydrate active enzymes including xylanases (Tamaru and Doi, 2001; van Dyk et al., 2010). In some literature there have been reported links between pectate lyase sequence similarity and predicted xylanase function (Brown et al., 2001; Jorge et al., 2006). *Thermoplasmata* that also shared sequence homology are all acidophiles and thermophiles, this along with the other BLASTp matches is encouraging for gene Gasu_31410 being acid and temperature resistant. The structure predictions from both i-TASSER and AlphaFold showed no matched with any other protein structures. Based on these various reasons, this uncharacterised protein is a good candidate for further investigation and may yield a previously undescribed xylanase.

4.4.6 Conclusion

This chapter aimed to identify putative enzymes produced from *G. sulphuraria* that could be involved in lignocellulosic degradation. In total, 26 genes were initially identified as putative enzymes involved specifically in the breakdown of the hemicellulose xylan.

Results presented were analysed for the selection of enzymes for further study. Target sequences are desired to be extracellularly secreted in order to assure that the enzymes could function under acidic conditions. This revealed 11 candidates suitable for further investigation, among these targets were two alpha-glucosidase, two beta-galactosidase, a peroxidase, a purple acid phosphatase and five uncharacterised proteins. In order to understand these putative polysaccharide degrading enzymes and their function regarding the degradation of lignocellulosic material it's necessary to express them in a heterologous system to assess their biochemical activity.

Chapter 5 - Cloning and Recombinant Protein Production

5.1 Introduction

The heterotrophic growth of *G. sulphuraria* on multiple carbon sources suggests a large repertoire of carbohydrate active enzymes (CAZymes) (see Chapters 2 and 3). It has yet to be confirmed how and when *G. sulphuraria* produces its CAZymes, the two schools of thought are either they are produced simultaneously or that there exists some type of feedback sensing system, where depending on the substrate the alga will produce the enzyme with the required catalytic function. Given the extraordinary bio-versatility demonstrated by *G. sulphuraria* its genome presents an excellent collection of targets with a focus on lignocellulosic degradation. Surprisingly, my previous analysis of the core six genomes of *G. sulphuraria* revealed on average only 124 enzymes that classified as CAZymes (Section 3.3.3). This was reduced to just 14 when looking at enzymes containing signal peptides and predicted hydrolase activity, a much smaller number that expected when considering the organisms vast growth capacities. Additionally, Chapter 4 described the identification of 26 putative polysaccharide degrading enzymes in the xylan grown secretome of *G. sulphuraria*. Based on their presence of a signal peptide and putative activities 11 candidates were chosen for further analysis. Combining these results provided a varied set of genes which contained novel proteins some with putative functions other with unknow possibly new functions. In order to further investigate these enzymes and discover their functions regarding the degradation of lignocellulosic material it's necessary to express them in a heterologous system to assess their biochemical activity.

5.1.1 Heterologous expression

Heterologous expression permits the introduction of a protein of interest from one species into a different host species. This allows for investigations into features of the target protein without the need for protein extraction from the original host, which is often non-trivial (Gomes et al., 2016). Utilising heterologous systems minimises issues in extracting proteins from the host organisms such as low protein levels or difficulties in extracting protein. Furthermore, heterologous approaches are readily scalable to industrial quantities if required. Therefore, as per the reasons given above this study chose heterologous expression over homologous gene expression.

The chosen host organism for heterologous expression can vary from single cell bacteria and yeast to more complex, multicellular fungal and mammalian systems (Peng et al., 2015; Gomes et al., 2016; Kumondai, et al., 2020). There are multiple commercially available expression systems, and the most suitable choice is decided on by factors such

as, post translation modifications, codon bias, presence/absence of disulphide bridges and as well as taking into consideration the characteristics of the target protein (Gellissen, 2006; Gomes et al., 2016). The bacterial recombinant expression system is typically the preferred expression host for many reasons including low cost, high growth, rapid biomass accumulation and a simple process to scale up (Sahdev et al., 2008; Chen, 2012; Khow and Suntrarachun, 2012). A huge amount of research and development has gone into this system resulting in numerous vectors and bacterial strains that have been tailored to benefit expression of various types of proteins and optimising expression (Khow and Suntrarachun, 2012).

The work detailed in this chapter utilised the expression host *Escherichia coli* using two pET28a expression vectors. For the pET28a vector, the target gene was inserted into plasmids multiple cloning site. Upstream of the inserted gene is a T7 promoter and adjacent to this a lac operator sequence (for repressing uninduced expression of the target gene), a 6x histidine tag for protein purification and a thrombin protease recognition site (TPS). Downstream of the target gene is the T7 terminator sequence (Studier and Moffatt, 1986, Rosenberg et al., 1987, William Studier et al.,1990) (Figure 5.1). The second vector pET28a-TIR-2+T7pCONS is an improved design where the restoration of the conserved T7 promoter (T7pCONS) and synthetically evolved translation initiation region (TIR) (TIR-2), showed a greater than two-fold increase in protein production (Shilling et al., 2020). Figure 5.2A highlights the T7 promoter consensus sequence was truncated in the T7*lac* promoter in pET28a and the new $T7p^{cons}$ sequence. Figure 5.2B shows the synthetic evolution of the pET28a-TIR in to the TIR-2 sequence variant (Shilling et al., 2020).

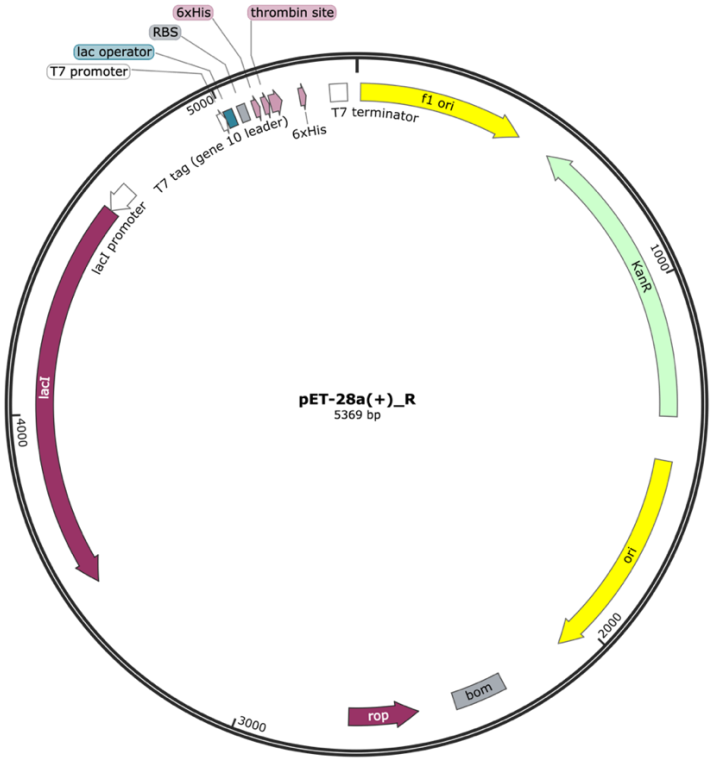


Figure 5.1: Map of the expression plasmid pET28a(+).

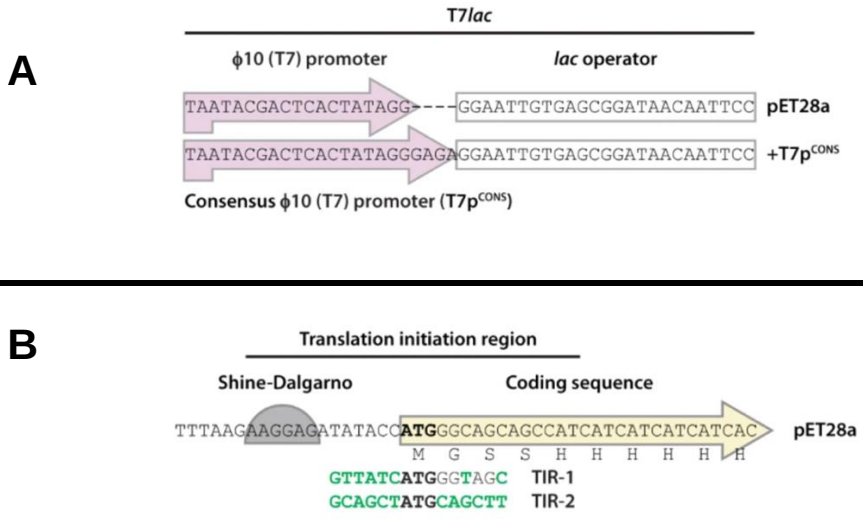


Figure 5.2: (A) The T7 promoter in pET28a is a truncated variant of the consensus T7 promoter (T7pCONS) figure adapted from Shilling et al., 2020. (B) Synthetic evolution of the pET28a-TIR in Shilling et al., 2020 gave two sequence variants (TIR-1 and TIR-2), these altered nucleotides for the TIR variants are shown in green (figure adapted from Shilling et al., 2020).

5.1.2 Solubility of proteins

Solubility is an important factor to consider when expressing and purifying recombinant proteins. The best-case scenario, is one where expressing using a strong promoter will result in the recombinant protein being highly soluble along with producing high yield and strong activity (Correa and Oppezzo, 2015; Singh et al., 2015). However, this is not always the case, and it has been estimated that over 30% of recombinant proteins expressed are insoluble (Papaneophytou and Kontopidis, 2014) The solubility of a recombinant protein can be influenced by multiple factors, such as amino acid sequence of target protein and issues arising from the protein structure having high degree of hydrophobic surface residues. Alongside this, other changeable conditions like the cell lines used, temperature, type of growth media, rate of protein synthesis, protease inhibitors and extraction buffers can also affect the solubility of a protein (Papaneophytou and Kontopidis, 2014; Correa and Oppezzo, 2015; Singh et al., 2015). Certain conditions can be incompatible, hence causing the protein product to be contained within insoluble inclusion bodies. Deconvoluting the appropriate conditions to solubilises protein that are formed in inclusion bodies can be difficult and time consuming (Singh et al., 2015). If this is impossible, then the process of achieving active protein from the inclusion bodies is equally challenging and often will result in low yields.

Alternatively, sometimes inclusion bodies can be beneficial in the production of purified recombinant protein. As the target protein is encapsulated within the aggregate inclusion bodies, the protein is protected from outside inference such as proteases or a change in cellular conditions (Singh et al., 2015). In such cases the target protein can reach a soluble form once the inclusion body is suitably denatured followed by refolding of the denatured target protein into an active native form. Equally purification via inclusions bodies can also function as a purification step, by separating and removing undesired *E. coli* proteins (Correa and Oppezzo, 2015; Singh et al., 2015). Nevertheless, it is not always possible to successfully recover a protein from inclusion bodies. Refolding of a protein under laboratory condition can often lead to misfolding events causing changes in the structure of the final protein product meaning insolubility or inactive protein. This is especially difficult to monitor or measure when working with proteins of unknow function. Taking this into consideration it is best to for each protein target to optimise expression conditions individually and push towards generating soluble forms of the target protein (Papaneophytou and Kontopidis, 2014; Correa and Oppezzo, 2015; Singh et al., 2015).

5.1.3 Aims

In Chapter 2 *G. sulphuraria* genomes were searched informatically for CAZymes, additionally in Chapter 3 the secretome of *G. sulphuraria* grown on xylan was assessed via mass spectrometry. Both these analyses were in the search enzymes linked to the degradation of lignocellulosic biomass that could possibly be industrially relevant. The next step in identifying and characterising these putative enzymes is to produce purified recombinant protein. This chapter will explain the process of attempted cloning three target genes, Gasu_17800, Gasu_27500 and Gasu_31410 into *E. coli* for heterologous expression, purification and testing the folding state of produced proteins.

5.2 Materials and Methods

5.2.1 Media

For the bacterial growth cultures both Super Optimal broth with Catabolite repression (SOC, Sigma Aldrich; 85469) and Lysogeny Broth (LB, Miller, 1972) medias were used.

5.2.2 Plasmids

The expression of the *G. sulphuraria* proteins was carried out using two pET vectors. Firstly pET28a(+) and an adapted version of this pET28a-TIR-2+T7pCONS (Shilling et al,. 2020). The features of each vector are displayed in Table 5.1.

Table 5.1: Information on the plasmids used for cloning and expression.

Vector	Antibiotic resistance	Features	Source
pET28a(+)	Kanamycin 50 μg/mL	Fuses a 6xHis affinity tag to the N-terminus of target protein which is cleavable using thrombin	Novagen

pET28a-TIR-2+T7pCONS	Kanamycin 50 μg/mL	Fuses a 6xHis affinity tag to the N-terminus which is cleavable using thrombin. Restoration of the conserved T7 promoter (T7pCONS) and (2) synthetically evolved TIRs (TIR-1, -2).	Shilling, P.J., Mirzadeh, K., Cumming, A.J. et al., Improved designs for pET expression plasmids increase protein production yield in Escherichia coli. Commun Biol 3, 214 (2020). https://doi.org/10.1038/s42003-020-0939-8

5.2.3 *E. coli* strains

Detailed in Table 5.2 are the *E. coli* strains used in this chapter.

Chemically competent Rosetta (DE3) cells were made according to the Inoue method (Im, 2011).

Table 5.2: Information on the E. coli strains used during cloning and expression.

Strain	Features	Source
DH5α	Routine cloning strain	Invitrogen
Rosetta (DE3)	BL21 derivative designed to enhance the expression of eukaryotic proteins that contain codons rarely used in E. coli	Home-brewed
Alpha select Silver efficiency	α-Select Competent Cells provide recA1 and endA1 markers to minimize recombination and enhance the quality of the plasmid DNA	Bioline

5.2.4 Cloning

For cloning of target proteins, a growth experiment was set up where *G. sulphuraria* was subject to different carbon sources with the aim of extracting RNA to use for cDNA synthesis.

5.2.4.1 Sample collection for RNA

G. sulphuraria 074W cultures were grown under a 12 h:12 h light:dark cycle under 42 µmol m-2 s-1 of light and at 37°C on an orbital shaker (130 rpm). The experimental design followed 11 conditions; details are shown in Table 5.3. Samples of approximately 100 mg were harvested at Day 5, 14 and 21. Upon sampling, cells were washed three times in a PBS pH 7.4 buffer then stored at -80 °C until RNA extractions were carried out.

5.2.4.2 RNA extraction and cDNA synthesis

Cells were ground into a fine powder with a pestle and drill in the presence of liquid nitrogen. RNA was isolated and cleaned up using the Monarch Total RNA Miniprep Kit (New England BioLabs, T2010S). All RNAs were treated with DNaseI (Qiagen) according to manufactures instructions, samples were then pooled. RNA integrity and concentration was assessed using agarose gel electrophoresis and NanoDrop 1000 Spectrophotometer (Thermo Fisher Scientific). Typically, 1.5 % w/v agarose gels were used in 1x Tris-borate-EDTA buffer (45 mM Tris, 45 mM boric acid, 1 mM EDTA - TBE) stained with ethidium bromide. A 1 kb DNA Ladder (Generuler) was used as a molecular weight marker. Nucleic acids were visualised with UV light. cDNA was synthesised from RNA collect as describe above. For the conversion of RNA to cDNA SuperScript® IV Reverse Transcriptase (Invitrogen) was used according to the manufacturer instructions.

5.2.4.3 Polymerase Chain reaction (PCR) Primers

Oligonucleotide primers for PCR were designed for the amplification of gene targets to meet several criteria. Each primer was required to have a melting temperature (T_m) between 50-60 °C where possible. Additionally, the T_m of primer pairs where possible were to be no more than 3 °C different. Again, where feasible each primers GC % should not surpass 60 % and ideally be kept as low as possible. To calculate T_m values as recommended for the Phusion DNA polymerase the Thermo Scientific T_m Calculator was used.

Table 5.3: Growth conditions for 074W cultures for RNA extraction. All media was pH 2. Allen medium made according to Table 2.1.

Condition	Carbon source	Media
Starvation	NA	dd H20
Autotrophic	NA	Allen's
Mixotrophic	10 g/L Sucrose	Allen's
Mixotrophic	20 g/L Xylan	Allen's
Mixotrophic	10 g/L Cellulose	Allen's
Mixotrophic	10 g/L Lignin	Allen's
Mixotrophic	4 g/L Xylan, Cellulose & Lignin	Allen's
Heterotrophic	10 g/L Sucrose	Allen's
Heterotrophic	20 g/L Xylan	Allen's
Heterotrophic	10 g/L Cellulose	Allen's
Heterotrophic	10 g/L Lignin	Allen's
Heterotrophic	4 g/L Xylan, Cellulose & Lignin	Allen's

5.2.4.4 PCR for cloning

Using Phusion® Hot Start High- Fidelity DNA Polymerase (Thermo Fisher Scientific) amplification of plasmid backbones and target genes was carried out with an Eppendorf Mastercycler. Conditions are shown in **Error! Not a valid bookmark self-reference.**. Primers for PCR were designed using the CDS sequences available from the 074W strain available on GenBank (GCA_000341285.1) not including the predicted signal peptides (SignalP 5.0) or the available plasmid map of pET28a(+). Due to the linearisation site of the plasmids only one pair of primers was required for both pET vectors. Primers for the genes of interest were designed with overhangs used for annealing the insert into the pET28a(+) vectors (Table 5.5).

5.2.4.5 Extraction and purification of PCR Products

PCR products were visualised by agarose gel electrophoresis, target bands were excised from the gel and contents were extracted and purified using a QIAquick® Gel Extraction kit (Qiagen) according to the manufacturer's instructions. Purified DNA fragments were eluted using 40 µl of Buffer EB (Qiagen).

Table 5.4: PCR reaction setup using Phusion® Hot Start High-Fidelity DNA Polymerase (Thermo Fisher Scientific) along with Eppendorf thermocycler conditions.

Reaction Mix (50 μl reaction)	Conditions
10 μl HF Buffer	1X: • 95 °C for 1 minute 25 X: • 96 °C for 15 seconds • T_m for 40 seconds • 72 °C for 30 seconds/kb 1X: • 72 °C for 10 minutes Store at 4 °C
1μl 10 μM DNTPs	
2.5 μl 10 μM Primers F/R	
1.5 μl Template	
0.25 μl Phusion	
35.5 μl ddH_2O	

Table 5.5: Primers used for amplification of gene targets and linearisation of plasmid backbones. Lowercase sequences indicate the overhang sequence on the pET28a(+) expression vectors.

Primer	Sequence (5' 3')
Gasu_31410_F	tggtgccgcgcggcagccatTTTCAAAAGTTTCCACATACTC
Gasu_31410_R	cagcttcctttcgggctttgCTAAGAAGACGAAATAATATTGAG
Gasu_17800_F	tggtgccgcgcggcagccatCAATGTTCGGAAGGTACTATTAAAG
Gasu_17800_R	cagcttcctttcgggctttgCTATATTGGGAAGAAAACAGG
Gasu_27500_R	tggtgccgcgcggcagccatTATAATGGTACAGGGGTACC
Gasu_27500_R	cagcttcctttcgggctttgTCAGTTGTTGCAACCACATC
pET28a-_F	caaagcccgaaaggaagctgagttg
pET28a-_R	atggctgccgcgcggcaccaggccgctg

5.2.4.6 Preparation of Plasmids

Using the appropriate antibiotics cultures of 5 mL of LB were inoculated with bacteria containing the plasmid of interest. Cultures were grown overnight for approximately 16 hours at 200 rpm and 37 °C. Cells were harvested via centrifugation then using

Monarch® Plasmid Miniprep Kit (New England BioLabs, T1010) the plasmid DNA was extracted according to manufactures instructions. During the final step plasmid DNA was eluted in 40 µl of Monarch® DNA Elution Buffer (New England BioLabs).

5.2.4.7 Transformation of *E. Coli* Cells

An aliquot of cells were removed form -80 °C and thawed on ice. Simultaneously 5 µl of a ligation reaction or 1 µl of a plasmid preparation were added to a sterile 1.5 mL tube and stored on ice. Then 50 µl of thawed cells were added to the 1.5 mL tube (containing the DNA) and the tube then gently mixed. The tubes were then incubated on ice 30 minutes on ice. Then tubes were heat shocked at 42 °C in a water bath for 30 seconds. Tubes were then returned to ice and incubated for another 2 minutes, before adding 500 µl of SOC medium. Cells were then incubated at 37 °C for 1 hour with shaking at 200 rpm. Finally, cells were plated onto 1.5 % LB agar plates containing the suitable antibiotic and incubated overnight (~16 hours) at 37 °C.

5.2.4.8 Colony Screen

To identify colonies that contained the transformed DNA, multiple colonies were screened using PCR. This was achieved by choosing an individual colony under sterile conditions and transferring it into 10 µl of deionised H_2O and using this as a template for a PCR. Table 5.6 shows the reaction mixture and conditions used for PCR. Afterwards 5 µl of each PCR product was assessed for the target DNA via an agarose gel.

5.2.4.1 NEBuilder® High-Fidelity DNA Assembly Cloning

G. sulphuraria target genes were cloned into both the pET28a(+) and the pET28a-TIR-2+T7pCONS vectors using the NEBuilder® Hi-Fi DNA Assembly Cloning kit (New England Biolabs, E5520S) according to manufacturer's instructions. Prior to the cloning reaction, the plasmid template used in the PCR reaction was degraded by a Dpn1 digest following the instructions in the NEBuilder® HiFi DNA Assembly Cloning Kit protocol. The resulting mixture was then transformed into a-select silver efficiency chemically competent *E. coli* Table 5.2 and any transformants confirmed via colony PCR as described above.

Table 5.6: PCR reaction setup using Taq Polymerase (Thermo Fisher Scientific) along with Eppendorf thermocycler conditions.

Reaction Mix (10 μl reaction)	Conditions
1 μl 10X Taq Buffer	1X: • 95 °C for51 minutes 30 X: • 94 °C for 30 seconds • 50 °C for 45 seconds • 72 °C for 30 seconds/kb 1X: • 72 °C for 10 minutes Store at 4 °C
1 μl 2.5 μM DNTPs	
0.5 μl 10 μM Primer F	
0.5 μl 10 μM Primer R	
1 μl Template	
0.05 μl Taq polymerase	
5.95 μl ddH_20	

5.2.4.2 DNA Sequencing and Analysis

Plasmid DNA was sequenced by Sanger sequencing services from GATC (Eurofins) following manufacturer's instructions. Primers used for sequencing are given in Table 5.7. The chromatograms obtained were analysed using 4Peaks software (Nucleobytes) to confirm correct gene sequence in plasmid.

Table 5.7: Primers used for DNA sequencing for confirmation of correct plasmid sequence.

Primer	Sequence (5' 3')
Gasu_31410_seq1_fwd	ACCTCAGTGAAGAAACCTG
Gasu_31410_seq1_rev	TTCCAACCTCCGGGAGTTG
Gasu_31410_seq2_fwd	AACTATATCGCAGATCTTC
Gasu_31410_seq2_rev	ATGGTATTCGAATCAATGC
Gasu_17800_seq1_fwd	TATCAGCTCTTATCTCTGC
Gasu_17800_seq1_rev	AAGGTATTATTATGACCAG
pET28a_seq_fwd	attgtgagcggataacaattc
pET28a_seq_rev	atccggatatagttcctcc

5.2.5 Protein expression

To express the protein, Rosetta DE3 were transformed with plasmids identified as containing the correct insertion of Gasu_31410 and Gasu_17800. Transformants were then grown in liquid LB cultures with appropriate antibiotics at 37 °C until the OD_{600} reached 0.5 - 0.8 (~2 hours). A 1 mL sample was then taken from each culture (T_0). To each of the remaining culture, isopropyl βD-1-thiogalactopyranoside (IPTG) was added to a 0.5 mM concentration for the induction of protein expression. Cultures were grown then at either 37 °C or 18 °C before samples were collected after various hours for each condition. When samples were collected, they were immediately centrifuged for 1 minute at 13,200 rpm and then the supernatant discarded. The pellet was then resuspended in 100 µl of SDS loading buffer (Laemmli, 1970) and heated 95 °C for 10 minutes. All samples were stored at -20 °C until the SDS-PAGE gel was run. After the final sample was taken, the remaining culture was aliquoted into 2 mL samples and centrifuged at 13,200 rpm. The supernatant was discarded, and samples snap frozen in liquid nitrogen before being stored at -80 °C for solubility testing.

5.2.5.1 Soluble Protein Expression Testing

E. coli samples from the -80 °C were tested for solubility of expressed protein. Samples were thawed on ice and resuspended in PBS pH 7.4 or pH 2.5 buffer at either 1x or 10x, then the cells were lysed through sonication (BANDELIN, 3x 15 s, 20% amplitude). One set of samples were then incubated at 60 °C for 1 hour. The soluble fraction was separated through centrifugation for 10 minutes at 20,000 g. Then 50 µl of supernatant was added to 50 µl of 2x Laemmli sample buffer (Laemmli, 1970), then heated to 95 °C for 5 minutes. Samples were then visualised with an SDS-PAGE as described in (Section 4.2.4).

5.2.5.2 Inclusion body preparation

Multiple different buffers were assessed for increasing the solubility of the target proteins. This involved the preparation of the inclusion bodies using buffer 1 from Table 5.8. Pellets from - 80 °C were resuspended on ice with 1 mL of buffer 1 with 1 % Triton X-100 for 30 minutes. Cells were then lysed through sonicating (BANDELIN, 3x 15 s, 20% amplitude), before centrifuging (4 °C, 30 m, 20,000 g) and removing the supernatant. Pellets were washed three times in buffer 1 with 2% Triton X-100, with a final wash step using buffer 1 (no Triton X-100). Supernatant was removed and pellets stored at - 80 °C, these are the inclusion bodies.

Table 5.8: Composition of base buffer used in multiple steps of purification.

Buffer 1
100 mM Tris-HCl pH
50 mM NaCl
10% Glycerol
10 mM Imidazole

5.2.5.3 Denaturing conditions

These inclusion bodies pellets were tested for solubility with buffers containing increasing concentration of urea, this was buffer 1 (Table 5.8) supplemented with either 5, 6, 7 or 8 M urea. The denaturing urea buffers were added, and samples incubated on ice for 30 minutes. Samples were then centrifuged (4 °C, 30 m, 20,000 g) and lastly 50 µl of supernatant was added to 50 µl of 2x Laemmli sample buffer (Laemmli, 1970), then heated to 95 °C for 5 minutes. Samples were then visualised with an SDS-PAGE as described in (Section 4.2.4).

5.2.6 Protein Purification

Successfully expressed protein targets Gasu_17800 and Gasu_31410 were purified using Ni-NTA Purification System (ThermoFisher), using 2 mL bed volume Poly-Prep® Chromatography Columns (BioRad), with Ni-NTA Agarose. Ni-NTA Agarose uses nitrilotriacetic acid (NTA), a tetradentate chelating ligand, in a highly cross-linked 6% agarose matrix. NTA binds Ni2+ ions by four coordination sites.

5.2.6.1 Preparing cell lysate under native conditions

Cell pellets were removed from -80 and resuspended in buffer 1 (Table 5.8), each sample was then incubated on ice for 30 minutes. Cells were then lysed through sonication (BANDELIN, 3x 15 s, 20% amplitude), before centrifuging (4 °C, 30 m, 20,000 g) and transferring supernatant to a pre-chilled tube.

5.2.6.2 Solubilisation/denaturation of proteins from inclusion bodies

Inclusion bodies were resuspended in buffer 1 (Table 5.8) supplemented with 7 M urea. Each sample was then incubated on ice for 30 minutes. Insoluble material was then

removed by centrifugation (4 °C, 30 m, 20,000 g) and soluble material transferred to a pre-chilled tube.

5.2.6.3 Preparation of column

To prepare the column, Ni-NTA Agarose was resuspended by gently inverting the bottle repeatedly. Then 1-2 mL was carefully pipetted into a Poly-Prep® column and the resin allowed to settle completely (10 minutes). The column was then washed three times with 10 x column volume (CV) of buffer 1 either with or without 7 M urea depending on purification conditions.

5.2.6.4 Purification under Native conditions

The cell lysate supernatant was slowly (~ 1mL/min) loaded onto the prepared column. The flowthrough was then collected and reloaded onto the column again. Next the column was washed with the lysis buffer (buffer 1; Table 5.8) (~10x CV). The protein was then eluted in three steps with buffer 1 (Table 5.8) containing increasing concentration of imidazole (50, 150 and 300 mM) (~5x CV). Samples stored at 4 °C

5.2.6.5 Denatured conditions with refolding

The prepared inclusion bodies were slowly (~ 1mL/min) loaded onto the prepared column. The flowthrough was then collected and reloaded onto the column again. Next the column was washed with buffer 1 (Table 5.8) supplemented with 7 M urea (~10x CV). The protein folding was then attempted on the column by washing the column with buffer 1 supplemented with a gradual decrease in urea concentration. This was done over 11 steps with concentrations of 7, 6.5, 6, 5.5, 5, 4.5, 4, 3.5, 3, 2, 0 M of urea. Then the protein was eluted in three steps with buffer 1 (Table 5.8) containing increasing concentration of imidazole (50, 150 and 300 mM) (~5x CV). Samples stored at 4 °C.

5.2.6.6 Qubit protein quantification

Protein quantification was determined using Qubit® Fluorometric Quantification (ThermoFisher) with the Qubit® Protein Assay (ThermoFisher, Q33212) according to manufactures instructions. Standard samples were freshly made at each assessment.

5.2.6.7 Protein melting temperature determination

The folded state and refolding of the protein was determined by assessing the protein's melting temperature using nanoscale differential scanning fluorimetry (nanoDSF). The Prometheus NT.48 instrument (NanoTemper Technologies) was used. The capillaries were filled with 10 µl sample and placed on the sample holder. A temperature gradient

of 1 °C·min^{-1} from 20 to 90 or 98 °C was applied and the intrinsic protein fluorescence was recorded at 330 and 350 nm.

5.2.7 Refolding screen

The setup of buffers tested is detailed in Figure 5.3 and compositions of these buffers are in Supplementary Table 4.1 (adapted from Wang et al., 2017). The denatured target protein was refolded by shock dilution at a ratio of 1:20. To each well of a 96-well plate (U-bottom, clear) 10 µl of denatured protein solution at ~5 mg/mL was added. Subsequently, from a pre-pared master plate 190 µl of each refolding buffer was then transferred to the plate, each well contained a total volume of 200 µl. The plate was subsequently incubated at room temperature with mild shaking for 1 hour. Absorbance was measured at 340, 360, and 600 nm in a Clario BiostarPLUS 96 well plate reader (BMG Labtech). A control plate of protein sample buffer and each buffer was also measured. The readings from the plates adjusted for background noise using the control plate. Subsequently, samples for further testing were determined by if there was >+/- 0.01 difference in measurements across all three wavelengths.

	pH	2	2.5	3	3.5	4	4.5	5	5.5	6	6.5	7	7.5	
Buffer		1	2	3	4	5	6	7	8	9	10	11	12	Arginine
GHC/MIB	A	GHC	GHC	GHC	GHC	MIB	MIB	MIB	MIB	MIB	MIB	MIB	MIB	-
PCB	B	PCB	PCB	PCB	PCB	PCB	PCB	PCB	PCB	PCB	PCB	PCB	PCB	
HCPC/PHP/MMT	C	HCPC	PHP	PHP	PHP	MMT	MMT	MMT	MMT	MMT	MMT	MMT	MMT	
***	D	PBS	PBS	PBS	Citric Acid	Citric Acid	Citric Acid	Citric Acid	MES	MES	MES	Tris	Tris	
GHC/MIB	E	GHC	GHC	GHC	GHC	MIB	MIB	MIB	MIB	MIB	MIB	MIB	MIB	+
PCB	F	PCB	PCB	PCB	PCB	PCB	PCB	PCB	PCB	PCB	PCB	PCB	PCB	
HCPC/PHP/MMT	G	HCPC	PHP	PHP	PHP	MMT	MMT	MMT	MMT	MMT	MMT	MMT	MMT	
***	H	PBS	PBS	PBS	Citric Acid	Citric Acid	Citric Acid	Citric Acid	MES	MES	MES	Tris	Tris	

Figure 5.3: Composition of the pH refolding screen in a 96-well plate (adapted from Wang et al., 2017). Buffer concentration: 50 mM, salt concentration 100 mM and Arginine concentration 0.4 M. GHC: Glycine, MIB: sodium malonate, imidazole and boric acid, PCB: Phosphate Citrate, HCPC: Potassium Chloride, PHP: Potassium Hydrogen Phthalate, MMT: DL-malic acid, MES and Tris-HCl, ***: pH 2-3 (PBS), pH 3.5–5 (Citric acid), pH 5.5–6.5 (MES), pH 7–7.5 (Tris-HCl).

5.3 Results

5.3.1 Selection for cloning

Using the results obtained from the CAZyme analysis (Chapter 3) and the xylan grown 074W secretome (Chapter 4), collectively there were 21 putative enzymes involved in the breakdown of carbohydrates and in particular xylan (Table 5.9). There were 10 enzymes only present in the CAZyme analysis, 7 enzymes only present in the secretome analysis and 4 enzymes that occurred in both lists. This list of enzymes provided a good list to investigate for finding potentially interesting enzymes for industrial contexts. However, due to time constraints, it was decided to focus on three enzymes from this list, so Gasu_31410, Gasu_17800 and Gasu_27500 were selected for cloning. The reasoning for this selection is detailed in Section 5.3.

5.3.2 Cloning

To isolate the required coding sequence (CDS) for heterologous expression, *G. sulphuraria* was grown on a mixture of different growth mediums for 21 days. After 5, 14 and 21 days, RNA was extracted from each of these growth conditions and pooled together. This RNA pool was subsequently used in cDNA synthesis to generate the template for PCR. Three different annealing temperatures (T_m = 50.2, 52.4, 55.2 °C) were tested in the PCR reaction. In all three, Gasu_17800 and Gasu_31410 successfully produced bands (Figure 5.4A). Gasu_31410 gene had an expected band size of 1105 bp, which is consistent with the band shown in Figure 5.4A. Alternatively Gasu_17800 had an expected band size of 931 bp and showed to have two bands one just above ~1000 bp and one just below, the band below 1000 bp was the desired band. Gasu_27500 showed no viable bands in any PCR reactions and hence was not taken further. All bands were extracted from the gel and purified and then quantified for the next steps.

To insert the target genes, I linearised the expression vectors (pET28a and pET28a-TIR-2+T7pCONS) via PCR where the primers were designed to amplify around the insertion site (Table 5.5), this is shown in Figure 5.4B both pET28a and pET28a-TIR-2+T7pCONS showed bands of expected size (~5252 bp). These bands were extracted from the gel then purified and then quantified for the next steps. To stop leftover template interfering with subsequent cloning steps the PCR product was Dpn1 treated to remove the methylated plasmid DNA template.

The amplified targets were then cloned using the NEBuilder® Hi-Fi DNA Assembly Cloning kit (New England Biolabs, E5520S) into both expression vectors. The resulting vectors were then transformed into a-select silver efficiency chemically competent *E. coli*. Several colonies for each of the target genes were chosen and validated through colony PCR to identify positive clones containing the plasmid with the insert (Table 5.5). Results of this are shown in Figure 5.5 where for target Gasu_17800 (Figure 5.5A) there were 5 positive colonies for pET28a and 4 for pET28a-TIR-2+T7pCONS. Similarly, Gasu_31410 (Figure 5.5B) showed 5 positive colonies for pET28a and 4 for pET28a-TIR-2+T7pCONS. Plasmids were purified from these positive colonies and sent for Sanger sequencing (Eurofins) to confirm correct target sequence. Sequencing primers were designed to have complete coverage over the whole of each insert (Table 5.7). Both Gasu_17800 and Gasu_31410 were successfully cloned into both pET28a and pET28a-TIR-2+T7pCONS vectors ready for recombinant protein production.

Table 5.9: Putative enzymes involved in lignocellulosic degradation from *G. sulphuraria*. Compiled from xylan grown secretome and informatic CAZyme analysis.

Gene Name	Gene Ontology	Present in secretome analysis	Present in CAZyme analysis	CAZyme Family
Gasu_01530	UDP-glucose:glycoprotein glucosyltransferase activity [GO:0003980]; protein glycosylation [GO:0006486]	×	✓	GT24
Gasu_05550	carbohydrate binding [GO:0030246]; glucan 1,3-alpha-glucosidase activity [GO:0033919]; carbohydrate metabolic process [GO:0005975]	×	✓	GH31
Gasu_06640	glucan 1,4-alpha-glucosidase activity [GO:0004339]; carbohydrate metabolic process [GO:0005975]	×	✓	GH15
Gasu_12000	integral component of membrane [GO:0016021]; transferase activity [GO:0016740]	×	✓	GT8
Gasu_17790	heme binding [GO:0020037]; metal ion binding [GO:0046872]; peroxidase activity [GO:0004601]; response to oxidative stress [GO:0006979]	×	✓	AA2
Gasu_17800	heme binding [GO:0020037]; peroxidase activity [GO:0004601]; response to oxidative stress [GO:0006979]	✓	✓	AA2
Gasu_21790	NA	✓	×	NA
Gasu_21980	NA	✓	×	NA
Gasu_24700	NA	✓	×	NA
Gasu_25520	glucan 1,4-alpha-glucosidase activity [GO:0004339]; polysaccharide metabolic process [GO:0005976]	✓	×	NA

Gasu_25530	glucan 1,4-alpha-glucosidase activity [GO:0004339]; polysaccharide metabolic process [GO:0005976]	✓	✓	GH15
Gasu_26360	integral component of membrane [GO:0016021]; transferase activity [GO:0016740]	×	✓	GT8
Gasu_27490	beta-galactosidase activity [GO:0004565]; carbohydrate metabolic process [GO:0005975]	✓	✓	GH35
Gasu_27500	beta-galactosidase activity [GO:0004565]; carbohydrate metabolic process [GO:0005975]	✓	✓	GH35
Gasu_29970	acid phosphatase activity [GO:0003993]; metal ion binding [GO:0046872]	✓	×	NA
Gasu_31410	NA	✓	×	NA
Gasu_41530	NA	✓	×	NA
Gasu_47280	hydrolase activity, hydrolyzing O-glycosyl compounds [GO:0004553]	×	✓	GH30_5
Gasu_48600	catalytic activity [GO:0003824]; carbohydrate metabolic process [GO:0005975]	×	✓	GH13
Gasu_52030	endoplasmic reticulum quality control compartment [GO:0044322]; membrane [GO:0016020]; calcium ion binding [GO:0005509]; mannosyl-oligosaccharide 1,2-alpha-mannosidase activity [GO:0004571]; carbohydrate metabolic process [GO:0005975]; endoplasmic reticulum mannose trimming [GO:1904380]; mannose trimming involved in glycoprotein ERAD pathway [GO:1904382]	×	✓	GH47
Gasu_64540	integral component of membrane [GO:0016021]; transferase activity, transferring glycosyl groups [GO:0016757]	×	✓	GT34

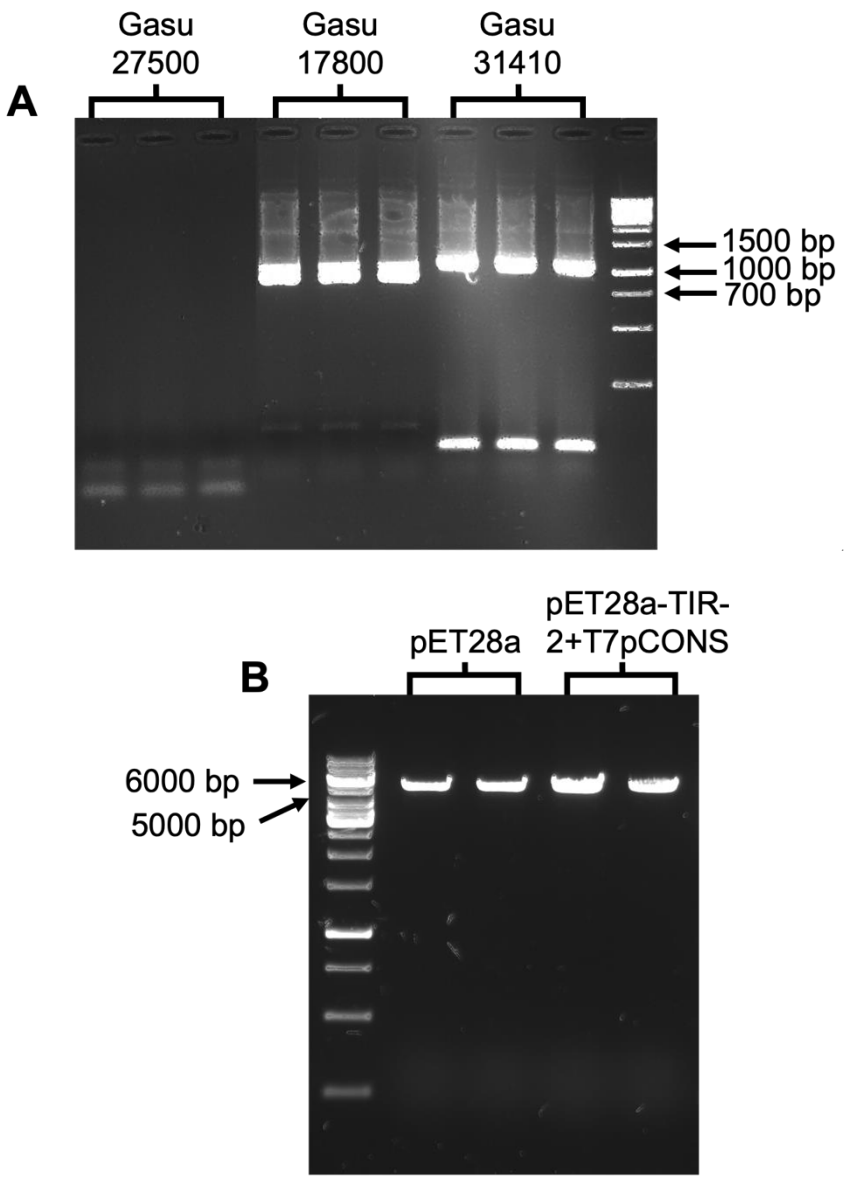


Figure 5.4: PCR products form cloning steps. (A) Amplification of target genes Gasu_17800, Gasu_27500 and Gasu_31410 under a range of annealing temperatures (Tm = 50.2, 52.4, 55.2 °C). The expected theoretical size of the targets are: Gasu_17800: ~931 bp; Gasu_27500: ~3094 bp; Gasu_31410 ~1105 bp. (B) PCR products of linearisation of the pET28a and pET28a-TIR-2+T7pCONS (Shilling et al., 2020) vectors. The expected theoretical size of the targets are: pET28a: ~5252 bp; pET28a-TIR-2+T7pCONS: ~5256 bp.

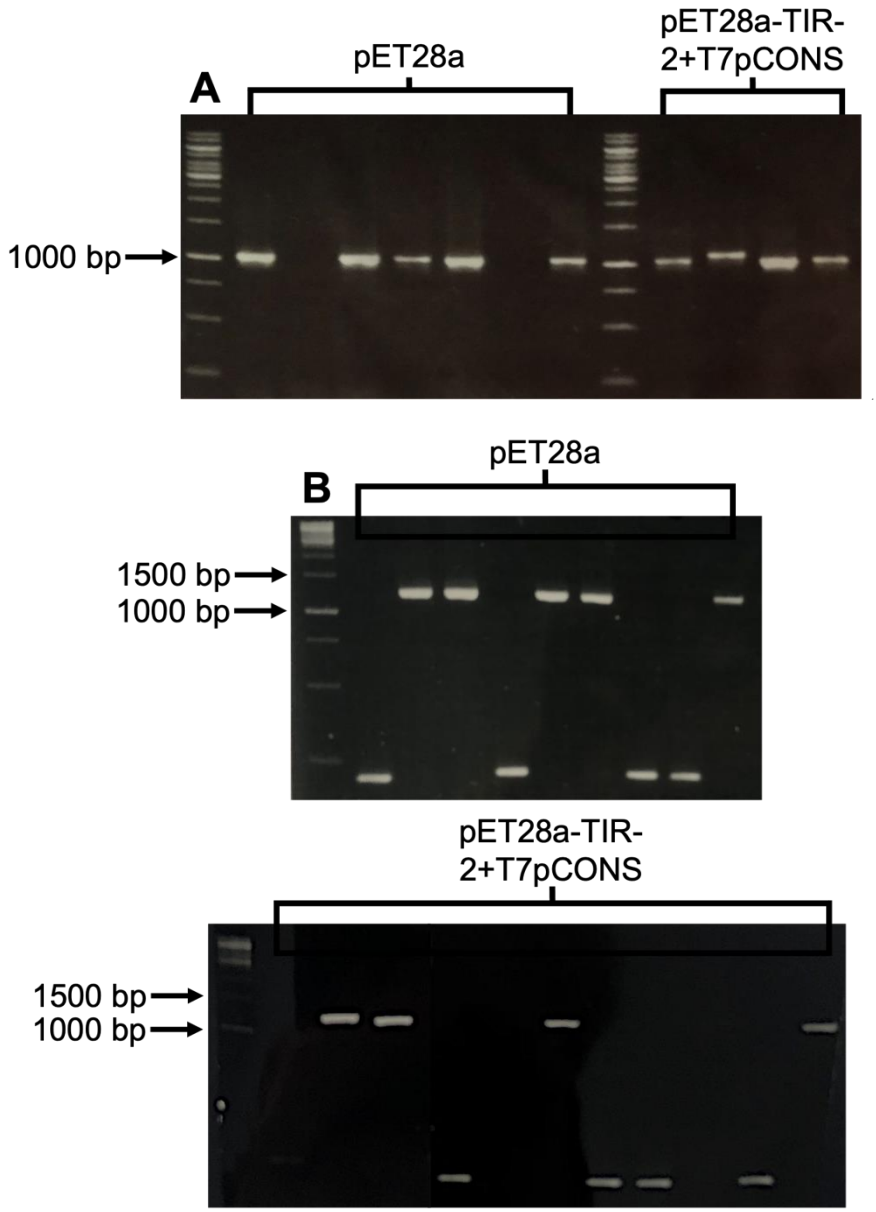


Figure 5.5: Products of a PCR reaction obtained from colony screen for the insertion of Gasu_17800 (A) and Gasu_31410 (B) into pET28a and pET28a-TIR-2+T7pCONS vectors. The PCR reaction was preformed using colonies selected from transformation as a template and primers stated in Table 4.5 and 4.8. The expected theoretical size of the targets are ~1036 bp and ~1200 bp retrospectively.

5.3.3 Recombinant protein production

5.3.3.1 Expression

To determine the best expression host for protein induction trials, I firstly analysed the coding sequence for the two targets to determine the use of rare codons as this can present problems for gene expression in *E. coli* (https://www.genscript.com/tools/rare-codon-analysis). This analysis revealed both Gasu_17800 and Gasu_31410 had a Codon Adaptation Index (CAI) of 0.64, meaning both genes had a high amount of rare codons present (CAI = 1.0 is ideal, while a CAI >0.8 is considered as good for expression *E. coli*). With this information, I decided to use Rosetta (DE3) (Table 5.2). This strain of *E. coli* is a BL21 derivative purposely designed to improve expression of eukaryotic proteins that use rare codons. To evaluate the most suitable expression vector to move forward with, expression trails were performed (data not shown). This revealed that pET28a and pET28a-TIR-2+T7pCONS vectors produced similar amounts of protein. Taking this into consideration along with the work published in Shilling et al., 2020 I decided to only focus on the pET28a-TIR-2+T7pCONS vector for further work.

To move forward it was necessary to check expression levels in both genes post induction for optimal harvesting time. Induced cultures were shifted to 18 °C to increase the likelihood of overexpressing soluble protein. To evaluate the rate of expression samples were taken over a 24-hour window and visualised via. SDS-PAGE, detailed in Figure 5.6. Both targets show the expression of the proteins increasing from the pre induction sample (T_0) and over the 24 hours. It is worth mentioning here that Figure 5.6A shows Gasu_17800 protein to appear just above the 40 kDa marker ~7 kDa higher than the expected size of 33.99 kDa for this protein.

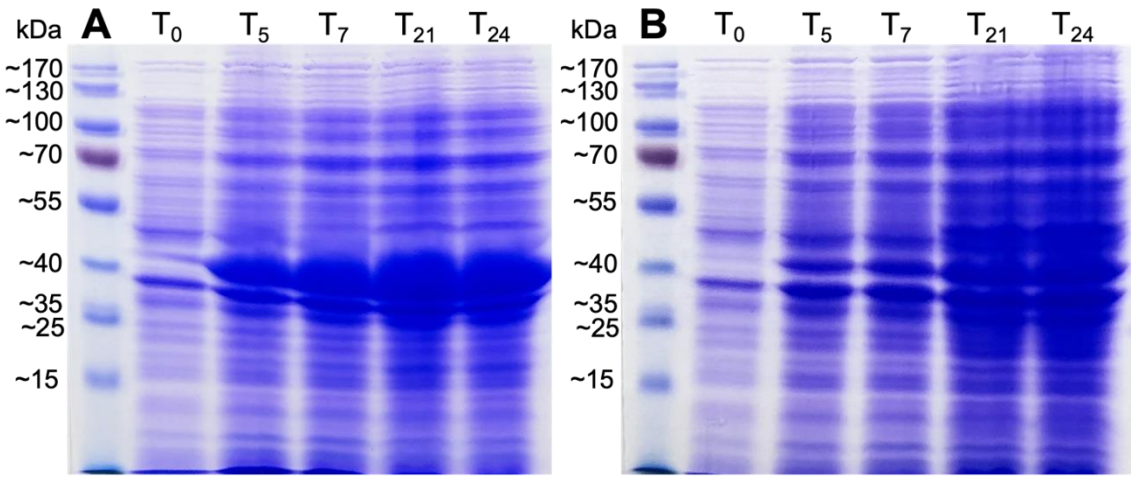


Figure 5.6: SDS-PAGE for ITPG induction at 18 °C for proteins Gasu_17800 (A) and Gasu_31410 (B). Ladder is PageRuler Plus Pre-stained Protein Ladder from Thermo Scientific; T0 is the samples pre induction then T5,7,21,24 are the hours post induction when samples were collected. The expected size for protein A is ~33.99 kDa and for protein B is ~40.58 kDa.

5.3.3.2 Solubility testing

To purify target genes, they firstly need to be soluble. To investigate the solubility of these expressed targets an initial experiment was set up. The results were visualised via SDS-PAGE in Figure 5.7 which shows the soluble fraction from both Gasu_17800 (Figure 5.7A) and Gasu_31410 (Figure 5.7B). All samples using PBS at pH 7.4 gave an increase in soluble material compared to pH 2.5, while all samples that underwent the 60 °C incubation or increasing the concentration to 10x decreased the amount of soluble material. There was no improvement in increasing solubility and thus the experiment was unsuccessful in providing soluble material to proceed with.

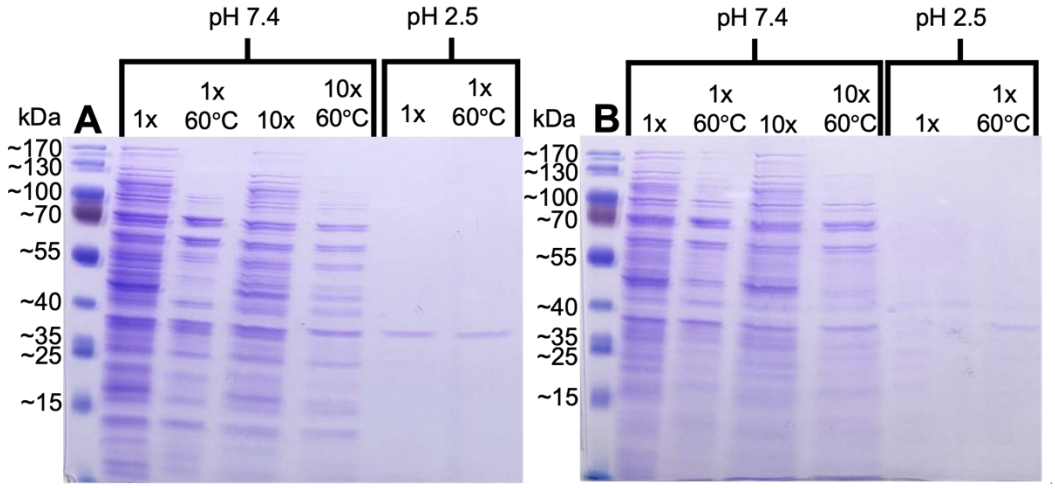


Figure 5.7: SDS-PAGE soluble fraction for room temperature Rosetta DE3 ITPG induced for proteins Gasu_17800 (A) and Gasu_31410 (B). Ladder is PageRuler Plus Prestained Protein Ladder from Thermo Scientific; Samples were sonicated in different buffers with and without an incubation at 60 °C, buffers were all PBS at different concentrations (1x and 10x) and pH (7.4 and 2.5). The expected size for protein A is ~33.99 kDa and for protein B is ~40.58 kDa.

From the results of the solubility assay it was determined that the recombinant proteins were formed in inclusion bodies (IB). In order to purify target proteins, it is necessary to isolate and purify the IBs before solubilising. This in turn will allow for purification and eventually refolding.

With the aim of isolating and purifying the IB cell samples were prepared according to Section 5.2.6.1. An assessment of this along with solubilisation of the IBs was carried out, this involved in testing different concentrations of urea (Section 5.2.5.3). Visualised in Figure 5.8 both targets Gasu_17800 (Figure 5.8A) and Gasu_31410 (Figure 5.8B) were solubilised in all concentrations of urea, and the amount of protein increased as the concentration of urea increased. The sample before IB preparation at T_{20} shows a good comparison of where the desired band was expected and indicates the amount of other *E. coli* proteins that were present. The IB preparation has decreased the number of other proteins present in the sample, which will likely make downstream purification easier. For further purification purposes, the IB solubilisation buffer that was used contained 7 M of urea.

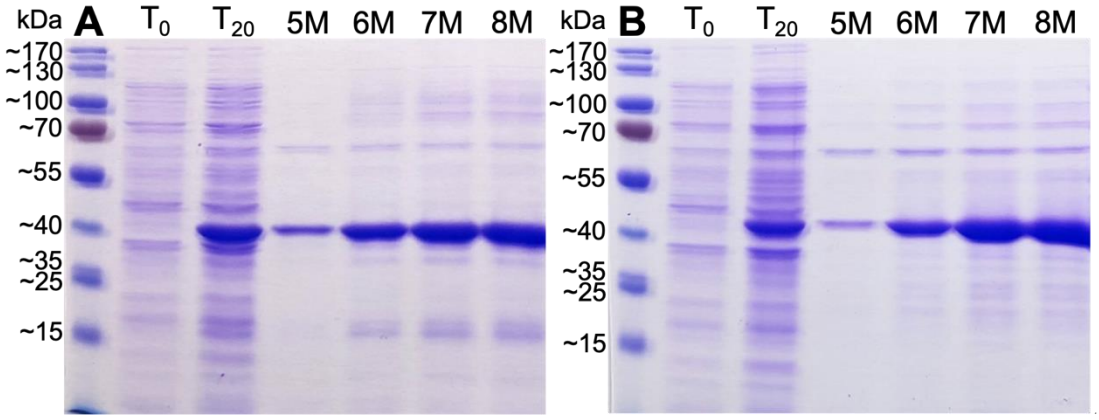


Figure 5.8: SDS-PAGE showing solubilisation of inclusion bodies from room temperature Rosetta DE3 ITPG induced for proteins Gasu_17800 (A) and Gasu_31410 (B). Ladder is PageRuler Plus Prestained Protein Ladder from Thermo Scientific. Inclusion bodies were solubilised in 5, 6, 7 and 8 M urea. T_0 is the cell samples pre induction and T_{20} are 20 hours post induction for comparison and guide. The expected size for protein A is ~33.99 kDa and for protein B is ~40.58 kDa.

5.3.3.3 Protein purification

In order to begin characterising these enzymes, first it was required to purify these proteins. Since the N-terminal of pET28a-TIR-2+T7pCONS vector consists of a hexahistidine tag (His-tag) the proteins were purified using affinity chromatography, namely immobilized metal ion affinity chromatography (IMAC). Here, a Ni-NTA resin was used to bind the his-tag containing proteins (under both native and denatured conditions) before a series of washes were applied and then the protein was eluted. Protein purification was carried out under denatured conditions using Ni-NTA Purification System (ThermoFisher) (details in section 5.2.6). Refolding of the proteins was attempted on the column by washing the bound protein in the buffer with decreasing concentrations of urea. For each purification step (IB pellet, flow through, wash and elution) an aliquot was collected and analysed by SDS-PAGE shown in Figure 5.9.

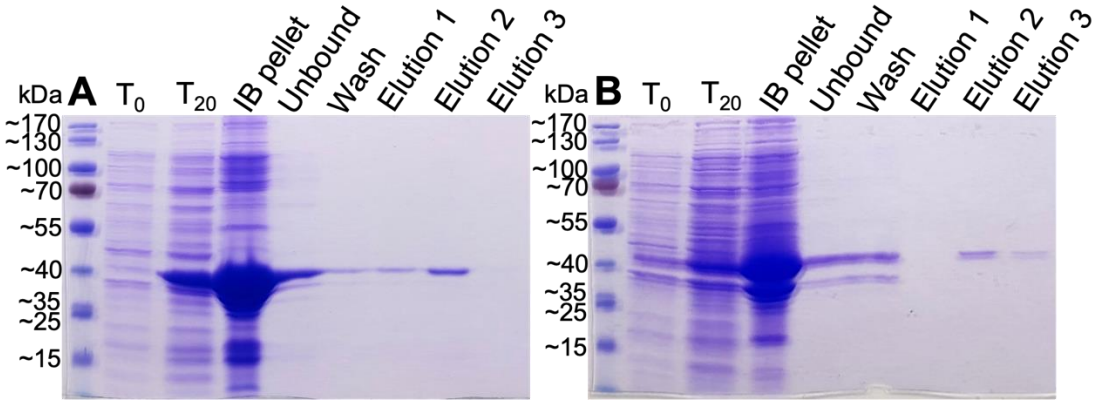


Figure 5.9: SDS-PAGE showing each step of purification of target proteins Gasu_17800 (A) and Gasu_31410 (B) in denatured conditions with refolding on the column. Ladder is PageRuler Plus Pre-stained Protein Ladder from Thermo Scientific. Elution steps 1, 2 and 3 were carried out with 50, 150 and 300 mM imidazole retrospectively. T_0 is the cell samples pre induction and T_{20} are 20 hours post induction for comparison and guide. The expected size for protein A is ~33.99 kDa and for protein B is ~40.58 kDa.

This showed a positive result for both targets, elution 2 was quantified using Qubit® protein assay. Target Gasu_17800 (Figure 5.9A) showed protein in both elution 1 (50 mM imidazole) and elution 2 (150 mM imidazole), the concentration of elution 2 sample was 130 µg/mL. Gasu_31410 (Figure 5.9B) showed protein in both elution 2 (150 mM imidazole) and elution 3 (300 mM imidazole), the concentration of elution 2 sample was 120 µg/mL. Although the eluted proteins were in low quantity there was enough protein to test the folded state of the samples and thus, if the refolding method was successful.

During testing of preparation and solubilisation of the IB there was a small amount of target Gasu_17800 that was soluble when cells were sonicated in buffer 1 (Table 5.8). This was then used to perform a purification under native conditions. For each step in the process an aliquot was taken and analysed by SDS-PAGE (Figure 5.10). From the sample of soluble material loaded onto the column, to the unbound material and wash step the target protein was bound to the column. There was elution of the target protein in all three elution steps, but the greatest quantity was in elution 2. Quantification of elution two determined a concentration of 182 µg/mL. This sample can be used for further analysis to look for the proteins folded state, though it should be noted there is some other protein in the elution 2 sample at ~20 kDa.

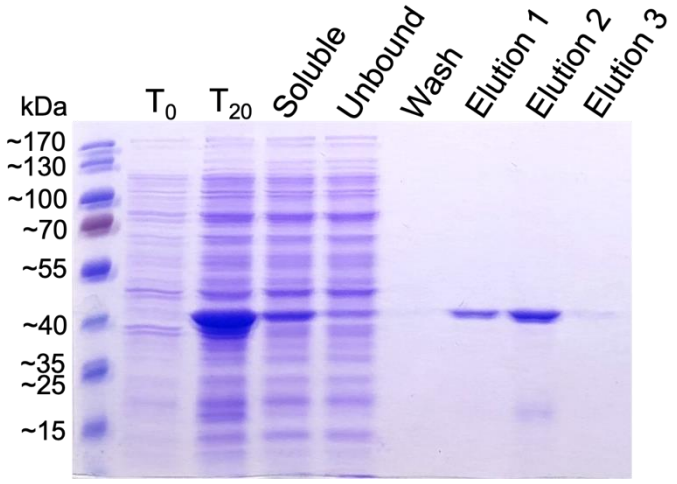


Figure 5.10: SDS-PAGE showing steps for purification of target protein Gasu_17800 under native conditions. Ladder is PageRuler Plus Pre-stained Protein Ladder from Thermo Scientific. Elution steps 1, 2 and 3 were carried out with 50, 150 and 300 mM imidazole retrospectively. T_0 is the cell samples pre induction and T_{20} are 20 hours post induction for comparison and guide. The expected size for the protein is ~33.99 kDa.

5.3.3.4 Protein melting temperature determination

The folded/unfolded nature of the target proteins were tested using NanoDSF (NanoTemper Technologies). Tyrosine fluorescence was used to monitor protein unfolding. Measuring the ratio at 350 nm and 330 nm of the fluorescence intensities allows for the detection of any changes in protein structure (for example due to protein unfolding). Figure 5.11 shows the ratio fluorescence at these wavelengths along with the first derivative against temperature for each of the samples. If a protein unfolds during this process typically you would expect to see a sigmoid shaped curve when looking and the ratios of fluorescence and a corresponding peak in the first derivative trace. Gasu_17800 in both native and denatured conditions (Figure 5.11A and Figure 5.11C) and Gasu_31410 (Figure 5.11B) in denatured conditions shown none of the features consistent with a protein unfolding.

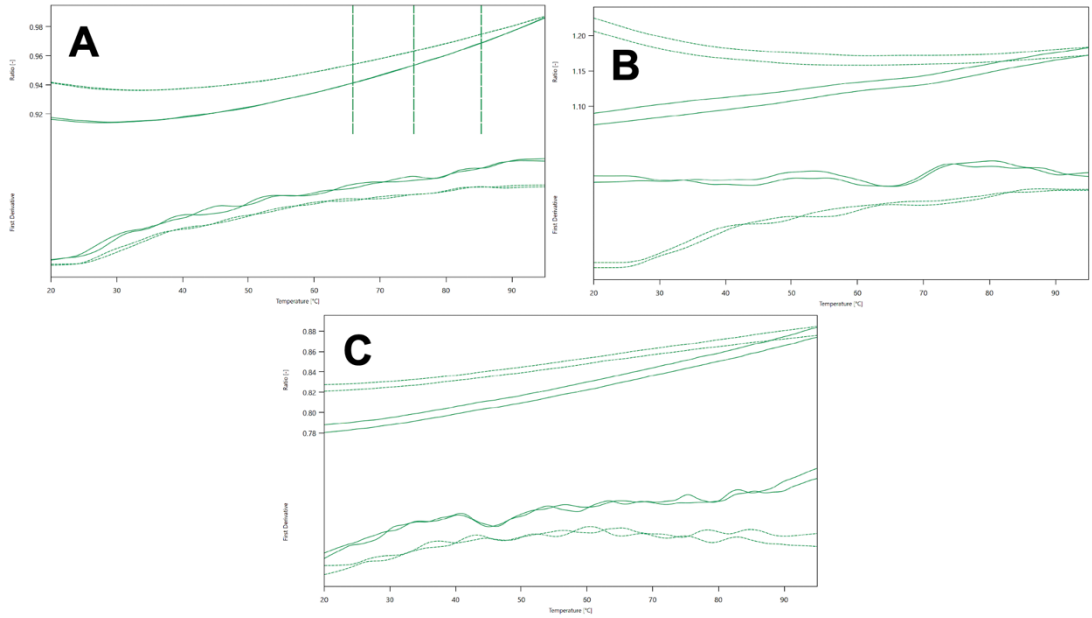


Figure 5.11: Ratio of fluorescence intensities at 350 nm vs 330 nm each line represents a sample and below are the equivalent first derivatives for a label-free nanoDSF experiment with target proteins Gasu_17800 purified under denaturing conditions (A), Gasu_31410 purified under denaturing conditions (B) and Gasu_17800 purified under native conditions (C).

Another option for producing correctly folded proteins is to elute the proteins under denaturing conditions and try to refold them after the purification. As before aliquots were taken from each step in the purification under denaturing conditions, these were analysed via SDS-PAGE shown in Figure 5.12. For both Gasu_17800 (Figure 5.12A) and Gasu_31410 (Figure 5.12B) elution steps 1 and 2 were pooled together and quantified using Qubit ® protein assay resulting in 1420 and 1762 µg/mL retrospectively. These samples were then stored at 4 °C ready for refolding test.

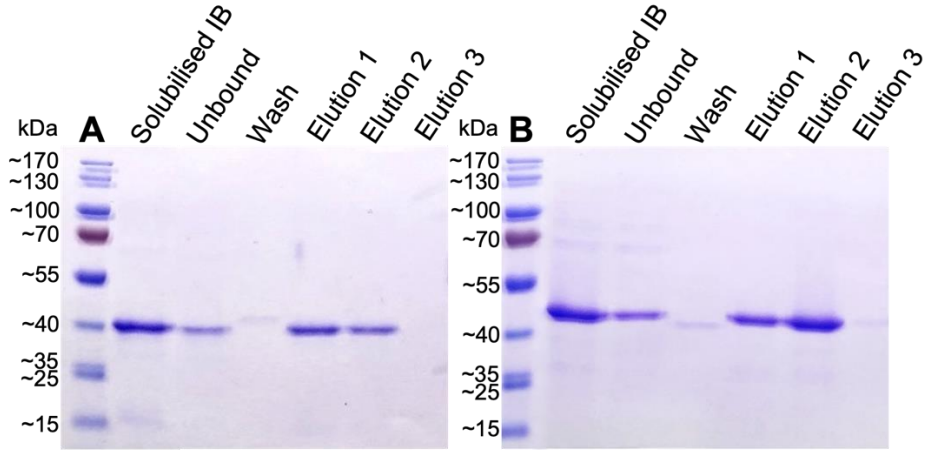


Figure 5.12: SDS-PAGE showing steps for purification of target proteins Gasu_17800 (A) and Gasu_31410 (B) in denatured conditions. Ladder is PageRuler Plus Prestained Protein Ladder from Thermo Scientific. Elution steps 1, 2 and 3 were carried out with 50, 150 and 300 mM imidazole retrospectively (with 7 M urea). The expected size for protein A is ~33.99 kDa and for protein B is ~40.58 kDa.

5.3.4 Refolding assay

In order to find a suitable buffer for refolding the target proteins an assay was designed to test 96 different conditions through shock dilution (Figure 5.3). Here the turbidity of the solution was measured after each protein was added to a buffer. This was measured against controls to then determine suitable buffers to further examine using nanoDS. Figure 5.13 shows the results of this assay, where 13 buffers were suitable for further testing with target Gasu_17800. For Gasu_31410, 38 buffers were compatible for further testing. Once again, the folded/unfolded nature of the target proteins were tested using NanoDSF (NanoTemper Technologies). For Gasu_31410, only one buffer gave a spectrum that was consistent with a protein unfolding this was buffer GHC (Glycine – HCl) at pH 2. This sample was then repeated with fresh protein to check for reproducibility. The spectra for the ratio of wavelengths and the first derivative for the five repeats are shown in Figure 5.14. The sigmoid shaped curve is clearly identifiable as is the peak in the derivative trace just after 90°C, this data reveals the protein Gasu_31410 unfolds at 93.2 °C.

Gasu 17800	pH	2	2.5	3	3.5	4	4.5	5	5.5	6	6.5	7	7.5	
Buffer		1	2	3	4	5	6	7	8	9	10	11	12	Arginine
GHC/MIB	A												MIB	-
PCB	B	PCB												
HCPC/PHP/MMT	C									MMT	MMT			
***	D													
GHC/MIB	E			GHC			MIB		MIB					+
PCB	F										PCB			
HCPC/PHP/MMT	G						MMT			MMT				
***	H				Citric Acid		Citric Acid	Citric Acid						

Gasu 31410	pH	2	2.5	3	3.5	4	4.5	5	5.5	6	6.5	7	7.5	
Buffer		1	2	3	4	5	6	7	8	9	10	11	12	Arginine
GHC/MIB	A	GHC		GHC										-
PCB	B	PCB		PCB										
HCPC/PHP/MMT	C													
***	D			PBS										
GHC/MIB	E	GHC	GHC	GHC	GHC	MIB	MIB	MIB	MIB	MIB	MIB	MIB		+
PCB	F			PCB							PCB			
HCPC/PHP/MMT	G	HCPC			PHP	MMT	MMT	MMT	MMT		MMT	MMT	MMT	
***	H	PBS	PBS		Citric Acid	Citric Acid	Citric Acid	Citric Acid	MES	MES	MES	Tris	Tris	

Figure 5.13: The selected buffers to undergo further testing for refolding of target Gasu_17800 (top) and Gasu_31410 (bottom). Buffer concentration: 50 mM, salt concentration 100 mM and Arginine concentration 0.4 M. GHC: Glycine, MIB: sodium malonate, imidazole and boric acid, PCB: Phosphate Citrate, HCPC: Potassium Chloride, PHP: Potassium Hydrogen Phthalate, MMT: DL-malic acid, MES and Tris-HCl, ***: pH 2-3 (PBS), pH 3.5–5 (Citric acid), pH 5.5–6.5 (MES), pH 7–7.5 (Tris-HCl).

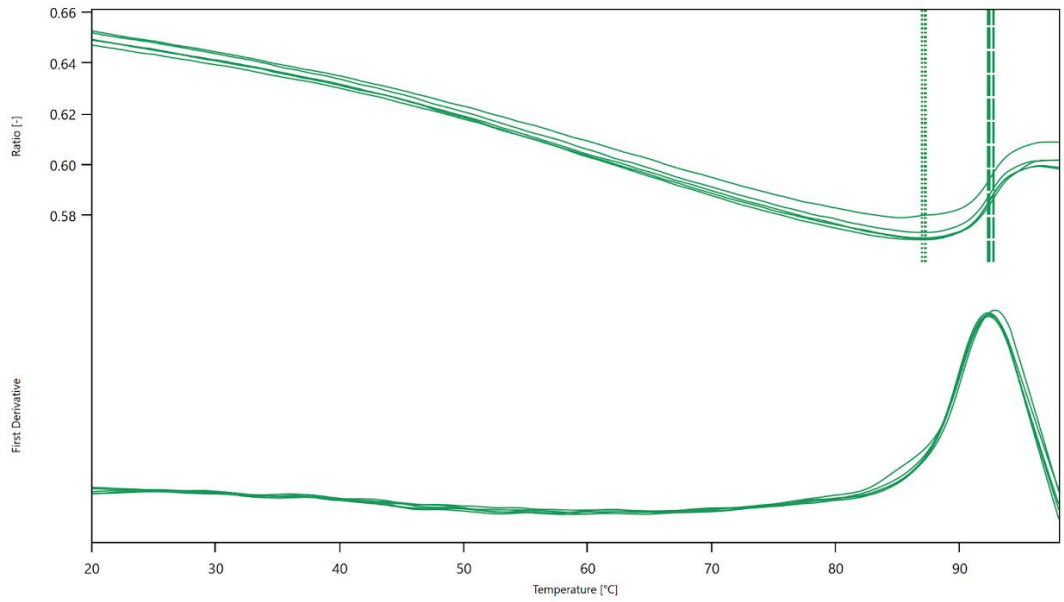


Figure 5.14: Ratio of fluorescence intensities at 350 nm vs 330 nm each line represents a repeated sample and below are the corresponding first derivatives for a label-free nanoDSF experiment with target protein Gasu_31410, purified under denaturing conditions then refolded using shock dilution with 50 mM Glycine – HCl pH 2 buffer. Vertical lines signify the identification of a change in the fluorescence ratios that is compatible with a change in protein state (denaturation).

5.4 Discussion

For the scope of this thesis, it was not possible to cover every protein from the list of putative enzymes (Table 5.9). Therefore, three were chosen to move forward with, Gasu_17800, Gasu_31410 and Gasu_27500. Of the four genes present in both CAZyme analysis (Chapter 2) and secretome analysis (Chapter 3) the predicted peroxidase Gasu_17800 was the smallest protein and therefore was predicted to be easier to clone and recombinantly express over longer proteins. As discussed previously peroxidase involvement with lignin degradation and cell wall modification, made this an interesting target to investigate and for these reasons it was selected. Gasu_27500 was also present in both types of analysis and has predicted beta-galactosidase activity, which has tremendous potential in an industrial setting making it a strong candidate. As previously discussed, (Section 4.4.5.3), beta-galactosidase can be used for the removal of lactose from dairy products and microbial native beta-galactosidases are popular due to their thermostability, thermoacidophilic properties (Asraf and Gunasekaran, 2010). There are also strong links to beta-galactosidases role in lignocellulose saccharification through hydrolysing o-glycosyl bonds in hemicellulose (Blumer-Schuette et al., 2014;

Zavrel et al., 2018; Fernández-Bayo et al., 2019) for these reasons this enzyme was selected. Lastly chosen was uncharacterised Gasu_31410, it was only present in the secretome analysis (Table 5.9) but showed relatively high unique peptide hits from the mass spectrometry analysis (Section 0). Furthermore, it showed promising ability to be acid tolerant and function under high temperature given the sequence matches even though these homology searches revealed no predicted function. Additionally given the possibility this gene was horizontally acquired from bacteria it was hypothesised that cloning into *E. coli* could be unproblematic.

For cloning and heterologous expression, the CDS sequence of each of the target was retrieved making sure to remove the predicted signal peptides, as it is likely these would be cleaved off during native protein production (Armenteros et al., 2019). The cloning work described in the chapter was largely successful, initially the amplification of the three targets was difficult and a series of optimisation PCR was required to find the correct conditions to amplify correctly (data not shown), unfortunately this was unsuccessful for target Gasu_27500 (Figure 5.4). Interestingly target Gasu_17800 amplification showed two bands, this is likely to be off target amplification of one of the other *Galdieria* peroxidases as they share similar sequences (Oesterhelt et al., 2008). The linearisation of the pET vectors was simple and successful, as was the ligation cloning of targets into the plasmids (Figure 5.5). For each gene and in each vector, there was a successful transformant that was confirmed through sequencing.

5.4.1 Recombinant protein production

Initially in expression trails there was no discernible difference in observed expression levels, so the decision to continue with the modified pET28a-TIR-2+T7pCONS expression vector was heavily based on published results that showed a greater than two-fold increase in protein production (Shilling et al., 2020). The expression strain Rosetta DE3 is a BL21 derivative purposely designed to improve expression of eukaryotic proteins that use rare codons. As Figure 5.6 shows, when induced at 18 °C there were good amounts of expressed protein from both genes. However, it proved very difficult for both genes to express soluble proteins (Figure 5.7). Also, the results highlighted an incongruence between the predicted molecular weight of the target proteins and what was observed. The difference in size could be due to the estimate being incorrect or the protein is in some folded form and thus was not fully denatured prior to electrophoresis. The native environment of these proteins is very acidic and can have temperatures up to 56 °C. Therefore, it was hypothesised that replicating these environments may help in increasing the solubility of the proteins. As Figure 5.7

highlights, both lower pH and incubation at 60 °C produced less soluble material. It is not uncommon for high levels of expressed recombinant proteins in *E. coli* to result in highly specific aggregation of the expressed protein into inclusion bodies (IBs). Due to the impractical and time-consuming nature of testing multiple buffers and expression conditions, it was decided to use the formation of IBs as an advantage. It should be noted the formation of inclusion bodies is not uncommon when expression eukaryotic genes in a prokaryotic and is influenced by multiple factors (Singh et al., 2015). The formation of inclusion bodies meant an easy isolation process, as they are mechanically stabile and protected the target proteins from proteolytic degradation (Singhvi et al., 2020). The isolation of inclusion bodies allowed for the removal of multiple *E. coli* genes, acting as a purification step in itself. After isolation of inclusion bodies, the next step in purification requires solubilisation of the aggregates. Typically, inclusion bodies are solubilised using strong denaturants and chaotropes in high concentrations, most commonly urea and guanidine hydrochloride are used (Singh et al., 2015). Figure 5.8 visualises this process where buffers containing increasing amounts of urea were used, here it is highlighted that the solubilisation of the inclusion bodies with higher concentrations of urea contain less background *E. coli* proteins compared to lane T_{20}.

5.4.2 Purification

As the purification was taking place under denatured conditions a strategy for refolding the proteins was required. Initially, this consisted of sequentially washing the protein that remained bound to the column with buffers containing decreasing amounts of urea. Ideally, this would allow the protein to fold slowly into its native condition while remaining soluble. Known as matrix-assisted protein refolding, this technique is useful as aggregation prone folding intermediates are partially suppressed during refolding so avoiding unwanted intermolecular interaction. As the purification and refolding steps are combined it is also convenient as the exchange of buffer conditions and the subsequent removal of the refolded target protein is relatively easy (Vallejo and Rinas, 2004; Singh et al., 2015). For both genes elution of the target protein was possible and enough protein to test the folded state. Both samples also showed from the unbound fraction and wash step that the column was likely at capacity, thus explain the low yield of eluted protein. During the preparations of inclusion bodies, it was noticed the buffer used cause some of Gasu_17800 to become soluble during cell lysis after expression. As such a purification was carried out, in this case as proteins were never denatured there was no need for any refolding steps and any eluted protein should be in a native form (Figure 5.10).

5.4.3 Refolding

Nanoscale differential scanning fluorimetry (nanoDSF) is a valuable tool for thermal unfolding assays and determining melting points of proteins (Magnusson et al., 2019). In all three cases the NanoDSF showed no indication of any folded protein material. It is possible that the temperature was not high enough to measure the denaturing of proteins but is most likely that the protein is in an aggregated form. The conditions at which a protein can be refolded can be very specific. Again, as discussed earlier, the native pH of these proteins is around pH 2. Therefore, it is feasible that the pH of the elution buffer (~ pH 7) is too basic and causing the proteins to precipitate and aggregate even under native purification. Using a high throughput systematic screening method to evaluate multiple refolding conditions in parallel is likely to be necessary.

The set-up of the refolding assay contained a range of buffering systems with and without 0.4 M L-arginine. Protein aggregation during the refolding process has been shown to be mitigated in the presence of L-arginine and is a useful additive to test in folding buffers (Bajorunaite et al., 2007; Chen et al., 2009; Soleymani et al., 2020). The screen covered different buffering systems that not only allowed the exploration into the effect of pH on the protein during refolding, but equally the influence that different compositions of buffers would have.

It is well known in regard to protein solubility that pH has a significant effect and based on the results in Figure 5.13 for Gasu_17800 and Gasu_31410 it is unclear on what an optimal pH could be. Though for Gasu_31410 there are more buffers with a pH of < 3 compared to Gasu_17800, thus it is concluded that the role of pH on protein refolding is specific to individual proteins. There are other variables to consider, aside from pH the different buffering systems themselves have an effect. This was demonstrated by the results in Figure 5.14, as Gasu_31410 could only be successfully refolded using pH 2 buffer GHC, other buffers of pH 2 did not show refolded protein (data not shown). These results are a strong indication that the composition of reagents making the buffering system influence the refolding process. With Gasu_31410 in a folded state at pH 2 and stable up to ~92 °C it is a perfect candidate for further testing on a range of different substrates for activity characterisation as is compatible with a thermostable protein. Analysis such as circular dichroism (CD) spectroscopy is widely used technique for analysing a proteins secondary structure (Berova et al., 2000).

5.4.4 Conclusion

During this study, all the genes tested have proven difficult either to amplify, solubly express and/or refold. Although the time constraints on this project did not allow testing of more enzymes, the original list of genes (Table 5.9) has interesting candidates worth further investigating for their industrially relevant lignocellulosic degrading properties. Due to the issues encountered while trying to express soluble protein it would be worth investigating different conditions to optimise *soluble* expression, such as induction OD, addition of solubility tag, additives, media and *E. coli* strains. Even exploring using a different expression host may yield positive as literature has previously showed some *G. sulphuraria* phosphate translocators have previously been cloned into yeast (Linka et al., 2008). However, without infinite time it may be more suitable to focus efforts in identifying suitable refolding buffers using the plate format described here. Different buffers and concentrations can be screened and even the addition of different types of protein stabilisers assessed relatively quickly.

In conclusion, it is clear that some of the enzymes of interest identified in *G. sulphuraria* are a good source for acid tolerant thermostable proteins for use in industry. These results show that *G. sulphuraria* and its environment could provide valuable possibilities for the discovery of new lignocellulosic degrading enzymes. These enzymes could be used in industrial processes such as biofuel production such as being used to create an enzymatic saccharification cocktail to act as part of a pre-treatment process. Although the experiments of heterologous expression and purification have been performed for successfully for target Gasu_31410, the question of the enzymes specific function and that of the other targets (Table 5.9) remains unanswered. Unfortunately, due to unforeseeable circumstances and time constraints those experiments were not able to take place within this project.

Chapter 6 - Discussion

6.1 Summary

My PhD aimed to explore the industrially relevant novel enzymes of the extremophile *G. sulphuraria*, focusing primarily on applications relevant to the degradation of lignocellulosic material for production of biofuels. Across all four data chapters, I have documented and demonstrated novel progress in the study of this organism. In Chapter 2, I carried out the largest sequencing and most extensive nuclear phylogenetic analysis performed in this field. Using this information, I was able to resolve the phylogenetic relationship among this extremely diverse species. This analysis clearly identified the divergence of the species into six lineages that have all been evolving separately. Additionally, I was then able to assess the rate of evolution on nuclear genes, this gave rise to numerous genes under positive selection that are likely essential *in G. sulphuraria*'s survival in extreme environments. This highlighted the importance to further examine each of the core six lineages for novel industrially relevant enzymes involved in the degradation of lignocellulosic material. I created transcriptomic data and long read DNA sequencing to facilitate uncovering the CAZyme repertoire of the species. The CAZyme analysis in Chapter 3 aimed to advance the knowledge of the collection of CAZymes within the *G. sulphuraria* genomes and aid in understanding its versatile metabolic abilities. I identified 14 putative secreted enzymes involved in the degradation of lignocellulosic material which were suspected to be both heat and acid tolerant. However, given the capability of *G. sulphuraria* to grown on multiple carbon sources this was a surprisingly low number and arose more questions than it answered. In Chapter 4 I investigated further through numerous growth studies and analysis of proteins present in the supernatant via LC-MS. Subsequently 11 genes were identified as suitable for further study among these targets were two alpha-glucosidase, two beta-galactosidase, a peroxidase, a purple acid phosphatase and five uncharacterised proteins. The uncharacterised proteins hold huge potential in the discovery of unseen and unique enzymes involved in carbohydrate degradation. Finally, Chapter 5 was intended to recombinantly express the genes identified in both Chapters 3 and 4 in order to understand these enzymes and their function regarding the degradation of lignocellulosic material. However, due to time constraints and difficulties encountered during the process only one target was successfully expressed, purified and refolded, this did show a high denaturation temperature at pH 2 which is supportive of enzyme function under high temperature and low pH. The culmination of my work has led to a greater understanding of G. sulphuraria and discovery of multiple novel enzymes that provide the essential basis for further exploration.

6.2 Adaptation and evolution of *Galdieria* to extreme conditions

To overcome the extreme conditions in *G. sulphuraria* environment the species has adapted for survival. They have evolved to survive extremely low pH where typically the majority of organism's protein denaturation would occur making survival unachievable. In order for a microorganism to keep metabolic activities active, the cytoplasmic pH of a cell is required to be near neutral ($5 \leq pH \leq 8.5$) (Oarga, 2009; Krulwich et al., 2011; Jin and Kirk, 2018; Merino et al., 2019). In *G. sulphuraria* the internal pH is neutral, the low outer pH (~pH 2) creates an inward acting proton gradient of approximately $1{:}1 \times 10^5$ (Enami et al., 1986). More research is required into the specifics of this mechanism, but it is believed that the alga uses a mechanism to temporarily make their plasma membrane impermeable, coupled with the use of passive proton pumps (Beardall and Entwisle, 1984). *G.* sulphuraria's ridged cell wall is resistant to the considerable osmotic stress caused by the high proton gradient. As a benefit, this rigid and resistant cell wall increases *G. sulphuraria*'s ability to withstand high salt environments. Another factor contributing to this is the acquisition of genes horizontally from archaea and prokaryotes. These have aided the alga's ability to withstand salt stress of up to 1.5 M NaCl (Rossoni et al., 2019). In high acidity environments, solubility of many heavy, precious and rare earth metals and minerals is increased. For example, in the environments where *G. sulphuraria* is found concentrations of arsenic and toxic heavy metals such as mercury can reach levels that at lethal for most organisms (Doemel and Brock 1971; Schönknecht et al., 2013; Rossoni et al., 2019). The bioaccumulation of these substances has been explored in *G. sulphuraria,* with the assumption that the algae has adapted and is therefore not inhibited by such levels. This has been explored multiple times for use in biotechnological remediation of metals via waste waters (Misumi et al. 2008; Osaki et al. 2009; Jalali et al 2018; Čížková et al. 2019; Cho et al. 2020).

These adaptations for survival are not solely attributed to horizontal gene transfer. *G. sulphuraria* have evolved and mutated and natural selection has favoured the more useful mutations. When a cell is replicating, mutations can occur in any region of the genome of an individual organism and there are several consequences. A mutation may be deleterious and so selected against until alleles of the mutated type are no longer present, mutations may have no effect and may persist in a population through genetic drift, or a mutation may concur a selective disadvantage and spread through the population until it becomes ubiquitous (Yang and Nelson, 2000; Del Amparo, 2019). Mutations that occur within the coding region of a gene may change the amino acid

sequence of a protein, in turn affecting the protein function. If changes occur elsewhere in the genome, the existing enhancer/promoter regions may be destabilised or new regions may be created entirely (Yang, 2000).

This thesis has shown the clear divergence of *G. sulphuraria* into six lineages, each of which has been evolving separately. The measuring of the synonymous to non-synonymous rates highlighted nuclear genes that are under adaptive evolution and could be key in the survival and adaptation of this alga to its extreme environment. Further to this I showed each linage contained sets of genes that are unique and potentially a mine for interesting enzymes.

6.3 Growth capacity of *G. sulphuraria*

G. sulphuraria's ability to grow heterotrophically on numerous carbon sources is a point of interest particularly related to the premise of my thesis. Genome analysis by Barbier et al. in 2005 revealed that *G. sulphuraria* encodes a large number of putative monosaccharide transporters. In their research 28 distinct sugar transported genes were identified that were not present in the sister species *C. merolae*. As *G. sulphuraria* can utilise polysaccharides and other polymers, it is suggested that there may be some extracellular enzymes at play. In order for these secreted enzymes to be stable in the environment *G. sulphuraria* are found, the enzymes must be both heat and acid resistant. Enzymes that are stable in very hot, acidic conditions are extremely beneficial for industrial purposes.

The search for the hypothesised extracellular carbohydrate acting enzymes has proven interesting. Previous research has revealed some examples such as a secreted glucoamylase with activity at pH 2 and 80 °C and a class III peroxidase (pH 1.8-2, 60-80 °C) (Shrestha and Weber, 2007; Oesterhelt et al., 2008). Similarly, this thesis in Chapter 5 showed a recombinantly expressed protein was refolded at pH 2 then showed denaturation at ~90 °C (though I was unable to test activity). There have been few studies done to investigate the likely noteworthy enzymes present in this alga. The abundance of sequence data created through this thesis allowed me to perform an extensive study into the CAZymes present in *G. sulphuraria* (Chapter 3) which were expected to be a rich source of exciting proteins especially in relation to degradation of lignocellulosic material. This yielded interesting results but were not as predicted, there were very few predicted secreted hydrolases given *G. sulphuraria* documented heterotrophic capabilities this was unusual. This lack of result is evidence supporting the

idea that some genes harboured by *G. sulphuraria* are likely to be multifunctional and/or be completely unique and therefore unseen sequence.

Phylogenetic analysis has supported the hypothesis that many genes contributing to the metabolic versatility observed by the alga (along with other environmental adaptations) are horizontally acquired. Schönknecht et al in 2013 highlighted that at least 5 % of the protein coding genes in *G. sulphuraria* were likely to be horizontally acquired (Schönknecht et al., 2013). This was also supported by work in this thesis from Chapter 4 where often homology sequencing searches from the secretome were similar to various bacteria and archaea. Having such high homology with prokaryotes is somewhat unexpected for a eukaryotic organism but just highlights the interesting evolution presented by this extremophilic microalga. Chapter 4 also highlighted once again the diversity across the species, there was clear evidence for different proteins present in the secretome for each of the six lineages on three different substrates.

6.4 Importance of studying non-model species

A model organism is defined as a non-human species that can be easily studied to examine biological theory with hopes that the collected data is applicable to a wider range of organisms (Leonelli and Ankeny 2013). Typically, model organisms contain attributes that make them easy to study and have features that can represent a wider group. For instance, they are typically relatively small in size, inexpensive and easy to culture under laboratory conditions. Additionally, they will ideally have a large number of offspring and a short reproductive cycle meaning a high volume of individuals can be accommodated in a single facility over a short period of time (Leonelli and Ankeny 2013; Mathews and Vosshall, 2020). The combination of these characteristics makes these species accessible and suitable as genetic tools for broad biological study.

G. sulphuraria is small in size (3-11 µm; Sentsova, 1991) and can be cultured to a high density with dry cell weight reaching up to 120 g/L under heterotopic growth (Schmidt et al., 2005). As comparison the model organism *Chlamydomonas reinhardtii* a green alga, has recently been shown to during overexpression of a recombinant lysine decarboylases grow up to 20 g/L dry weight (Freudenberg et al., 2021). *G. sulphuraria* can release from 4-32 endospores at a time during cell multiplication (Sentsova, 1991; Pinto et al., 2003). Cell doubling time is very dependent on the *G. sulphuraria* strain but has been reported from anywhere between 16 – 95 hours (Graziani et al., 2013). Although this is not comparable to other model organism such as *E. coli*, they are still able to produce a high density of cells in a relatively short time when compared to other

model organisms such as mice or plants. The versatile nature of the algae means the system requirements are adaptable. They can grow as an aerobic autotroph in the light and as a heterotroph in the dark, with growth supported essentially from any carbon source; they will also grow anaerobically in 100% N_2 in the dark if supplied a carbon source and in 100% CO_2 in the light. The required pH of the media can present some difficulty and risk during cultivation. However, this is also a benefit in some ways, as this pH is toxic to most microorganisms, they can be cultured in non-sterile conditions with little consequence (Scherhag and Ackermann, 2021). The heterotrophic growth capacity also means that *G. sulphuraria* can be grown on various waste streams containing a carbon source (Henkanatte-Gedera et al., 2017; Ende and Noke, 2019; Pleissner et al., 2021).

Increased affordability and ease of genome sequencing has increased the number of classified model organisms has grown rapidly (Hedges 2002). Having access to an organism genome sequence significantly increases the amount and quality of research that can be carried out. The depth of genome information available for *G. sulphuraria* has been extensively extended during this thesis, with six fully annotated *G. sulphuraria* genomes each representing one of the clear lineages as shown in Chapter 2 now available.

However, there are issues with working on this organism, as detailed in this thesis the diversity across the lineages in both phylogenetic analysis as well as the CAZyme and secretome analysis show vast differences dependent on the strain. In reality there is strong evidence to suggest that each lineage may even constitute a separate species altogether. This means that any given strain may be too variable to be applicable to others. Additionally, G. sulphuraria can be extremely difficult to work with in laboratory settings, in order to attain the data presented in this thesis many molecular techniques needed to be adapted or established specifically for this extremophile. Equally many standard protocols such as enzymes assays are difficult to adapt to work at pH 2. *G. sulphuraria* is an extremely interesting organism but for the reasons detailed above does not suit research as a model organism. These algae are too unique and complex to serve as an easy model for the study of general biological phenomena.

To investigate *G. sulphuraria* enzymes and their function, purified folded proteins is required. Chapter 5 of this thesis detailed an attempt as this using heterologous expression in *E. coli*. Issues arose around protein solubility, which is not unusual for expression of eukaryotic gene. In many cases trying to create soluble recombinant protein is futile. A solution to this problem could be developing *G. sulphuraria* itself as an

expression host. The secretome results shown in this thesis support the hypothesis of secreted extracellular enzymes. Creating overexpression of genes of interest with signal peptide sequences could mean the production of large amount of secreted protein in a native folded state. The advantage of this is the purification steps would be simplified as no cell lysis would be required just filtering the correct proteins from the supernatant. In order to achieve this transformation protocols, need to be established in the organism.

Due to their ease of study, biological research has focused on studying model organisms leading to the creation of well documented databases and protocols insert (Leonelli and Ankeny 2013). Model organisms have therefore been invaluable in enhancing scientific principles, such as evolution theory. However, the choice to study model organisms does in some cases restricted our knowledge as the data collected is from the examination of few species and information learned is not always widely applicable. Using non-model organisms such as extremophiles like *G. sulphuraria* allow for the study of naturally occurring biological niches, which enables broader comparisons and greater understanding of core biological functions and adaptive traits (Crawford 2001). Additionally, it provides the opportunity to solve real world problems using features identified in extreme nature conditions.

6.5 Overall conclusions

Presented with environmental challenges such as global warming it is undeniable that there is a need for sustainable replacements for fossil fuels. Lignocellulose biomass is a potentially rich source of fermentable sugars and the abundance of this feedstock make it an attractive source for the production of biofuels. The conversion of lignocellulosic material into useful sugars is difficult due to its recalcitrant nature. It is important to look for examples in nature that could offer solutions. The main objective of my work was to try to find novel acid resistant and heat tolerant enzymes that would be able to act on lignocellulosic material. Any enzymes identified could benefit yields in biofuel production by aiding the saccharification of the lignocellulosic matrix. As they would be heat and acid tolerant, they could potentially be used during some of the harsh pre-treatments steps already in place. It is clear that *G. sulphuraria* is unique organism that given its growth capacity heterotrophically on number carbon source would possess interesting enzymes. However, findings did not reveal any such astonishing enzymes with predicted ligninase, cellulase or xylanase function, just highlighted the diversity across the species and confirmed the presence of secreted enzymes under heterotrophic growth of lignocellulosic material. Many of the enzymes discovered were uncharacterised and of

unknown function. Attempts to purify recombinant proteins proved difficult and the question of the functions of these target proteins remains unanswered. This highlights the importance of further research to better understand this extraordinary organism and the enzymes it possesses. The resolved phylogeny, sequence data, CAZyme and secretome analysis presented in this thesis mean that future study of this species, is now much more feasible.

References

Abdel-Hamid, A.M., Solbiati, J.O. and Cann, I.K., 2013. Insights into lignin degradation and its potential industrial applications. *Advances in applied microbiology*, *82*, pp.1-28.

Ahmed, S., Riaz, S. and Jamil, A., 2009. Molecular cloning of fungal xylanases: an overview. *Applied microbiology and biotechnology*, *84*(1), pp.19-35.

Akoh, C.C., Lee, G.C., Liaw, Y.C., Huang, T.H. and Shaw, J.F., 2004. GDSL family of serine esterases/lipases. *Progress in lipid research*, *43*(6), pp.534-552.

Akhtar, N., Gupta, K., Goyal, D., and Goyal, A. (2016). Recent advances in pretreatment technologies for efficient hydrolysis of lignocellulosic biomass. *Environ. Prog. Sustain Energy*. 35, 489–511. doi: 10.1002/ep.12257

Alalwan, H.A., Alminshid, A.H. and Aljaafari, H.A., 2019. Promising evolution of biofuel generations. Subject review. *Renewable Energy Focus*, *28*, pp.127-139.

Albertano, P., Ciniglia, C., Pinto, G. and Pollio, A., 2000. The taxonomic position of Cyanidium, Cyanidioschyzon and Galdieria: an update. *Hydrobiologia*, *433*(1), pp.137-143.

Allen, M. and Stanier, R. (1968). Selective Isolation of Blue-green Algae from Water and Soil. *Journal of General Microbiology*, 51(2), pp.203-209.

Allen, M.B., 1959. Studies with Cyanidium caldarium, an anomalously pigmented chlorophyte. *Archiv für Mikrobiologie*, *32*(3), pp.270-277.

Amiri, H. and Karimi, K., 2018. Pretreatment and hydrolysis of lignocellulosic wastes for butanol production: challenges and perspectives. *Bioresource technology*, *270*, pp.702-721.

Andlar, M., Rezić, T., Marđetko, N., Kracher, D., Ludwig, R. and Šantek, B., 2018. Lignocellulose degradation: an overview of fungi and fungal enzymes involved in lignocellulose degradation. *Engineering in Life Sciences*, *18*(11), pp.768-778.

Andolfatto, P., 2005. Adaptive evolution of non-coding DNA in Drosophila. *Nature*, *437*(7062), pp.1149-1152.

Arad, S., 1988. Production of sulfated polysaccharides from red unicellular algae. *Algal biotechnology/edited by T. Stadler...[et al.]*.

Arad, S.M. and Levy-Ontman, O., 2010. Red microalgal cell-wall polysaccharides: biotechnological aspects. *Current opinion in biotechnology*, *21*(3), pp.358-364.

Arantes, V. and Saddler, J.N., 2010. Access to cellulose limits the efficiency of enzymatic hydrolysis: the role of amorphogenesis. *Biotechnology for biofuels*, *3*(1), pp.1-11.

Armenta, S., Moreno-Mendieta, S., Sánchez-Cuapio, Z., Sánchez, S. and Rodríguez-Sanoja, R., 2017. Advances in molecular engineering of carbohydrate-binding modules. *Proteins: Structure, Function, and Bioinformatics*, *85*(9), pp.1602-1617.

Armenteros, J.J.A., Tsirigos, K.D., Sønderby, C.K., Petersen, T.N., Winther, O., Brunak, S., von Heijne, G. and Nielsen, H., 2019. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nature biotechnology*, *37*(4), pp.420-423

Armendáriz-Ruiz, M., Rodríguez-González, J.A., Camacho-Ruíz, R.M. and Mateos-Díaz, J.C., 2018. Carbohydrate esterases: An overview. *Lipases and Phospholipases*, pp.39-68.

Armenteros, J.J.A., Tsirigos, K.D., Sønderby, C.K., Petersen, T.N., Winther, O., Brunak, S., von Heijne, G. and Nielsen, H., 2019. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nature biotechnology*, *37*(4), pp.420-423.

Aspeborg, H., Schrader, J., Coutinho, P.M., Stam, M., Kallas, A., Djerbi, S., Nilsson, P., Denman, S., Amini, B., Sterky, F. and Master, E., 2005. Carbohydrate-active enzymes involved in the secondary cell wall biogenesis in hybrid aspen. *Plant Physiology*, *137*(3), pp.983-997.

Asraf, S.S. and Gunasekaran, P., 2010. Current trends of ß-galactosidase research and application. *Current research, technology and education topics in applied microbiology and microbial biotechnology. Microbiology book series Formatex Research Center, Spain*, pp.880-890.

Bai FW, Shihui Yang, Nancy W.Y. Ho., 2019. 3.05 - Fuel Ethanol Production From Lignocellulosic Biomass. *Comprehensive Biotechnology (Third Edition), pp.49-65*

Bajorunaite, E., Sereikaite, J. and Bumelis, V.A., 2007. L-arginine suppresses aggregation of recombinant growth hormones in refolding process from E. coli inclusion bodies. *The protein journal*, *26*(8), pp.547-555.

Bajpai, P., 2016. Structure of lignocellulosic biomass. In *Pretreatment of lignocellulosic biomass for biofuel production* (pp. 7-12). Springer, Singapore.

Bajwa, D.S., Pourhashem, G., Ullah, A.H. and Bajwa, S.G., 2019. A concise review of current lignin production, applications, products and their environmental impact. *Industrial Crops and Products*, *139*, p.111526.

Bala, A. and Singh, B., 2019. Cellulolytic and xylanolytic enzymes of thermophiles for the production of renewable biofuels. *Renewable Energy*, *136*, pp.1231-1244.

Barbier, G., Oesterhelt, C., Larson, M.D., Halgren, R.G., Wilkerson, C., Garavito, R.M.,

Baruah, J., Nath, B.K., Sharma, R., Kumar, S., Deka, R.C., Baruah, D.C. and Kalita, E., 2018. Recent trends in the pretreatment of lignocellulosic biomass for value-added products. *Frontiers in Energy Research*, *6*, p.141.

Bastawde, K.B., 1992. Xylan structure, microbial xylanases, and their mode of action. *World Journal of Microbiology and Biotechnology*, *8*(4), pp.353-368

Beardall, J. and Entwisle, L., 1984. Internal pH of the obligate acidophile Cyanidium caldarium Geitler (Rhodophyta?). *Phycologia*, *23*(3), pp.397-399.

Beaulieu, J.M., Leitch, I.J., Patel, S., Pendharkar, A. and Knight, C.A., 2008. Genome size is a strong predictor of cell size and stomatal density in angiosperms. *New Phytologist*, *179*(4), pp.975-986.

Behera, S., Singh, R., Arora, R., Sharma, N.K., Shukla, M. and Kumar, S., 2015. Scope of algae as third generation biofuels. *Frontiers in bioengineering and biotechnology*, *2*, p.90.

Benning, C. and Weber, A.P., 2005. Comparative genomics of two closely related unicellular thermo-acidophilic red algae, Galdieria sulphuraria and Cyanidioschyzon merolae, reveals the molecular basis of the metabolic flexibility of Galdieria sulphuraria and significant differences in carbohydrate metabolism of both algae. *Plant physiology*, *137*(2), pp.460-474.

Berlemont, R. and Martiny, A.C., 2016. Glycoside hydrolases across environmental microbial communities. *PLoS computational biology*, *12*(12), p.e1005300.

Bernardes, A., Pellegrini, V. O. A., Curtolo, F., Camilo, C. M., Mello, B. L., Johns, M. A., et al., (2019). Carbohydrate binding modules enhance cellulose enzymatic hydrolysis by increasing access of cellulases to the substrate. *Carbohydr. Polym.* 211, 57–68. doi: 10.1016/j.carbpol.2019.01.108

Berova, N., Nakanishi, K. and Woody, R.W. eds., 2000. *Circular dichroism: principles and applications*. John Wiley & Sons.
Bewley JD, Krochko JE. 1982. Desiccation tolerance. In *Encyclopedia of Plant Physiology: Physiological Plant Ecology II: Water Relations and Carbon Assimilation*, ed. OL Lange, PS Nobel, CB Osmond, H Ziegler, pp. 325–78. Berlin: Springer

Bhatia, S.K., Jagtap, S.S., Bedekar, A.A., Bhatia, R.K., Patel, A.K., Pant, D., Banu, J.R., Rao, C.V., Kim, Y.G. and Yang, Y.H., 2020. Recent developments in pretreatment technologies on lignocellulosic biomass: effect of key parameters, technological improvements, and challenges. *Bioresource technology*, *300*, p.122724.

Bhatia, R.K., Ullah, S., Hoque, M.Z., Ahmad, I., Yang, Y.H., Bhatt, A.K. and Bhatia, S.K., 2021. Psychrophiles: A source of cold-adapted enzymes for energy efficient biotechnological industrial processes. *Journal of Environmental Chemical Engineering*, *9*(1), p.104607.

Bhattacharya, D., Price, D.C., Chan, C.X., Qiu, H., Rose, N., Ball, S., Weber, A.P., Cecilia Arias, M., Henrissat, B., Coutinho, P.M. and Krishnan, A., 2013. Genome of the red alga Porphyridium purpureum. *Nature communications*, *4*(1), pp.1-10.

Bharathiraja B, Jayamuthunagai J, Sudharsanaa T, Bharghavi A, Praveenkumar R, Chakravarthy M & Yuvaraj D (2017) Biobutanol - an impending biofuel for future: a review on upstream and downstream processing techniques. *Renew Sust Energ Rev* **68**, 788– 807.

Biely, P. Microbial carbohydrate esterases deacetylating plant polysaccharides. *Biotechnol. Adv.* 2012, *30*, 1575–1588.

Billi, D. and Potts, M., 2002. Life and death of dried prokaryotes. *Research in microbiology*, *153*(1), pp.7-12.

Blommaert, J., 2020. Genome size evolution: Towards new model systems for old questions. *Proceedings of the Royal Society B*, *287*(1933), p.20201441.

Blumer-Schuette, S.E., Brown, S.D., Sander, K.B., Bayer, E.A., Kataeva, I., Zurawski, J.V., Conway, J.M., Adams, M.W. and Kelly, R.M., 2014. Thermophilic lignocellulose deconstruction. *FEMS Microbiology Reviews*, *38*(3), pp.393-448.

Bourne, Y. and Henrissat, B., 2001. Glycoside hydrolases and glycosyltransferases: families and functional modules. *Current opinion in structural biology*, *11*(5), pp.593-600.

Breton, C.; Šnajdrová, L.; Jeanneau, C.; Koča, J.; Imberty, A. Structures and mechanisms of glycosyltransferases. *Glycobiology* 2006, *16*, 29R–37R.

Breton, C., Fournel-Gigleux, S. and Palcic, M.M., 2012. Recent structures, evolution and mechanisms of glycosyltransferases. *Current opinion in structural biology*, *22*(5), pp.540-549.

Brosse, N., Mohamad Ibrahim, M.N. and Abdul Rahim, A., 2011. Biomass to bioethanol: Initiatives of the future for lignin. *International Scholarly Research Notices*, *2011*.

Burns, T., 2020. *High-density heterotrophic cultivation of Galdieria sulphuraria for the production of high-stability phycocyanin* (Doctoral dissertation, University of Sheffield).

Byrne, R. T., Klingele, A. J., Cabot, E. L., Schackwitz, W. S., Martin, J. A., Martin, J., et al., (2014). Evolution of extreme resistance to ionizing radiation via genetic adaptation of DNA repair. *eLife* 3:e01322. doi: 10.7554/eLife.01322

COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS Stepping up Europe's 2030 climate ambition Investing in a climate-neutral future for the benefit of our people

Cantarel, B.L., Coutinho, P.M., Rancurel, C., Bernard, T., Lombard, V. and Henrissat, B., 2009. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. *Nucleic acids research*, *37*(suppl_1), pp.D233-D238.

Chakraborty, S., Rani, A., Dhillon, A. and Goyal, A., 2017. Polysaccharide lyases. In *Current Developments in Biotechnology and Bioengineering* (pp. 527-539). Elsevier.

Chen, R., 2012. Bacterial expression systems for recombinant protein production: E. coli and beyond. *Biotechnology advances*, *30*(5), pp.1102-1107.

Chen, J., Liu, Y., Li, X., Wang, Y., Ding, H., Ma, G. and Su, Z., 2009. Cooperative effects of urea and L-arginine on protein refolding. *Protein expression and purification*, *66*(1), pp.82-90.

Chen, H., Liu, J., Chang, X., Chen, D., Xue, Y., Liu, P., Lin, H. and Han, S., 2017. A review on the pretreatment of lignocellulose for high-value chemicals. *Fuel Processing Technology*, *160*, pp.196-206.

Chettri, D., Verma, A.K., Sarkar, L. and Verma, A.K., 2021. Role of extremophiles and their extremozymes in biorefinery process of lignocellulose degradation. *Extremophiles*, pp.1-17.

Correa, A. and Oppezzo, P., 2015. Overcoming the solubility problem in E. coli: available approaches for recombinant protein production. *Insoluble proteins*, pp.27-44.

Cho, Chung Hyun, Seung In Park, Claudia Ciniglia, Eun Chan Yang, Louis Graf, Debashish Bhattacharya, and Hwan Su Yoon. 2020. “Potential Causes and Consequences of Rapid Mitochondrial Genome Evolution in Thermoacidophilic Galdieria (Rhodophyta).” BMC Evolutionary Biology 20 (112): 1–15. https://doi.org/10.21203/rs.3.rs-36820/v1.

Christiansen, C., Abou Hachem, M., Janeček, Š., Viksø-Nielsen, A., Blennow, A. and Svensson, B., 2009. The carbohydrate-binding module family 20–diversity, structure, and function. *The FEBS journal*, *276*(18), pp.5006-5029.

Ciniglia, C., Yoon, H.S., Pollio, A., Pinto, G. and Bhattacharya, D., 2004. Hidden biodiversity of the extremophilic Cyanidiales red algae. *Molecular Ecology*, *13*(7), pp.1827-1838.

Ciniglia, C., Yang, E.C., Pollio, A., Pinto, G., Iovinella, M., Vitale, L. and Yoon, H.S., 2014. Cyanidiophyceae in Iceland: plastid rbc L gene elucidates origin and dispersal of extremophilic Galdieria sulphuraria and G. maxima (Galdieriaceae, Rhodophyta). *Phycologia*, *53*(6), pp.542-551

Čížková, M, K Bišová, V Zachleder, D Mezricky, M Rucki, and M Vítová. 2019. “Recovery of Rare Earth Elements from Luminophores Using the Red Alga Galdieria,” no. September: 2018–19.

Coker, J.A., 2016. Extremophiles and biotechnology: current uses and prospects. *F1000Research*, *5*.

Cragg, S.M., Beckham, G.T., Bruce, N.C., Bugg, T.D., Distel, D.L., Dupree, P., Etxabe, A.G., Goodell, B.S., Jellison, J., McGeehan, J.E. and McQueen-Mason, S.J., 2015. Lignocellulose degradation mechanisms across the Tree of Life. *Current opinion in chemical biology*, *29*, pp.108-119.

Crawford, D.L., 2001. Functional genomics does not have to be limited to a few select organisms. *Genome Biology*, *2*(1), pp.1-2.

Curien, G., Lyska, D., Guglielmino, E., Westhoff, P., Janetzko, J., Tardif, M., Hallopeau, C., Brugière, S., Dal Bo, D., Decelle, J. and Gallet, B., 2021. Mixotrophic growth of the extremophile galdieria sulphuraria reveals the flexibility of its carbon assimilation metabolism. *New Phytologist*.

Dahadha, S., Amin, Z., Bazyar Lakeh, A. A., and Elbeshbishy, E. (2017). Evaluation of different pretreatment processes of lignocellulosic biomass for enhanced biomethane production. *Energy Fuels.* 31, 10335–10347. doi: 10.1021/acs.energyfuels.7b02045

Dahman, Y., Syed, K., Begum, S., Roy, P. and Mohtasebi, B., 2019. Biofuels: Their characteristics and analysis. In *Biomass, Biopolymer-Based Materials, and Bioenergy* (pp. 277-325). Woodhead Publishing.

DasSarma, P. and DasSarma, S., 2018. Survival of microbes in Earth's stratosphere. *Current opinion in microbiology*, *43*, pp.24-30.

Del Amparo, R., Branco, C., Arenas, J., Vicens, A. and Arenas, M., 2021. Analysis of selection in protein-coding sequences accounting for common biases. *Briefings in Bioinformatics*.

De Luca, P., Taddei, R. and Varano, L., 1978. «Cyanidioschyzon merolae»: a new alga of thermal acidic environments. *Webbia*, *33*(1), pp.37-44.

De Oliveira, I.P. and Martínez, L., 2020. The shift in urea orientation at protein surfaces at low pH is compatible with a direct mechanism of protein denaturation. *Physical Chemistry Chemical Physics*, *22*(1), pp.354-367.

Demirbas, A., 2009. Biofuels securing the planet's future energy needs. *Energy conversion and management*, *50*(9), pp.2239-2249.

Demirbas, A., 2008. Biofuels sources, biofuel policy, biofuel economy and global biofuel projections. *Energy conversion and management*, *49*(8), pp.2106-2116.

Dhakar, K. and Pandey, A., 2016. Wide pH range tolerance in extremophiles: towards understanding an important phenomenon for future biotechnology. *Applied microbiology and biotechnology*, *100*(6), pp.2499-2510.

Dodds, W. and Whiles, M., 2018. *Freshwater ecology*. Amsterdam [i pozostałe]: Academic Press, an imprint of Elsevier, pp.375-398.

Dong, W.L., Wang, R.N., Zhang, N.Y., Fan, W.B., Fang, M.F. and Li, Z.H., 2018. Molecular evolution of chloroplast genomes of orchid species: insights into phylogenetic relationship and adaptive evolution. *International Journal of Molecular Sciences*, *19*(3), p.716.

Dos Santos, A.C., Ximenes, E., Kim, Y. and Ladisch, M.R., 2019. Lignin–enzyme interactions in the hydrolysis of lignocellulosic biomass. *Trends in biotechnology*, *37*(5), pp.518-531.

Drouault, S., Anba, J., Bonneau, S., Bolotin, A., Ehrlich, S.D. and Renault, P., 2002. The peptidyl-prolyl isomerase motif is lacking in PmpA, the PrsA-like protein involved in the secretion machinery of Lactococcus lactis. *Applied and Environmental Microbiology*, *68*(8), pp.3932-3942

Drouin, G., Daoud, H. and Xia, J., 2008. Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants. *Molecular phylogenetics and evolution*, *49*(3), pp.827-831

Drummond, D.A., Bloom, J.D., Adami, C., Wilke, C.O. and Arnold, F.H., 2005. Why highly expressed proteins evolve slowly. *Proceedings of the National Academy of Sciences*, *102*(40), pp.14338-14343.

Dumorne K, Cordova DC, Astorga-Elo M & Renganathan P (2017) Extremozymes: a potential source for industrial applications. *J Microbiol Biotechn* **27**, 649– 659.

Duroux L, Welinder KG (2003) The peroxidase gene family in plants: a phylogenetic overview. J Mol Evol 57:397–407

Dutta, K., Daverey, A. and Lin, J.G., 2014. Evolution retrospective for alternative fuels: First to fourth generation. *Renewable energy*, *69*, pp.114-122.

Dutta, S. and Wu, K.C.W., 2014. Enzymatic breakdown of biomass: enzyme active sites, immobilization, and biofuel production. *Green chemistry*, *16*(11), pp.4615-4626.

Edbeib, M.F., Wahab, R.A. and Huyop, F., 2016. Halophiles: biology, adaptation, and their role in decontamination of hypersaline environments. *World Journal of Microbiology and Biotechnology*, *32*(8), pp.1-23.

Edgar, R.C., 2010. Quality measures for protein alignment benchmarks. *Nucleic acids research*, *38*(7), pp.2145-2153.

Eibinger, M., Sattelkow, J., Ganner, T., Plank, H. and Nidetzky, B., 2017. Single-molecule study of oxidative enzymatic deconstruction of cellulose. *Nature communications*, *8*(1), pp.1-7.

Eichler, J., 2001. Biotechnological uses of archaeal extremozymes. *Biotechnology advances*, *19*(4), pp.261-278.

Eisele, L.E., Bakhru, S.H., Liu, X., MacColl, R. and Edwards, M.R., 2000. Studies on C-phycocyanin from Cyanidium caldarium, a eukaryote at the extremes of habitat. *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, *1456*(2-3), pp.99-107.

Emms, D.; Kelly, S. OrthoFinder: Solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **2015**, *16*, 157

Emms, D.M. and Kelly, S., 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome biology*, *20*(1), pp.1-14.

Enami, I., Akutsu, H. and Kyogoku, Y., 1986. Intracellular pH regulation in an acidophilic unicellular alga, Cyanidium caldarium: 31P-NMR determination of intracellular pH. *Plant and cell physiology*, *27*(7), pp.1351-1359.

Ende, S.S. and Noke, A., 2019. Heterotrophic microalgae production on food waste and by-products. *Journal of Applied Phycology*, *31*(3), pp.1565-1571.

Eren, A., Iovinella, M., Yoon, H.S., Cennamo, P., de Stefano, M., de Castro, O. and Ciniglia, C., 2018. Genetic structure of Galdieria populations from Iceland. *Polar Biology*, *41*(9), pp.1681-1691.

Espliego, J.M.E., Saiz, V.B., Torregrosa-Crespo, J., Luque, A.V., Carrasco, M.L.C., Pire, C., Bonete, M.J. and Martínez-Espinosa, R.M., 2018. Extremophile enzymes and biotechnology. In *Extremophiles* (pp. 227-248). CRC Press.

Eyre-Walker, A., 2006. The genomic rate of adaptive evolution. *Trends in ecology & evolution*, *21*(10), pp.569-575.

Eyre-Walker, A. and Keightley, P.D., 2007. The distribution of fitness effects of new mutations. *Nature Reviews Genetics*, *8*(8), pp.610-618.

Ezeilo, U.R., Zakaria, I.I., Huyop, F. and Wahab, R.A., 2017. Enzymatic breakdown of lignocellulosic biomass: the role of glycosyl hydrolases and lytic polysaccharide monooxygenases. *Biotechnology & Biotechnological Equipment*, *31*(4), pp.647-662.

Fawal, N., Li, Q., Savelli, B., Brette, M., Passaia, G., Fabre, M., Mathe, C. and Dunand, C., 2012. PeroxiBase: a database for large-scale evolutionary analysis of peroxidases. *Nucleic acids research*, *41*(D1), pp.D441-D444.

Fernández-Bayo, J.D., Hestmark, K.V., Claypool, J.T., Harrold, D.R., Randall, T.E., Achmon, Y., Stapleton, J.J., Simmons, C.W. and VanderGheynst, J.S., 2019. The initial soil microbiota impacts the potential for lignocellulose degradation during soil solarization. *Journal of applied microbiology*, *126*(6), pp.1729-1741.

Fouke, B.W., 2011. Hot-spring Systems Geobiology: abiotic and biotic influences on travertine formation at Mammoth Hot Springs, Yellowstone National Park, USA. *Sedimentology*, *58*(1), pp.170-219.

Freshwater, D.W., Fredericq, S., Butler, B.S., Hommersand, M.H. and Chase, M.W., 1994. A gene phylogeny of the red algae (Rhodophyta) based on plastid rbcL. *Proceedings of the National Academy of Sciences*, *91*(15), pp.7281-7285.

Freudenberg, R.A., Baier, T., Einhaus, A., Wobbe, L. and Kruse, O., 2021. High cell density cultivation enables efficient and sustainable recombinant polyamine production in the microalga Chlamydomonas reinhardtii. *Bioresource Technology*, *323*, p.124542.

Frösler, J., Panitz, C., Wingender, J., Flemming, H.C. and Rettberg, P., 2017. Survival of Deinococcus geothermalis in biofilms under desiccation and simulated space and martian conditions. *Astrobiology*, *17*(5), pp.431-447.

Gabani, P. and Singh, O.V., 2013. Radiation-resistant extremophiles and their potential in biotechnology and therapeutics. *Applied microbiology and biotechnology*, *97*(3), pp.993-1004.

Geismar, H.N., McCarl, B.A. and Searcy, S.W., 2021. Optimal Design and Operation of a Second-Generation Biofuels Supply Chain. *IISE Transactions*, (just-accepted), pp.1-35.

Gellissen, G. ed., 2006. *Production of recombinant proteins: Novel microbial and eukaryotic expression systems*. John Wiley & Sons.

Gomes, A.R., Byregowda, S.M., Veeregowda, B.M. and Balamurugan, V., 2016. An overview of heterologous expression host systems for the production of recombinant proteins.

Gourlay, K., Hu, J., Arantes, V., Andberg, M., Saloheimo, M., Penttilä, M. and Saddler, J., 2013. Swollenin aids in the amorphogenesis step during the enzymatic hydrolysis of pretreated biomass. *Bioresource technology*, *142*, pp.498-503.

Gimmler, H. and Weis, U., 1992. Dunaliella acidophila–life at pH 1.0. *Dunaliella: Physiology, Biochemistry and Biotechnology*, pp.99-133.

Glenn, E.P., Brown, J.J. and Blumwald, E., 1999. Salt tolerance and crop potential of halophytes. *Critical reviews in plant sciences*, *18*(2), pp.227-255.

Gonsior, M., Hertkorn, N., Hinman, N., Dvorski, S.E.M., Harir, M., Cooper, W.J. and Schmitt-Kopplin, P., 2018. Yellowstone Hot Springs are organic chemodiversity hot spots. *Scientific reports*, *8*(1), pp.1-13.

Graverholt, O.S. and Eriksen, N.T., 2007. Heterotrophic high-cell-density fed-batch and continuous-flow cultures of Galdieria sulphuraria and production of phycocyanin. *Applied microbiology and biotechnology*, *77*(1), pp.69-75.

Graziani, G., Schiavo, S., Nicolai, M.A., Buono, S., Fogliano, V., Pinto, G. and Pollio, A., 2013. Microalgae as human food: chemical and nutritional characteristics of the thermo-acidophilic microalga Galdieria sulphuraria. *Food & function*, *4*(1), pp.144-152.

Gregory, T.R., 2005. Genome size evolution in animals. In *The evolution of the genome* (pp. 3-87). Academic Press.

Gross, W., 2000. Ecophysiology of algae living in highly acidic environments. *Hydrobiologia*, *433*(1), pp.31-37.

Gross W and Schnarrenberger C. Heterotrophic Growth of Two Strains of the Acido-

Thermophilic Red Alga Galdieria sulphuraria. Plant. Cell. Physiol., 36, 633-638 (1995).

Gross, W., Kuever, J., Tischendorf, G., Bouchaala, N. and Büsch, W., 1998. Cryptoendolithic growth of the red alga Galdieria sulphuraria in volcanic areas. *European Journal of Phycology*, *33*(1), pp.25-31.

Gross, W. and Oesterhelt, C., 1999. Ecophysiological studies on the red alga Galdieria sulphuraria isolated from southwest Iceland. *Plant biology*, *1*(06), pp.694-700

Gunde-Cimerman, N., Plemenitaš, A. and Oren, A., 2018. Strategies of adaptation of microorganisms of the three domains of life to high salt concentrations. *FEMS microbiology reviews*, *42*(3), pp.353-375

Hamdan, A., 2018. Psychrophiles: Ecological significance and potential industrial application. *South African Journal of Science*, *114*(5-6), pp.1-6.

Han, Y., Liu, X., Nan, F., Feng, J., Lv, J., Liu, Q. and Xie, S., 2021. Analysis of Adaptive Evolution and Coevolution of rbc L Gene in the Genus Galdieria (Rhodophyta). *Journal of Eukaryotic Microbiology*, *68*(2), p.e12838.

Harrison, M.D., Geijskes, J., Coleman, H.D., Shand, K., Kinkema, M., Palupe, A., Hassall, R., Sainz, M., Lloyd, R., Miles, S. and Dale, J.L., 2011. Accumulation of recombinant cellobiohydrolase and endoglucanase in the leaves of mature transgenic sugar cane. *Plant Biotechnology Journal*, *9*(8), pp.884-896.

Hassan, S.S., Williams, G.A. and Jaiswal, A.K., 2018. Emerging technologies for the pretreatment of lignocellulosic biomass. *Bioresource Technology*, *262*, pp.310-318.

Henkanatte-Gedera, S.M., Selvaratnam, T., Karbakhshravari, M., Myint, M., Nirmalakhandan, N., Van Voorhies, W. and Lammers, P.J., 2017. Removal of dissolved organic carbon and nutrients from urban wastewaters by Galdieria sulphuraria: Laboratory to field scale demonstration. *Algal Research*, *24*, pp.450-456.

Henrissat, B., 1991. A classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochemical journal*, *280*(2), pp.309-316.

Hill, T., Nordström, K.J., Thollesson, M., Säfström, T.M., Vernersson, A.K., Fredriksson, R. and Schiöth, H.B., 2010. SPRIT: Identifying horizontal gene transfer in rooted phylogenetic trees. *BMC evolutionary biology*, *10*(1), pp.1-9.

Hilpmann, G., Becher, N., Pahner, F.A., Kusema, B., Mäki-Arvela, P., Lange, R., Murzin, D.Y. and Salmi, T., 2016. Acid hydrolysis of xylan. *Catalysis Today*, *259*, pp.376-380.

Hiraga, S., Sasaki, K., Ito, H., Ohashi, Y. and Matsui, H., 2001. A large family of class III plant peroxidases. *Plant and Cell Physiology*, *42*(5), pp.462-468.

Hittinger, C.T. and Carroll, S.B., 2007. Gene duplication and the adaptive evolution of a classic genetic switch. *Nature*, *449*(7163), pp.677-681.

Hoffmann, L., 1994. Cyanidium-like algae from caves. In *Evolutionary Pathways and Enigmatic Algae: Cyanidium Caldarium (Rhodophyta) and Related Cells* (pp. 175-182). Springer, Dordrecht.

Holzinger, A. and Karsten, U., 2013. Desiccation stress and tolerance in green algae: consequences for ultrastructure, physiological and molecular mechanisms. *Frontiers in plant science*, *4*, p.327.

Horikoshi, K., 2016. Alkaliphiles. In *Extremophiles* (pp. 53-78). Springer, Tokyo.

Houfani, A.A., Anders, N., Spiess, A.C., Baldrian, P. and Benallaoua, S., 2020. Insights from enzymatic degradation of cellulose and hemicellulose to fermentable sugars–a review. *Biomass and Bioenergy*, *134*, p.105481.

Hsieh, Chia Jung, Shing Hei Zhan, Yiching Lin, Sen Lin Tang, and Shao Lun Liu. 2015. “Analysis of RbcL Sequences Reveals the Global Biodiversity, Community Structure, and Biogeographical Pattern of Thermoacidophilic Red Algae (Cyanidiales).” Journal of Phycology 51: 682–94. https://doi.org/10.1111/jpy.12310.

Husain, Q., 2010. β Galactosidases and their potential applications: a review. *Critical reviews in biotechnology*, *30*(1), pp.41-62.

Ihsan, N., 2017. *Identifying Novel Lignocellulosic Processing Enzymes from Cellulomonas fimi using Transcriptomic, Proteomic and Evolution Adaptive Studies* (Doctoral dissertation, University of York).

Isenberg, I., 1979. Histones. *Annual review of biochemistry*, *48*(1), pp.159-191.

Isikgor, F.H. and Becer, C.R., 2015. Lignocellulosic biomass: a sustainable platform for the production of bio-based chemicals and polymers. *Polymer Chemistry*, *6*(25), pp.4497-4559.

Janeček, Š., Svensson, B. and MacGregor, E.A., 2011. Structural and evolutionary aspects of two families of non-catalytic domains present in starch and glycogen binding proteins from microbes, plants and animals. *Enzyme and microbial technology*, *49*(5), pp.429-440.

Janeček, Š, Mareček, F., MacGregor, E. A., and Svensson, B. (2019). Starch-binding domains as CBM families–history, occurrence, structure, function and evolution. *Biotechnol. Adv.* 37:107451. doi: 10.1016/j.biotechadv.2019.107451

Jeffares, D.C., Tomiczek, B., Sojo, V. and dos Reis, M., 2015. A beginners guide to estimating the non-synonymous to synonymous rate ratio of all protein-coding genes in a genome. In *Parasite genomics protocols* (pp. 65-90). Humana Press, New York, NY.

Jiang, L., Wu, N., Zheng, A., Zhao, Z., He, F. and Li, H., 2016. The integration of dilute acid hydrolysis of xylan and fast pyrolysis of glucan to obtain fermentable sugars. *Biotechnology for biofuels*, *9*(1), pp.1-10.

Jiang, Q., Qin, S. and Wu, Q.Y., 2010. Genome-wide comparative analysis of metacaspases in unicellular and filamentous cyanobacteria. *BMC genomics*, *11*(1), pp.1-11.

Jin, Q. and Kirk, M.F., 2018. pH as a primary control in environmental microbiology: 1. thermodynamic perspective. *Frontiers in Environmental Science*, *6*, p.21.

Jin, M., Gai, Y., Guo, X., Hou, Y. and Zeng, R., 2019. Properties and applications of extremozymes from deep-sea extremophilic microorganisms: A mini review. *Marine drugs*, *17*(12), p.656.

Jo, B.S. and Choi, S.S., 2015. Introns: the functional benefits of introns in genomes. *Genomics & informatics*, *13*(4), p.112.

Kao, O.H., Edwards, M.R. and Berns, D.S., 1975. Physical-chemical properties of C-phycocyanin isolated from an acido-thermophilic eukaryote, Cyanidium caldarium. *Biochemical Journal*, *147*(1), pp.63-70.

Katayama, T., Nagao, N., Kasan, N.A., Khatoon, H., Rahman, N.A., Takahashi, K., Furuya, K., Yamada, Y., Abd Wahid, M.E. and Jusoh, M., 2020. Bioprospecting of indigenous marine microalgae with ammonium tolerance from aquaculture ponds for microalgae cultivation with ammonium-rich wastewaters. *Journal of Biotechnology*, *323*, pp.113-120.

Karpievitch, Y.V., Polpitiya, A.D., Anderson, G.A., Smith, R.D. and Dabney, A.R., 2010. Liquid chromatography mass spectrometry-based proteomics: biological and technological aspects. *The annals of applied statistics*, *4*(4), p.1797.

Keeling, P.J. and Slamovits, C.H., 2005. Causes and effects of nuclear genome reduction. *Current opinion in genetics & development*, *15*(6), pp.601-608.

Khow, O. and Suntrarachun, S., 2012. Strategies for production of active eukaryotic proteins in bacterial expression system. *Asian Pacific journal of tropical biomedicine*, *2*(2), pp.159-162.

Kinnunen, A., Maijala, P., JArvinen, P. and Hatakka, A., 2017. Improved efficiency in screening for lignin-modifying peroxidases and laccases of basidiomycetes. *Current Biotechnology*, *6*(2), pp.105-115.

Koren S, Walenz BP, Berlin K, Miller JR, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Research. (2017).

Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A (2008) Patterns of positive selection in six Mammalian genomes. PLoS Genet 4:e1000144

Kristjansson, J.K. and Hreggvidsson, G.O., 1995. Ecology and habitats of extremophiles. *World Journal of Microbiology and Biotechnology*, *11*(1), pp.17-25.

Krulwich, T.A., Sachs, G. and Padan, E., 2011. Molecular aspects of bacterial pH sensing and homeostasis. *Nature Reviews Microbiology*, *9*(5), pp.330-343.

Kryazhimskiy, S. and Plotkin, J.B., 2008. The population genetics of dN/dS. *PLoS genetics*, *4*(12), p.e1000304.

Kumondai, M., Hishinuma, E., Rico, E.M.G., Ito, A., Nakanishi, Y., Saigusa, D., Hirasawa, N. and Hiratsuka, M., 2020. Heterologous expression of high-activity cytochrome P450 in mammalian cells. *Scientific reports*, *10*(1), pp.1-13.

Lage, C.A., Dalmaso, G.Z., Teixeira, L.C., Bendia, A.G., Paulino-Lima, I.G., Galante, D., Janot-Pacheco, E., Abrevaya, X.C., Azúa-Bustos, A., Pelizzari, V.H. and Rosado, A.S., 2012. Mini-Review: Probing the limits of extremophilic life in extraterrestrial environment-simulated experiments. *International Journal of Astrobiology*, *11*(4), pp.251-256.

Lai, C.P., Huang, L.M., Chen, L.F.O., Chan, M.T. and Shaw, J.F., 2017. Genome-wide analysis of GDSL-type esterases/lipases in Arabidopsis. *Plant molecular biology*, *95*(1), pp.181-197.

Lairson, L.L., Henrissat, B., Davies, G.J. and Withers, S.G., 2008. Glycosyltransferases: structures, functions, and mechanisms. *Annu. Rev. Biochem.*, *77*, pp.521-555.

Laksanalamai, P. and Robb, F.T., 2004. Small heat shock proteins from extremophiles: a review. *Extremophiles*, *8*(1), pp.1-11.

Lane, N. and Martin, W., 2010. The energetics of genome complexity. *Nature*, *467*(7318), pp.929-934.

Larracuente, A.M., Sackton, T.B., Greenberg, A.J., Wong, A., Singh, N.D., Sturgill, D., Zhang, Y., Oliver, B. and Clark, A.G., 2008. Evolution of protein-coding genes in Drosophila. *Trends in Genetics*, *24*(3), pp.114-123.

Last, W.M., 2002. Geolimnology of salt lakes. *Geosciences Journal*, *6*(4), pp.347-369.

Lee, J., Ghosh, S. and Saier Jr, M.H., 2017. Comparative genomic analyses of transport proteins encoded within the red algae Chondrus crispus, Galdieria sulphuraria, and Cyanidioschyzon merolae11. *Journal of phycology*, *53*(3), pp.503-521.

Leliaert, F., Smith, D.R., Moreau, H., Herron, M.D., Verbruggen, H., Delwiche, C.F. and De Clerck, O., 2012. Phylogeny and molecular evolution of the green algae. *Critical reviews in plant sciences*, *31*(1), pp.1-46.

Leonelli S, Ankeny RA. What makes a model organism?. Endeavour. 2013 Dec 1;37(4):209-12.

Li, F., Ma, F., Zhao, H., Zhang, S., Wang, L., Zhang, X. and Yu, H., 2019. A lytic polysaccharide monooxygenase from a white-rot fungus drives the degradation of lignin by a versatile peroxidase. *Applied and environmental microbiology*, *85*(9), pp.e02803-18.

Li, H. and Durbin, R., 2010. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, *26*(5), pp.589-595.

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

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

Li, Q., et al., Plant biotechnology for lignocellulosic biofuel production. Plant Biotechnol J, 2014. 12(9): p. 1174-92.

Li, X. and Zheng, Y., 2020. Biotransformation of lignin: Mechanisms, applications and future work. *Biotechnology progress*, *36*(1), p.e2922.

Liguori, R., Ventorino, V., Pepe, O. and Faraco, V., 2016. Bioreactors for lignocellulose conversion into fermentable sugars for production of high added value products. *Applied microbiology and biotechnology*, *100*(2), pp.597-611.

Limayem A & Ricke SC (2012) Lignocellulosic biomass for bioethanol production: current perspectives, potential issues and future prospects. Prog Energ Combust 38, 449–467.

Linka, M., Jamai, A. and Weber, A.P., 2008. Functional characterization of the plastidic phosphate translocator gene family from the thermo-acidophilic red alga Galdieria sulphuraria reveals specific adaptations of primary carbon partitioning in green plants and red algae. *Plant physiology*, *148*(3), pp.1487-1496.

Lu, H., Wu, S., Li, A. and Ruan, J., 2021. SMARTdenovo: A de novo assembler using long noisy reads. *Gigabyte*, *2021*, pp.1-9.

Liu, S.L., Chiang, Y.R., Yoon, H.S. and Fu, H.Y., 2020. Comparative genome analysis reveals Cyanidiococcus gen. nov., a new extremophilic red algal genus sister to Cyanidioschyzon (Cyanidioschyzonaceae, Rhodophyta). *Journal of Phycology*, *56*(6), pp.1428-1442.

Liu, Z. and Zhang, J., 2018. Most m6A RNA modifications in protein-coding regions are evolutionarily unconserved and likely nonfunctional. *Molecular biology and evolution*, *35*(3), pp.666-675.

Levasseur, A., Drula, E., Lombard, V., Coutinho, P.M. and Henrissat, B., 2013. Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes. *Biotechnology for biofuels*, *6*(1), pp.1-14.

Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P.M. and Henrissat, B., 2014. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic acids research*, *42*(D1), pp.D490-D495.

Lopes, A.D.M., Ferreira Filho, E.X. and Moreira, L.R.S., 2018. An update on enzymatic cocktails for lignocellulose breakdown. *Journal of applied microbiology*, *125*(3), pp.632-645.

Lu, H., Giordano, F. and Ning, Z., 2016. Oxford Nanopore MinION sequencing and genome assembly. *Genomics, proteomics & bioinformatics*, *14*(5), pp.265-279.

Lynch, M. and Conery, J.S., 2003. The origins of genome complexity. *science*, *302*(5649), pp.1401-1404.

Lynch, M. and Walsh, B., 2007. *The origins of genome architecture* (Vol. 98). Sunderland, MA: Sinauer associates.

A.O., Szekrenyi, A., Joosten, H.J., Finnigan, J., Charnock, S. and Fessner, W.D., 2019. nanoDSF as screening tool for enzyme libraries and biotechnology development. *The FEBS journal*, *286*(1), pp.184-204.

Madern, D., Ebel, C. and Zaccai, G., 2000. Halophilic adaptation of enzymes. *Extremophiles*, *4*(2), pp.91-98.

Malerba, M.E., Ghedini, G. and Marshall, D.J., 2020. Genome size affects fitness in the eukaryotic alga Dunaliella tertiolecta. *Current Biology*, *30*(17), pp.3450-3456.

Mallick, N. and Mohn, F.H., 2000. Reactive oxygen species: response of algal cells. *Journal of Plant Physiology*, *157*(2), pp.183-193

Maloy, E., Hughes, K. and Maloy, S., 2013. *Brenner's Encyclopedia of Genetics (Second Edition)*. [Place of publication not identified]: Academic Press, pp.301-305.

Mamayev, O.I., 2010. *Temperature-salinity analysis of world ocean waters*. Elsevier.

Marriott, P.E., L.D. Gomez, and S.J. McQueen-Mason, Unlocking the potential of lignocellulosic biomass through plant science. New Phytol, 2016. 209(4): p. 1366-81.

Martinez-Fleites, C., Proctor, M., Roberts, S., Bolam, D.N., Gilbert, H.J. and Davies, G.J., 2006. Insights into the synthesis of lipopolysaccharide and antibiotics through the structures of two retaining glycosyltransferases from family GT4. *Chemistry & biology*, *13*(11), pp.1143-1152.

Matsuzaki, M., Misumi, O., Shin-i, T., Maruyama, S., Takahara, M., Miyagishima, S.Y., Mori, T., Nishida, K., Yagisawa, F., Nishida, K. and Yoshida, Y., 2004. Genome sequence of the ultrasmall unicellular red alga Cyanidioschyzon merolae 10D. *Nature*, *428*(6983), pp.653-657.

Matthews, B.J. and Vosshall, L.B., 2020. How to turn an organism into a model organism in 10 'easy'steps. *Journal of Experimental Biology*, *223*(Suppl_1), p.jeb218198.

McClure, C.P., Rusche, K.M., Peariso, K., Jackman, J.E., Fierke, C.A. and Penner-Hahn, J.E., 2003. EXAFS studies of the zinc sites of UDP-(3-O-acyl)-N-acetylglucosamine deacetylase (LpxC). *Journal of inorganic biochemistry*, *94*(1-2), pp.78-85.

Merino, N., Aronson, H.S., Bojanova, D.P., Feyhl-Buska, J., Wong, M.L., Zhang, S. and Giovannelli, D., 2019. Living at the extremes: extremophiles and the limits of life in a planetary context. *Frontiers in microbiology*, *10*, p.780.

Merola, A., Castaldo, R., Luca, P.D., Gambardella, R., Musacchio, A. and Taddei, R., 1981. Revision of Cyanidium caldarium. Three species of acidophilic algae. *Plant Biosystem*, *115*(4-5), pp.189-195.

Merino, N., Aronson, H.S., Bojanova, D.P., Feyhl-Buska, J., Wong, M.L., Zhang, S. and Giovannelli, D., 2019. Living at the extremes: extremophiles and the limits of life in a planetary context. *Frontiers in microbiology*, *10*, p.780.

Miller, J.H., 1972. Experiments in molecular genetics.

Minh, B.Q., Hahn, M.W. and Lanfear, R., 2020. New methods to calculate concordance factors for phylogenomic datasets. *Molecular biology and evolution*, *37*(9), pp.2727-2733.

Mittal, A., Pilath, H.M., Parent, Y., Chatterjee, S.G., Donohoe, B.S., Yarbrough, J.M., Black, S.K., Himmel, M.E., Nimlos, M.R. and Johnson, D.K., 2019. Chemical and structural effects on the rate of xylan hydrolysis during dilute acid pretreatment of poplar wood. *ACS Sustainable Chemistry & Engineering*, *7*(5), pp.4842-4850.

Moreira, L.R.S., 2016. Insights into the mechanism of enzymatic hydrolysis of xylan. *Applied microbiology and biotechnology*, *100*(12), pp.5205-5214

Moreira, D., López-Archilla, A.I., Amils, R. and Marín, I., 1994. Characterization of two new thermoacidophilic microalgae: genome organization and comparison with Galdieria sulphuraria. *FEMS microbiology letters*, *122*(1-2), pp.109-114.

Moutinho, A.F., Bataillon, T. and Dutheil, J.Y., 2020. Variation of the adaptive substitution rate between species and within genomes. *Evolutionary Ecology*, *34*(3), pp.315-338.

Mukhtar, S. and Aslam, M., 2020. Biofuel Synthesis by Extremophilic Microorganisms. In *Biofuels Production–Sustainability and Advances in Microbial Bioresources* (pp. 115-138). Springer, Cham.

Mulder, S.J., 2016. *Cultivating Galdieria sulphuraria on a lactose waste stream to improve sustainability* (Doctoral dissertation, Faculty of Science and Engineering).

Náhlík, V., Zachleder, V., Čížková, M., Bišová, K., Singh, A., Mezricky, D., Řezanka, T. and Vítová, M., 2021. Growth under Different Trophic Regimes and Synchronization of the Red Microalga Galdieria sulphuraria. *Biomolecules*, *11*(7), p.939.

Naik, S.N., Goud, V.V., Rout, P.K. and Dalai, A.K., 2010. Production of first and second generation biofuels: a comprehensive review. *Renewable and sustainable energy reviews*, *14*(2), pp.578-597.

Nakamura, A.M., Nascimento, A.S. and Polikarpov, I., 2017. Structural diversity of carbohydrate esterases. *Biotechnology Research and Innovation*, *1*(1), pp.35-51.

Naumoff, D.G., 2011. Hierarchical classification of glycoside hydrolases. *Biochemistry (Moscow)*, *76*(6), pp.622-635.

Navarro, F., Forján, E., Vázquez, M., Toimil, A., Montero, Z., Ruiz-Domínguez, M.D.C., Garbayo, I., Castaño, M.Á., Vílchez, C. and Vega, J.M., 2017. Antimicrobial activity of the acidophilic eukaryotic microalga Coccomyxa onubensis. *Phycological research*, *65*(1), pp.38-43.

Nicolaus, B., Kambourova, M. and Oner, E.T., 2010. Exopolysaccharides from extremophiles: from fundamentals to biotechnology. *Environmental Technology*, *31*(10), pp.1145-1158.

Nishida, K., Yagisawa, F., Kuroiwa, H., Nagata, T. and Kuroiwa, T., 2005. Cell cycle-regulated, microtubule-independent organelle division in Cyanidioschyzon merolae. *Molecular biology of the cell*, *16*(5), pp.2493-2502.

Nosaka, Y. and Nosaka, A.Y., 2017. Generation and detection of reactive oxygen species in photocatalysis. *Chemical reviews*, *117*(17), pp.11302-11336.

Nguyen, S.T., Freund, H.L., Kasanjian, J. and Berlemont, R., 2018. Function, distribution, and annotation of characterized cellulases, xylanases, and chitinases from CAZy. *Applied microbiology and biotechnology*, *102*(4), pp.1629-1637.

Nühse, T.S., Bottrill, A.R., Jones, A.M. and Peck, S.C., 2007. Quantitative phosphoproteomic analysis of plasma membrane proteins reveals regulatory mechanisms of plant innate immune responses. *The Plant Journal*, *51*(5), pp.931-940.

Oarga, A., 2009. Life in extreme environments. *Revista de Biologia e ciencias da Terra*, *9*(1), pp.1-10.

Oesterhelt, C., Schnarrenberger, C. and Gross, W., 1999. Characterization of a sugar/polyol uptake system in the red alga Galdieria sulphuraria. *European journal of Phycology*, *34*(3), pp.271-277.

Oesterhelt C and Gross W. Different Sugar Kinases Are Involved in the Sugar Sensing of Galdieria Sulphuraria. Plant. Physiol. 128, 291-299 (2002).

Oesterhelt, C., Schmälzlin, E., Schmitt, J.M. and Lokstein, H., 2007. Regulation of photosynthesis in the unicellular acidophilic red alga Galdieria sulphuraria. *The Plant Journal*, *51*(3), pp.500-511.

Oesterhelt, C., Vogelbein, S., Shrestha, R.P., Stanke, M. and Weber, A.P.M., 2008. The genome of the thermoacidophilic red microalga Galdieria sulphuraria encodes a small family of secreted class III peroxidases that might be involved in cell wall modification. *Planta*, *227*(2), pp.353-362.

Oliver, M.J., Farrant, J.M., Hilhorst, H.W., Mundree, S., Williams, B. and Bewley, J.D., 2020. Desiccation tolerance: avoiding cellular damage during drying and rehydration. *Annual Review of Plant Biology*, *71*, pp.435-460.

Onofri, S., Selbmann, L., Pacelli, C., De Vera, J.P., Horneck, G., Hallsworth, J.E. and Zucconi, L., 2018. Integrity of the DNA and cellular ultrastructure of cryptoendolithic fungi in space or Mars conditions: a 1.5-year study at the International Space Station. *Life*, *8*(2), p.23.

Oren, A., 2020. Ecology of extremely halophilic microorganisms. In *The biology of halophilic bacteria* (pp. 25-53). CRC Press.

Papaneophytou, C.P. and Kontopidis, G., 2014. Statistical approaches to maximize recombinant protein expression in Escherichia coli: a general review. *Protein expression and purification*, *94*, pp.22-32.

Paquola, A.C., Asif, H., de Bragança Pereira, C.A., Feltes, B.C., Bonatto, D., Lima, W.C. and Menck, C.F.M., 2018. Horizontal gene transfer building prokaryote genomes: genes related to exchange between cell and environment are frequently transferred. *Journal of molecular evolution*, *86*(3), pp.190-203.

Park, Y.J., Jeong, Y.U. and Kong, W.S., 2018. Genome sequencing and carbohydrate-active enzyme (CAZyme) repertoire of the white rot fungus Flammulina elastica. *International journal of molecular sciences*, *19*(8), p.2379.

Parmley, J.L. and Hurst, L.D., 2007. How do synonymous mutations affect fitness?. *Bioessays*, *29*(6), pp.515-519.

Peng, B., Williams, T.C., Henry, M., Nielsen, L.K. and Vickers, C.E., 2015. Controlling heterologous gene expression in yeast cell factories on different carbon substrates and across the diauxic shift: a comparison of yeast promoter activities. *Microbial cell factories*, *14*(1), pp.1-11.

Petersen, T.N.; Brunak, S.; von Heijne, G.; Nielsen, H. SignalP 4.0: Discriminating signal peptides from transmembrane regions. *Nat. Methods* **2011**, *8*, 785–786.

Peterson, C.L. and Laniel, M.A., 2004. Histones and histone modifications. *Current Biology*, *14*(14), pp.R546-R551.

Piccinni, F.E., Ontañon, O.M., Ghio, S., Sauka, D.H., Talia, P.M., Rivarola, M.L., Valacco, M.P. and Campos, E., 2019. Secretome profile of Cellulomonas sp. B6 growing on lignocellulosic substrates. *Journal of applied microbiology*, *126*(3), pp.811-825.

Pinto, G., Albertano, P., Ciniglia, C., Cozzolino, S., Pollio, A., Yoon, H.S. and Bhattacharya, D., 2003. Comparative approaches to the taxonomy of the genus Galdieria merola (Cyanidiales, Rhodophyta). *Cryptogamie-Algologie*, *24*(1), pp.13-32.

Pinto, G., Ciniglia, C., Cascone, C. and Pollio, A., 2007. Species composition of Cyanidiales assemblages in Pisciarelli (Campi Flegrei, Italy) and description of Galdieria phlegrea sp. nov. In *Algae and cyanobacteria in extreme environments* (pp. 487-502). Springer, Dordrecht.

Pleissner, D., Lindner, A.V. and Händel, N., 2021. Heterotrophic cultivation of Galdieria sulphuraria under non-sterile conditions in digestate and hydrolyzed straw. *Bioresource Technology*, *337*, p.125477.

Pray, L., 2008. Eukaryotic genome complexity. *Nature Education*, *1*(1), p.96.

Qiu, H., Rossoni, A.W., Weber, A.P., Yoon, H.S. and Bhattacharya, D., 2018. Unexpected conservation of the RNA splicing apparatus in the highly streamlined genome of Galdieria sulphuraria. *BMC evolutionary biology*, *18*(1), pp.1-11.

Raddadi, N., Cherif, A., Daffonchio, D., Neifar, M. and Fava, F., 2015. Biotechnological applications of extremophiles, extremozymes and extremolytes. *Applied microbiology and biotechnology*, *99*(19), pp.7907-7913.

Rahman, D.Y., Sarian, F.D. and van der Maarel, M.J., 2020. Biomass and phycocyanin content of heterotrophic Galdieria sulphuraria 074G under maltodextrin and granular starches–feeding conditions. *Journal of Applied Phycology*, *32*(1), pp.51-57.

Rajendran, K., Drielak, E., Varma, V. S., Muthusamy, S., and Kumar, G. (2017). Updates on the pretreatment of lignocellulosic feedstocks for bioenergy production–a review. *Biomass Conver. Biorefin*. 8, 471–483. doi: 10.1007/s13399-017-0269-3

Rampelotto, P.H., 2013. Extremophiles and extreme environments.

Raud, M., Kikas, T., Sippula, O. and Shurpali, N.J., 2019. Potentials and challenges in lignocellulosic biofuel production technology. *Renewable and Sustainable Energy Reviews*, *111*, pp.44-56.

Rennie, E.A. and Scheller, H.V., 2014. Xylan biosynthesis. *Current opinion in biotechnology*, *26*, pp.100-107.

Reyes-Ortiz, V., Heins, R. A., Cheng, G., Kim, E. Y., Vernon, B. C., Elandt, R. B., et al., (2013). Addition of a carbohydrate-binding module enhances cellulase penetration into cellulose substrates. *Biotechnol. Biofuels* 6:93. doi: 10.1186/1754-6834-6-93

Rezania, S., Oryani, B., Cho, J., Talaiekhozani, A., Sabbagh, F., Hashemi, B., Rupani, P.F. and Mohammadi, A.A., 2020. Different pretreatment technologies of lignocellulosic biomass for bioethanol production: An overview. *Energy*, *199*, p.117457.

Rigano, C., Aliotta, G., Rigano, V.D.M., Fuggi, A. and Vona, V., 1977. Heterotrophic growth patterns in the unicellular alga Cyanidium caldarium. *Archives of microbiology*, *113*(3), pp.191-196.

Rigano, C., Fuggi, A., Rigano, V.D.M. and Aliotta, G., 1976. Studies on utilization of 2-ketoglutarate, glutamate and other amino acids by the unicellular alga Cyanidium caldarium. *Archives of microbiology*, *107*(2), pp.133-138.

Rocha, E.P., Smith, J.M., Hurst, L.D., Holden, M.T., Cooper, J.E., Smith, N.H. and Feil, E.J., 2006. Comparisons of dN/dS are time dependent for closely related bacterial genomes. *Journal of theoretical biology*, *239*(2), pp.226-235.

Rodionova, M.V., Poudyal, R.S., Tiwari, I., Voloshin, R.A., Zharmukhamedov, S.K., Nam, H.G., Zayadan, B.K., Bruce, B.D., Hou, H.J.M. and Allakhverdiev, S.I., 2017. Biofuel production: challenges and opportunities. *International Journal of Hydrogen Energy*, *42*(12), pp.8450-8461.

Rodolfi, L., Chini Zittelli, G., Bassi, N., Padovani, G., Biondi, N., Bonini, G. and Tredici, M.R., 2009. Microalgae for oil: Strain selection, induction of lipid synthesis and outdoor mass cultivation in a low-cost photobioreactor. *Biotechnology and bioengineering*, *102*(1), pp.100-112.

Rosenberg, A.H., Lade, B.N., Dao-shan, C., Lin, S.W., Dunn, J.J. and Studier, F.W., 1987. Vectors for selective expression of cloned DNAs by T7 RNA polymerase. *Gene*, *56*(1), pp.125-135.

Rossoni, A.W., Schönknecht, G., Lee, H.J., Rupp, R.L., Flachbart, S., Mettler-Altmann, T., Weber, A.P. and Eisenhut, M., 2019. Cold acclimation of the thermoacidophilic red alga Galdieria sulphuraria: Changes in gene expression and involvement of horizontally acquired genes. *Plant and Cell Physiology*, *60*(3), pp.702-712.

Rothschild, L.J. and Mancinelli, R.L., 2001. Life in extreme environments. *Nature*, *409*(6823), pp.1092-1101.

Rytioja, J.; Hildén, K.; Yuzon, J.; Hatakka, A.; de Vries, R.P.; Mäkelä, M.R. Plant-polysaccharide-degrading enzymes from basidiomycetes. *Microbiol. Mol. Biol. Rev.* **2014**, *78*, 614–649.

Saha, B.C., 2003. Hemicellulose bioconversion. *Journal of industrial microbiology and biotechnology*, *30*(5), pp.279-291.

Sahdev, S., Khattar, S.K. and Saini, K.S., 2008. Production of active eukaryotic proteins through bacterial expression systems: a review of the existing biotechnology strategies. *Molecular and cellular biochemistry*, *307*(1), pp.249-264.

Sano, S., Ueda, M., Kitajima, S., Takeda, T., Shigeoka, S., Kurano, N., Miyachi, S., Miyake, C. and Yokota, A., 2001. Characterization of ascorbate peroxidases from unicellular red alga Galdieria partita. *Plant and Cell Physiology*, *42*(4), pp.433-440.

Sarkar, N., Ghosh, S.K., Bannerjee, S. and Aikat, K., 2012. Bioethanol production from agricultural wastes: an overview. *Renewable energy*, *37*(1), pp.19-27.

Satari, B., Karimi, K. and Kumar, R., 2019. Cellulose solvent-based pretreatment for enhanced second-generation biofuel production: a review. *Sustainable energy & fuels*, *3*(1), pp.11-62.

Saqib, S., Akram, A., Halim, S.A. and Tassaduq, R., 2017. Sources of β-galactosidase and its applications in food industry. *3 Biotech*, *7*(1), p.79.

Schenk, G., Mitić, N., Hanson, G.R. and Comba, P., 2013. Purple acid phosphatase: A journey into the function and mechanism of a colorful enzyme. *Coordination Chemistry Reviews*, *257*(2), pp.473-482.

Scherhag, P. and Ackermann, J.U., 2021. Removal of sugars in wastewater from food production through heterotrophic growth of Galdieria sulphuraria. *Engineering in Life Sciences*, *21*(3-4), pp.233-241.

Schmidt, R.A., Wiebe, M.G. and Eriksen, N.T., 2005. Heterotrophic high cell-density fed-batch cultures of the phycocyanin-producing red alga Galdieria sulphuraria. *Biotechnology and bioengineering*, *90*(1), pp.77-84.

Seckbach, J. and Chapman, D.J. eds., 2010. *Red algae in the genomic age* (Vol. 13).

Seckbach, J. and Rampelotto, P.H., 2015. 8 Polyextremophiles. In *Microbial evolution under extreme conditions* (pp. 153-170). De Gruyter.

SENTSOVA, U., 1991. On the diversity of acido-thermophilic unicellular algae of the genus Galdieria (Rhodophyta, Cyanidiophyceae). *Botaničeskij žurnal*, *76*(1), pp.69-78.

Serrano, A., Perez-Castineira, J.R., Baltscheffsky, H. and Baltscheffsky, M., 2004. Proton-pumping inorganic pyrophosphatases in some archaea and other extremophilic prokaryotes. *Journal of bioenergetics and biomembranes*, *36*(1), pp.127-133.

Setubal, J.C. and Stadler, P.F., 2018. Gene phylogenies and orthologous groups. In *Comparative genomics* (pp. 1-28). Humana Press, New York, NY.

Sharma, H.K., Xu, C. and Qin, W., 2019. Biological pretreatment of lignocellulosic biomass for biofuels and bioproducts: an overview. *Waste and Biomass Valorization*, *10*(2), pp.235-251.

Shirkavand, E., Baroutian, S., Gapes, D.J. and Young, B.R., 2016. Combination of fungal and physicochemical processes for lignocellulosic biomass pretreatment–A review. *Renewable and Sustainable Energy Reviews*, *54*, pp.217-234.

Shoseyov, O., Shani, Z. and Levy, I., 2006. Carbohydrate binding modules: biochemical properties and novel applications. *Microbiology and molecular biology reviews*, *70*(2), pp.283-295.

Shuter, B.J., Thomas, J.E., Taylor, W.D. and Zimmerman, A.M., 1983. Phenotypic correlates of genomic DNA content in unicellular eukaryotes and other cells. *The American Naturalist*, *122*(1), pp.26-44.

Sidar, A., Albuquerque, E.D., Voshol, G.P., Ram, A.F., Vijgenboom, E. and Punt, P.J., 2020. Carbohydrate binding modules: diversity of domain architecture in amylases and cellulases from filamentous microorganisms. *Frontiers in bioengineering and biotechnology*, *8*.

Siddiqui, K.S., Williams, T.J., Wilkins, D., Yau, S., Allen, M.A., Brown, M.V., Lauro, F.M. and Cavicchioli, R., 2013. Psychrophiles. *Annual Review of Earth and Planetary Sciences*, *41*, pp.87-115.

Simate, G.S. and Ndlovu, S., 2014. Acid mine drainage: Challenges and opportunities. *Journal of Environmental Chemical Engineering*, *2*(3), pp.1785-1803.

Sims, R.E., Mabee, W., Saddler, J.N. and Taylor, M., 2010. An overview of second generation biofuel technologies. *Bioresource technology*, *101*(6), pp.1570-1580.

Sindhu, R., Binod, P. and Pandey, A., 2016. Biological pretreatment of lignocellulosic biomass–An overview. *Bioresource technology*, *199*, pp.76-82.

Singh, A., Upadhyay, V., Upadhyay, A.K., Singh, S.M. and Panda, A.K., 2015. Protein recovery from inclusion bodies of Escherichia coli using mild solubilization process. *Microbial cell factories*, *14*(1), pp.1-10.

Singh, H., 2018. Desiccation and radiation stress tolerance in cyanobacteria. *Journal of basic microbiology*, *58*(10), pp.813-826

Singhvi, P., Saneja, A., Srichandan, S. and Panda, A.K., 2020. Bacterial inclusion bodies: a treasure trove of bioactive proteins. *Trends in biotechnology*, *38*(5), pp.474-486.

Skorupa, D.J., Reeb, V., Castenholz, R.W., Bhattacharya, D. and McDermott, T.R., 2013. Cyanidiales diversity in Yellowstone national park. *Letters in applied microbiology*, *57*(5), pp.459-466.

Sloth, J.K., Wiebe, M.G. and Eriksen, N.T., 2006. Accumulation of phycocyanin in heterotrophic and mixotrophic cultures of the acidophilic red alga Galdieria sulphuraria. *Enzyme and Microbial Technology*, *38*(1-2), pp.168-175.

Sloth, J.K., Jensen, H.C., Pleissner, D. and Eriksen, N.T., 2017. Growth and phycocyanin synthesis in the heterotrophic microalga Galdieria sulphuraria on substrates made of food waste from restaurants and bakeries. *Bioresource technology*, *238*, pp.296-305.

Smith, D.R., Hua, J., Lee, R.W. and Keeling, P.J., 2012. Relative rates of evolution among the three genetic compartments of the red alga Porphyra differ from those of green plants and do not correlate with genome architecture. *Molecular phylogenetics and evolution*, *65*(1), pp.339-344.

Soleymani, B., Barzegari, E., Mansouri, K., Karami, K., Mohammadi, P., Kiani, S., Moasefi, N., Tabar, M.S. and Mostafaie, A., 2020. Heterologous expression, purification, and refolding of SRY protein: role of l-arginine as analyzed by simulation and practical study. *Molecular biology reports*, *47*(8), pp.5943-5951.

Som, A., 2015. Causes, consequences and solutions of phylogenetic incongruence. *Briefings in Bioinformatics*, *16*(3), pp.536-548.

Sticklen, M., 2006. Plant genetic engineering to improve biomass characteristics for biofuels. *Current opinion in biotechnology*, *17*(3), pp.315-319.

Stone, B.A., Jacobs, A.K., Hrmova, M., Burton, R.A. and Fincher, G.B., 2018. Biosynthesis of plant cell wall and related polysaccharides by enzymes of the GT2 and GT48 families. *Annual Plant Reviews online*, pp.109-165.

Somers, M.D., Chen, P., Clippinger, J., Cruce, J.R., Davis, R., Lammers, P.J. and Quinn, J.C., 2021. Techno-economic and life-cycle assessment of fuel production from mixotrophic Galdieria sulphuraria microalgae on hydrolysate. *Algal Research*, *59*, p.102419.

Spielman, S.J. and Wilke, C.O., 2015. The relationship between dN/dS and scaled selection coefficients. *Molecular biology and evolution*, *32*(4), pp.1097-1108.

Strazzulli, A., Cobucci-Ponzano, B., Iacono, R., Giglio, R., Maurelli, L., Curci, N., Schiano-di-Cola, C., Santangelo, A., Contursi, P., Lombard, V. and Henrissat, B., 2020. Discovery of hyperstable carbohydrate-active enzymes through metagenomics of extreme environments. *The FEBS journal*, *287*(6), pp.1116-1137.

Studier FW & Moffatt BA (1986) Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. Journal of Molecular Biology 189(1):113-130. 218.

Suleiman, M., Krüger, A. and Antranikian, G., 2020. Biomass-degrading glycoside hydrolases of archaeal origin. *Biotechnology for biofuels*, *13*(1), pp.1-14.

Sun, S., Sun, S., Cao, X. and Sun, R., 2016. The role of pretreatment in improving the enzymatic hydrolysis of lignocellulosic materials. *Bioresource technology*, *199*, pp.49-58.

Sun, Y. and Cheng, J., 2002. Hydrolysis of lignocellulosic materials for ethanol production: a review. *Bioresource technology*, *83*(1), pp.1-11.

Suyama, M., Torrents, D. and Bork, P., 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic acids research*, *34*(suppl_2), pp.W609-W612.

Suzuki, K., Ehara, T., Osafune, T., Kuroiwa, H., Kawano, S. and Kuroiwa, T., 1994. Behavior of mitochondria, chloroplasts and their nuclei during the mitotic cycle in the ultramicroalga Cyanidioschyzon merolae. *European journal of cell biology*, *63*(2), pp.280-288.

Świątek, K., Gaag, S., Klier, A., Kruse, A., Sauer, J. and Steinbach, D., 2020. Acid hydrolysis of lignocellulosic biomass: Sugars and furfurals formation. *Catalysts*, *10*(4), p.437.

Takahara, M., Takahashi, H., Matsunaga, S., Sakai, A., Kawano, S. and Kuroiwa, T., 1999. Two types of ftsZ genes isolated from the unicellular primitive red alga Galdieria sulphuraria. *Plant and cell physiology*, *40*(8), pp.784-791.

Talebnia, F., Karakashev, D. and Angelidaki, I. (2010). Production of bioethanol from wheat straw: An overview on pretreatment, hydrolysis and fermentation. *Bioresource Technology*, 101(13), pp.4744-4753.

Telesh, I., Schubert, H. and Skarlato, S., 2013. Life in the salinity gradient: discovering mechanisms behind a new biodiversity pattern. *Estuarine, Coastal and Shelf Science*, *135*, pp.317-327.

Tian, S.Q., Zhao, R.Y. and Chen, Z.C., 2018. Review of the pretreatment and bioconversion of lignocellulosic biomass from wheat straw materials. *Renewable and Sustainable Energy Reviews*, *91*, pp.483-489.

Tilden, J.E., 1898. Observations on some west American thermal algae. *Botanical Gazette*, *25*(2), pp.89-105.

Toda, K., Takahashi, H., Itoh, R. and Kuroiwa, T., 1995. DNA contents of cell nuclei in two Cyanidiophyceae: Cyanidioschyzon merolae and Cyanidium caldarium Forma A. *Cytologia*, *60*(2), pp.183-188.

Toplin, J.A., Norris, T.B., Lehr, C.R., McDermott, T.R. and Castenholz, R., 2008. Biogeographic and phylogenetic diversity of thermoacidophilic cyanidiales in Yellowstone National Park, Japan, and New Zealand. *Applied and environmental microbiology*, *74*(9), pp.2822-2833.

Ulvskov, P., Paiva, D.S., Domozych, D. and Harholt, J., 2013. Classification, naming and evolutionary history of glycosyltransferases from sequenced green and red algal genomes. *PLoS One*, *8*(10), p.e76511.

Union, E., 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. *Official Journal of the European Union*, *5*, pp.82-209.

UniProt: the universal protein knowledgebase in 2021

United Nations, Department of Economic and Social Affairs, Population Division, 2019. World Population Prospects: The 2019 Revision (Medium variant)

Vallejo, L.F. and Rinas, U., 2004. Strategies for the recovery of active proteins through refolding of bacterial inclusion body proteins. *Microbial cell factories*, *3*(1), pp.1-12.

Varshney, P., Mikulic, P., Vonshak, A., Beardall, J. and Wangikar, P.P., 2015. Extremophilic micro-algae and their potential contribution in biotechnology. *Bioresource technology*, *184*, pp.363-372.

Vaser, R. and Sikic, M., 2021. Raven: a de novo genome assembler for long reads. *BioRxiv*, pp.2020-08.

Vinogradov, A.E. and Anatskaya, O.V., 2006. Genome size and metabolic intensity in tetrapods: a tale of two lines. *Proceedings of the Royal Society B: Biological Sciences*, *273*(1582), pp.27-32.

Vohra, M., Manwar, J., Manmode, R., Padgilwar, S. and Patil, S., 2014. Bioethanol production: feedstock and current technologies. Journal of Environmental Chemical Engineering, 2(1), pp.573-584.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K. and Earl, A.M., 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS one*, *9*(11), p.e112963.

Wall, D.P., Hirsh, A.E., Fraser, H.B., Kumm, J., Giaever, G., Eisen, M.B. and Feldman, M.W., 2005. Functional genomic analysis of the rates of protein evolution. *Proceedings of the National Academy of Sciences*, *102*(15), pp.5483-5488.

Wang, H., Squina, F., Segato, F., Mort, A., Lee, D., Pappan, K. and Prade, R., 2011. High-temperature enzymatic breakdown of cellulose. *Applied and environmental microbiology*, *77*(15), pp.5199-5206.

Wang, L., Littlewood, J. and Murphy, R.J., 2013. Environmental sustainability of bioethanol production from wheat strsaw in the UK. *Renewable and Sustainable Energy Reviews*, *28*, pp.715-725.

Wang, S., Dai, G., Yang, H. and Luo, Z., 2017. Lignocellulosic biomass pyrolysis mechanism: a state-of-the-art review. *Progress in energy and combustion science*, *62*, pp.33-86.

Wang, Y., Van Oosterwijk, N., Ali, A.M., Adawy, A., Anindya, A.L. and Dömling, A.S.S., *A Systematic Protein Refolding Screen Method using the DGR Approach Reveals that Time and Secondary TSA are Essential Variables. Sci Rep. 2017; 7 (1): 9355*. Epub 2017/08/24. doi: 10.1038/s41598-017-09687-z. PubMed PMID: 28839267.

Weber, K.A., Achenbach, L.A. and Coates, J.D., 2006. Microorganisms pumping iron: anaerobic microbial iron oxidation and reduction. *Nature Reviews Microbiology*, *4*(10), pp.752-764.

William Studier F, Rosenberg AH, Dunn JJ, & Dubendorff JW (1990) Use of T7 RNA polymerase to direct expression of cloned genes. Methods in Enzymology, ed David VG (Academic Press), Vol Volume 185, pp 60-89.

Woiciechowski, A.L., Neto, C.J.D., de Souza Vandenberghe, L.P., de Carvalho Neto, D.P., Sydney, A.C.N., Letti, L.A.J., Karp, S.G., Torres, L.A.Z. and Soccol, C.R., 2020. Lignocellulosic biomass: Acid and alkaline pretreatments and their effects on biomass recalcitrance–Conventional processing and recent advances. *Bioresource technology*, *304*, p.122848.

Xu, S., Wang, J., Guo, Z., He, Z. and Shi, S., 2020. Genomic convergence in the adaptation to extreme environments. *Plant communications*, p.100117.

Yampolsky, L., 2016. *Mutation and Genome Evolution*. Academic Press, pp.77-83.

Yang, A.S. and Honig, B., 1993. On the pH dependence of protein stability. *Journal of molecular biology*, *231*(2), pp.459-474.

Yang, Z., 2000. Phylogenetic analysis by maximum likelihood (PAML).

Yang, Z., Nielsen, R., Goldman, N. and Pedersen, A.M.K., 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics, 155(1), pp.431-449.

Yang, E.C., Boo, S.M., Bhattacharya, D., Saunders, G.W., Knoll, A.H., Fredericq, S., Graf, L. and Yoon, H.S., 2016. Divergence time estimates and the evolution of major lineages in the florideophyte red algae. Scientific reports, 6(1), pp.1-11.

Yang, J., Li, Q., Du, W., Yao, Y., Shen, G., Jiang, W. and Pang, Y., 2021. Genome-Wide Analysis of Glycoside Hydrolase Family 35 Genes and Their Potential Roles in Cell Wall Development in Medicago truncatula. Plants, 10(8), p.1639.

Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER Suite: Protein structure and function prediction. Nature Methods, 12: 7-8 (2015)

Yang J, Zhang Y. I-TASSER server: new development for protein structure and function predictions. Nucleic Acids Research, 43: W174-W181 (2015)

Yin, Y.; Mao, X.; Yang, J.C.; Chen, X.; Mao, F.; Xu, Y. dbCAN: A web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. 2012, 40, W445–W451.

Yoon, H.S., Müller, K.M., Sheath, R.G., Ott, F.D. and Bhattacharya, D., 2006. Defining the major lineages of red algae (RHODOPHYTA) 1. Journal of phycology, 42(2), pp.482-492.

Yoon, H.S., Ciniglia, C., Wu, M., Comeron, J.M., Pinto, G., Pollio, A. and Bhattacharya, D., 2006. Establishment of endolithic populations of extremophilic Cyanidiales (Rhodophyta). BMC Evolutionary Biology, 6(1), pp.1-12.

Zavrel, M., Markus, Z., Bartuch, J. and Verhuelsdonk, M. (2018) Process for the Hydrolysis of Biomass. Google Patents.

Zeldes, B.M., Keller, M.W., Loder, A.J., Straub, C.T., Adams, M.W. and Kelly, R.M., 2015. Extremely thermophilic microorganisms as metabolic engineering platforms for production of fuels and industrial chemicals. Frontiers in microbiology, 6, p.1209.

Zhao, Y., Yi, Z., Warren, A. and Song, W.B., 2018. Species delimitation for the molecular taxonomy and ecology of the widely distributed microbial eukaryote genus Euplotes (Alveolata, Ciliophora). Proceedings of the Royal Society B: Biological Sciences, 285(1871), p.20172159.

Zhao, Y., Norouzi, H., Azarderakhsh, M. and AghaKouchak, A., 2021. Global Patterns of Hottest, Coldest and Extreme Diurnal Variability on Earth. Bulletin of the American Meteorological Society, pp.1-23.

Zheng W, Zhang C, Li Y, Pearce R, Bell EW, Zhang W. Folding non-homology proteins by coupling deep-learning contact maps with I-TASSER assembly simulations. Cell Reports Methods, 1: 100014 (2021).

Zoghlami, A. and Paës, G., 2019. Lignocellulosic biomass: understanding recalcitrance and predicting hydrolysis. Frontiers in chemistry, 7, p.874.

Appendix

Supplementary Table 1: Collection information on all strains used in the phylogenetic analysis.

Taxa	Lineage	Strain	Sampling site (Country)	Habitat	pH	Temperature (°C)	Source (Reference)
Cyanidiophyceae							
Galdieria sulphuraria	1	138	San salvador (SV)	NA	NA	NA	ACUF (Gross et al., 2001)
	2	002	Piscarelli (IT)	Dry cripto-endolithic site	1	18-30	ACUF (Pinto et al., 2003)
		011	Caserta (IT)	Acidic rock	0.8	15	ACUF (Ciniglia et al., 2004)
		017	Solfatara (IT)	Fumarols	1	38	ACUF (Ciniglia et al., 2004)
		021	Vulcano Island (IT)	NA	NA	NA	ACUF (Ciniglia et al., 2004)
		638	Güglükonak (TR)	Thermal bath	1	54	ACUF (Iovinella et al., 2018)
		660	Güglükonak (TR)	Thermal bath	1	54	ACUF (Iovinella et al., 2018)
		PISC 6	Piscarelli (IT)	Dry cripto-endolithic site	1	18-30	This study
		RI 1	Rio tinto (ES)	Acidic periodic water flow	0.85-1.55	12.2-19.4	N/A (Aguilera et al., 2007)
		SOL 1	Solfatara (IT)	Fumarols	1	38	This study
		SOL 2	Solfatara (IT)	Fumarols	1	38	This study
		SOL 3	Solfatara (IT)	Fumarols	1	38	This study
	3	136	Mexicali (MX)	NA	NA	NA	ACUF (Gross et al., 2001)

Taxa	Lineage	Strain	Sampling site (Country)	Habitat	pH	Temperature (°C)	Source (Reference)
Galdieria sulphuraria	3	141	Yellostone National Park (US)	Acidic hot spring	NA	N/A	ACUF (Pinto et al., 2003)
		142	N/A (IS)	NA	NA	N/A	ACUF (Gross et al., 2001)
		1067	Azores (PT)	Porous sandstone, endolithic	2.1	N/A	CCALA (Gross et al., 2001)
		965	Soos (CZE)	Diatom field	0.8-2.0	<30	CCALA (Gross et al., 2002)
		5573	Yellostone National Park (US)	Acidic soil	1	55	CCMEE (Toplin et al., 2008)
		5610	Yellostone National Park (US)	Acidic crust	4	40	CCMEE (Toplin et al., 2008)
		5657	Owakudani (JP)	Acidic pool edge	2.5	>45	CCMEE (Toplin et al., 2008)
		5658	Owakudani (JP)	Acidic pool edge	2.5	>45	CCMEE (Toplin et al., 2008)
		5665	Kusatsu (JP)	Acidic pool	2	49	CCMEE (Toplin et al., 2008)
		5672	Owakudani (JP)	Acidic pool edge	3	42-55	CCMEE (Toplin et al., 2008)
		5712	Craters of the Moon (NZ)	Acid steam hole	NA	N/A	CCMEE (Toplin et al., 2008)
		5720	White Island (NZ)	Acidic stream	2.5-3.0	45	CCMEE (Toplin et al., 2008)
		P503	Kamchatka (RU)	NA	NA	N/A	IPPAS (Sentsova 1991)
		107.79	California (US)	Acidic hot water	1	70-75	SAG (Allen 1959)
	4	074W	Java (ID)	Fumarols	NA	35	ACUF (Pinto et al., 2003)
	5	21.92	Yangmingshan National Park (TW)	Hot spring	NA	N/A	SAG (Gross et al., 2001)
		033	GengZiPeng (TW)	Acidic Stream	2.6	45	THAL (Hsieh et al., 2015)
		054	DaYouKeng (TW)	Acidic pool	2.2	54	THAL (Hsieh et al., 2015)

Taxa	Lineage	Strain	Sampling site (Country)	Habitat	pH	Temperature (°C)	Source (Reference)
Galdieria sulphuraria	6	388	Landmannalaugar (IS)	Acid soil	1	NA	ACUF (Ciniglia et al., 2014)
		407	Niasjvellir (IS)	Acidic soil and stream	0.0-4.5	42-47	ACUF (Ciniglia et al., 2014)
		427	Gunnhuver (IS)	Fumarole, acidic soil and mud	0.0-1.0	31.2-47	ACUF (Ciniglia et al., 2014)
		455	Viti (IS)	Acidic soil and mud	1.0-1.5	25-29	ACUF (Ciniglia et al., 2014)
		P501	Kamchatka (RU)	NA	NA	25-40	IPPAS (Sentsova 1991)
Galdieria Phlegrea		009	Nepi (IT)	Sulphur spring	0.8	N/A	ACUF (Pinto et al., 2003)
		647	Çermik (TR)	Thermal bath	7	12	ACUF (Ciniglia et al., 2018)
		663	Güglükonak (TR)	Thermal bath	1	24.6	ACUF (Ciniglia et al., 2018)
		735	Biloris (TR)	Thermal bath	7	54	ACUF (Ciniglia et al., 2018)
		788	Diyadin (TR)	Hot spring, pool and soil	6.5	25.8	ACUF (Ciniglia et al., 2018)
		AG1	Rio tinto (ES)	Acidic stream	2.32-2.88	45	N/A (Aguilera et al., 2007)
		CEMI	Rio tinto (ES)	Acidic stream coming from	2.38-2.62	12.3-27.3	N/A (Aguilera et al., 2007)
Cyanidioschyzon merolae		10D	Sardinia (IT)	Acidic hot spring	1.5	12.4-22.6	ATCC (Kuroiwa et al., 1994)
Porphyra umbilicalis		LB 2951	Schoodic Point, Maine (US)	NA	NA	45	N/A (Brawley et al., 2017)
Pyropia haitanensis		PH-38	NA	NA	NA	N/A	N/A (Wang et al., 2013)

Supplementary Table 2: DNA extraction buffer compositions

DNA Extraction buffers	
Buffer 1.1	**Buffer 2.2**
200mM Tris-HCl pH8	100mM Tris-HCl pH8
200mM NaCl	700mM NaCl
100mM LiCl	20mM EDTA pH8
25mM EDTA pH8	2% CTAB
1M Urea	0.0125mM PVP-40
1% SDS	
1% NP-40	

Supplementary Table 3: Sequencing and assembly statistics of all strains used in the phylogenetic analysis. Strains 017, 074 read files are missing due to being sequenced pre-the beginning of this project and unavailable.

Sample	Read pairs	Read 1 bases	Read 2 bases	Contigs	Assembly size (bp)	Contig N50
ACUF 002	1,412,780	320,297,288	290,433,295	11,989	15,427,022	48,215
ACUF 009	6,235,759	1,434,113,440	1,340,425,250	5,371	13,825,730	106,076
ACUF 011	2,715,009	590,009,367	583,434,258	11,370	14,840,803	25,766
ACUF 017	NA	NA	NA	10,036	17,520,760	63,170
ACUF 021	4,624,837	1,088,890,939	1,030,169,327	10,744	15,778,202	41,375
THAL 054	22,398,586	3,382,186,486	3,382,186,486	28,507	20,754,823	2,829
ACUF 074	NA	NA	NA	8,299	16,157,023	106,469
CCALA 1067	2,962,566	694,359,924	639,222,999	9,613	14,577,605	44,288
SAG 107.79	2,607,426	783,421,676	783,146,529	8,341	19,210,403	4,251
ACUF 136	2,468,554	524,982,332	486,784,536	13,675	15,846,950	60,582
ACUF 138	2,147,163	501,547,923	456,401,363	13,731	16,283,394	56,051
ACUF 141	4,104,615	976,696,898	916,953,428	17,137	17,422,955	48,462
ACUF 142	9,757,200	1,473,337,200	1,473,337,200	16,441	17,281,166	49,142
SAG 21.92	31,500,159	4,756,524,009	4,756,524,009	45,415	30,039,158	2,241
ACUF 735	3,150,564	746,578,122	693,626,302	9,355	26,495,857	36,861
ACUF 388	3,471,734	826,672,299	778,213,835	8,408	13,630,313	123,414
ACUF 402	4,107,600	976,578,967	934,083,497	7,697	13,443,952	125,296
ACUF 427	4,319,515	1,026,830,889	974,283,076	6,003	13,036,834	127,172
ACUF 660	4,133,424	926,184,421	858,689,246	17,133	17,063,585	16,261
ACUF 455	3,551,648	847,039,146	798,416,986	6,431	13,104,113	129,872
CCMEE 5573	2,088,344	483,787,198	413,957,014	5,692	13,269,278	67,360

CCMEE 5610	34,831,048	7,999,192,478	7,423,602,002	7,475	13,314,561	76,570
CCMEE 5657	2,049,799	487,223,023	477,425,511	18,459	14,957,230	29,002
CCMEE 5658	25,757,583	3,889,395,033	3,889,395,033	32,479	21,639,394	1,432
CCMEE 5665	15,570,432	2,351,135,232	2,351,135,232	28,737	20,676,297	1,755
CCMEE 5672	12,743,681	1,924,295,831	1,924,295,831	17,288	16,315,258	57,845
CCMEE 5712	6,520,585	984,608,335	984,608,335	24,061	19,795,830	33,864
CCMEE 5720	18,977,191	2,865,555,841	2,865,555,841	23,864	19,367,444	16,107
ACUF 638	12,693,309	1,916,689,659	1,916,689,659	16,976	16,239,889	90,061
ACUF 647	1,807,928	405,213,286	398,695,333	10,072	14,230,489	59,457
ACUF 663	2,010,294	450,314,554	406,579,114	11,416	26,049,968	15,610
ACUF 788	2,404,689	580,077,820	563,151,772	15,730	27,141,156	23,398
CCALA 965	2,817,420	651,308,242	599,834,789	9,622	14,537,236	38,530
AG1	4,338,286	700,929,312	696,920,952	15,741	15,384,994	39,997
CEMI_1	1,491,661	330,202,698	322,151,248	8,951	13,985,505	84,140
p501	13,717,008	2,071,268,208	2,071,268,208	36,090	23,499,960	1,286
p503	11,502,092	1,736,815,892	1,736,815,892	24,631	19,336,870	15,406
PISC6	1,593,260	361,819,502	325,010,524	8,151	14,704,879	66,768
RI1	3,168,556	614,680,656	611,440,727	10,807	15,711,356	12,917
SOL_2	1,816,332	388,383,510	375,899,780	15,743	15,315,393	23,366
SOL_3	1,308,307	290,654,397	285,885,638	11,757	14,603,277	40,019
SOL_1	1,067,344	217,024,007	200,869,967	5,449	13,489,107	91,627
THAL033	6,517,103	984,082,553	984,082,553	18,161	17,127,862	85,154

Supplementary Table 4: Information on assembly and source for outgroups used in phylogenetic analysis.

Species	median total length (Mb)	median GC %	Source (Reference)
Cyanidioschyzon merolae	16.429	54.7229	Matsuzaki et al., 2004
Porphyra umbilicalis	87.889	65.7288	Brawley et al., 2017
Pyropia haitanensis	53.2546	67.8	Cao et al., 2019

Supplementary Table 5: Genes tested under positive selection with, protein names, gene name, Test statistic value, E.C number, Length, Mass, Gene ontology and predicted signal peptide presence.

Protein names	Gene	Test	EC	Length	Mass	Gene ontology (GO)	Signal
Ribonuclease P (EC 3.1.26.5)	Gasu_00210	21.37	3.1.26.5	624	72,678	nucleolar ribonuclease P complex [GO:0005655]; ribonuclease MRP complex	
4-methyl-5(B-hydroxyethyl)-thiazole	Gasu_00530	44.39		277	29,616		
O-acyltransferase	Gasu_00890	27.37		464	54,802	endoplasmic reticulum membrane [GO:0005789]; integral component of membrane	
Metallo-beta-lactamase family protein	Gasu_01370	15.77		299	33,954		
Uncharacterized protein	Gasu_01470	13.82		297	33,458		
Fructokinase (EC 2.7.1.4)	Gasu_01610	113.7	2.7.1.4	299	32,054	fructokinase activity [GO:0008865]	
Bifunctional enzyme involved in thiolation and	Gasu_01730	17.32	3.6.1.27	594	67,609	4 iron, 4 sulfur cluster binding [GO:0051539]; metal ion binding [GO:0046872];	
CBF domain-containing protein	Gasu_02050	17		497	57,322	ribosome biogenesis [GO:0042254]	
Long-chain acyl-CoA synthetase (EC 6.2.1.3)	Gasu_02100	15.72	6.2.1.3	553	60,917	long-chain fatty acid-CoA ligase activity [GO:0004467]	
Inositol-1,3,4-trisphosphate 5/6-kinase (EC	Gasu_02310	95.11	2.7.1.159	475	53,985	integral component of membrane [GO:0016021]; ATP binding [GO:0005524];	
Protein transporter	Gasu_02720	14.22		1025	118,541	small GTPase binding [GO:0031267]; intracellular protein transport [GO:0006886]	
Methyltransferase	Gasu_02830	14.88		392	46,861	THO complex part of transcription export complex [GO:0000445];	
CLP1_P domain-containing protein	Gasu_02980	21.85		539	61,522	ATP binding [GO:0005524]	
Imidazole glycerol-phosphate synthase (EC	Gasu_03320	25.09	4.3.2.10	334	35,985	chloroplast [GO:0009507]; imidazoleglycerol-phosphate synthase activity	
Alanine--tRNA ligase (EC 6.1.1.7) (Alanyl-tRNA	Gasu_03470	18.66	6.1.1.7	961	107,907	mitochondrion [GO:0005739]; alanine-tRNA ligase activity [GO:0004813]; ATP	
Transcription initiation factor TFIID subunit D1	Gasu_03630	31.64		1401	161,239	nucleus [GO:0005634]; translation initiation factor activity [GO:0003743]	
MYB-related protein	Gasu_03640	16.09		251	28,589	DNA binding [GO:0003677]; transcription coactivator activity [GO:0003713]	
Uncharacterized protein	Gasu_03670	27.23		789	92,373	nucleus [GO:0005634]; ribosome biogenesis [GO:0042254]	
ATP-dependent Clp protease ATP-binding	Gasu_03700	31.57		922	102,620	chloroplast [GO:0009507]; ATP binding [GO:0005524]; peptidase activity	
Uncharacterized protein	Gasu_04170	57.03		975	111,207	integral component of membrane [GO:0016021]	
Uncharacterized protein	Gasu_04370	40.87		608	68,280	integral component of membrane [GO:0016021]	
3-isopropylmalate dehydratase (EC 4.2.1.33)	Gasu_04390	24.76	4.2.1.33	666	74,120	chloroplast stroma [GO:0009570]; cytoplasmic vesicle [GO:0031410]; integral	
Tyrosine aminotransferase (EC 2.6.1.5)	Gasu_04500	114.3	2.6.1.5	418	46,026	L-tyrosine:2-oxoglutarate aminotransferase activity [GO:0004838]; pyridoxal	
Preprotein translocase, Oxa1 family	Gasu_04560	44.46		357	40,240	integral component of membrane [GO:0016021]; membrane insertase activity	
tRNA modification GTPase isoform 1	Gasu_04710	59.36		552	62,924	GTP binding [GO:0005525]; GTPase activity [GO:0003924]; tRNA modification	Yes
Elongation factor G, mitochondrial (EF-Gmt)	Gasu_04930	37.7		755	85,483	chloroplast [GO:0009507]; mitochondrion [GO:0005739]; GTP binding	
Uncharacterized protein	Gasu_04940	17.19		621	70,676	phosphatidylinositol binding [GO:0035091]; ubiquitin binding [GO:0043130]	
Nipped-B_C domain-containing protein	Gasu_05250	20.54		1351	153,886		
Transcription factor, zinc ion binding protein	Gasu_05270	21.92		620	69,697	DNA-binding transcription factor activity, RNA polymerase II-specific	
Uncharacterized protein	Gasu_05480	133.1		689	79,744		
Transmembrane 9 superfamily member	Gasu_05800	36.18		627	71,646	integral component of membrane [GO:0016021]	Yes
Uncharacterized protein	Gasu_06260	20.2		447	49,760		
Threonyl-tRNA synthetase (EC 6.1.1.3)	Gasu_06310	26.5	6.1.1.3	705	82,214	cytoplasm [GO:0005737]; ATP binding [GO:0005524]; threonine-tRNA ligase	
Lipid-A-disaccharide synthase (EC 2.4.1.182)	Gasu_06440	16.21	2.4.1.182	501	56,243	integral component of membrane [GO:0016021]; lipid-A-disaccharide synthase	
Chloroplast inner membrane import protein	Gasu_06510	14.82		322	36,073	chloroplast [GO:0009507]; protein transport [GO:0015031]	
MFS transporter, PHS family, inorganic	Gasu_06570	25.86		621	69,913	integral component of membrane [GO:0016021]; transmembrane transporter	
Metal ion (Mn2+-iron) transporter, Nramp family	Gasu_06870	16.06		486	53,847	integral component of membrane [GO:0016021]; metal ion transmembrane	
Uncharacterized protein	Gasu_07740	14.28		770	91,140		
Uncharacterized protein	Gasu_07920	17.34		543	62,895		
tRNA-dihydrouridine synthase (EC 1.3.1.-)	Gasu_08250	43.7	1.3.1.-	314	35,750	chloroplast [GO:0009507]; flavin adenine dinucleotide binding [GO:0050660]; tRNA	
Solute carrier, DMT family	Gasu_08360	32.45		324	35,940	integral component of membrane [GO:0016021]	
Ethanolaminephosphotransferase (EC 2.7.8.1)	Gasu_08520	62.4	2.7.8.1	1063	122,114	integral component of membrane [GO:0016021]; transcription factor TFIIA complex	
tRNA-dihydrouridine synthase 3	Gasu_08660	14.88		508	58,304	chloroplast [GO:0009507]; flavin adenine dinucleotide binding [GO:0050660]; tRNA	

Cryptochrome, DASH family	Gasu_08770	143.4		570	66,359		
Glutamate N-acetyltransferase (EC 2.3.1.35)	Gasu_08880	16.19	2.3.1.35	636	71,581	glutamate N-acetyltransferase activity [GO:0004358]	
Uncharacterized protein	Gasu_09740	70.64		766	87,716	lipid binding [GO:0008289]; lipid transport [GO:0006869]	
Uncharacterized protein	Gasu_09850	15.12		261	25,746		Yes
FANCI_S4 domain-containing protein	Gasu_10660	25.56		1374	160,147	DNA repair [GO:0006281]	
FHA domain-containing protein	Gasu_10720	17.96		451	51,065	Mre11 complex [GO:0030870]; double-strand break repair [GO:0006302]; mitotic	
Phenazine biosynthesis PhzC/PhzF family	Gasu_11300	16.25		248	27,584	catalytic activity [GO:0003824]; biosynthetic process [GO:0009058]	
PRP4 pre-mRNA processing factor 4-like protein	Gasu_11920	61.49		520	59,391		
DUF1995 domain-containing protein	Gasu_12130	29.14		273	31,928		
Uncharacterized protein	Gasu_12890	31.85		707	81,558	nucleus [GO:0005634]	
Mannosyltransferase (EC 2.4.1.-)	Gasu_13520	21.96	2.4.1.-	637	72,603	endoplasmic reticulum membrane [GO:0005789]; integral component of membrane	
Uncharacterized protein	Gasu_13680	43.8		833	96,370	integral component of membrane [GO:0016021]	
Uncharacterized protein	Gasu_14440	27.04		221	25,803	integral component of membrane [GO:0016021]	
Exportin 1 (Xpo1)	Gasu_14650	57.81		1098	127,005	nucleus [GO:0005634]; nuclear export signal receptor activity [GO:0005049]; small	
5-amino-6-(5-phosphoribosylamino)uracil	Gasu_14780	39.54	1.1.1.193;	465	51,410	5-amino-6-(5-phosphoribosylamino)uracil reductase activity [GO:0008703];	
R3H-assoc domain-containing protein	Gasu_14850	29.89		232	26,499		
eIF-2B GDP-GTP exchange factor subunit alpha	Gasu_15580	28.83		322	35,747	translation initiation factor activity [GO:0003743]	
Uncharacterized protein	Gasu_15590	19.3		322	34,617	integral component of membrane [GO:0016021]	
3-hydroxyisobutyrate dehydrogenase (EC	Gasu_15760	15.19	1.1.1.31	351	37,412	cytosol [GO:0005829]; 3-hydroxyisobutyrate dehydrogenase activity	Yes
Replication factor C subunit 1	Gasu_15980	26.32		758	84,646	chloroplast [GO:0009507]; DNA replication factor C complex [GO:0005663]; ATP	
Eukaryotic translation initiation factor 3 subunit	Gasu_16220	18.97		543	62,407	eukaryotic 43S preinitiation complex [GO:0016282]; eukaryotic 48S preinitiation	
GTP-binding protein HflX	Gasu_16470	190.9		453	50,639	GTP binding [GO:0005525]; metal ion binding [GO:0046872]	
Uncharacterized protein	Gasu_16510	105.3		713	82,477	Golgi apparatus [GO:0005794]; protein transport [GO:0015031]	
Vacuolar protein sorting-associated protein 54	Gasu_17570	17.55		939	107,981	cytosol [GO:0005829]; GARP complex [GO:0000938]; protein transport	
Ubiquinol-cytochrome c reductase cytochrome	Gasu_17710	17.15	1.10.2.2	325	36,507	integral component of membrane [GO:0016021]; mitochondrial inner membrane	
BAR protein	Gasu_17730	14.79		303	34,746		
Peroxisomal membrane MPV17/PMP22-like	Gasu_17910	31.03		289	32,178	integral component of membrane [GO:0016021]	
Sucrose transporter, GPH family isoform 1	Gasu_18190	53.35		471	51,308	integral component of membrane [GO:0016021]; transmembrane transporter	
POT1-like telomere end-binding protein	Gasu_18520	85.28		566	64,317	chromosome, telomeric region [GO:0000781]; single-stranded telomeric DNA	
Monovalent cation:H+ antiporter-1, CPA1 family	Gasu_18700	13.93		569	63,385	integral component of membrane [GO:0016021]; solute:proton antiporter activity	
Methionine aminopeptidase 2 (MAP 2) (MetAP	Gasu_18760	14.75	3.4.11.18	412	46,752	cytoplasm [GO:0005737]; metal ion binding [GO:0046872]; metalloaminopeptidase	
Cysteine desulfurase (EC 2.8.1.7)	Gasu_18770	16.2	2.8.1.7	477	52,959	cysteine desulfurase activity [GO:0031071]; lyase activity [GO:0016829]; pyridoxal	
DNA mismatch repair protein MutS2	Gasu_18840	28.95		902	103,315	ATP binding [GO:0005524]; endonuclease activity [GO:0004519]; mismatched	
Calcium-transporting ATPase (EC 7.2.2.10)	Gasu_18920	22.17	7.2.2.10	1089	120,265	integral component of membrane [GO:0016021]; ATP binding [GO:0005524];	
Probable ATP-dependent transporter ycf16	Gasu_19480	28.64		798	91,614	chloroplast [GO:0009507]; ATP binding [GO:0005524]; ATPase-coupled	
DNA helicase (EC 3.6.4.12)	Gasu_19890	15.79	3.6.4.12	807	92,712	ATP binding [GO:0005524]; ATP hydrolysis activity [GO:0140603]; DNA binding	
Tyrosine--tRNA ligase (EC 6.1.1.1) (Tyrosyl-	Gasu_20480	25.66	6.1.1.1	501	56,384	chloroplast stroma [GO:0009570]; mitochondrion [GO:0005739]; ATP binding	
Deoxyribodipyrimidine photo-lyase /	Gasu_20740	27.09		1042	120,253	lyase activity [GO:0016829]; protein-chromophore linkage [GO:0018298]	
Pyrophosphate--fructose 6-phosphate 1-	Gasu_20900	81.39	2.7.1.90	569	63,167	cytoplasm [GO:0005737]; 6-phosphofructokinase activity [GO:0003872]; ATP	
ATP-dependent Lon protease (EC 3.4.21.53)	Gasu_21000	19.61	3.4.21.53	1229	138,512	ATP binding [GO:0005524]; ATP-dependent peptidase activity [GO:0004176];	
MFS transporter, SP family, sugar:H+ symporter	Gasu_21540	21.55		541	59,929	integral component of membrane [GO:0016021]; transmembrane transporter	
Cell division cycle 2, cofactor of APC complex	Gasu_21580	88.24		490	55,754	anaphase-promoting complex binding [GO:0010997]; ubiquitin-protein transferase	
Proteasome subunit alpha type	Gasu_21750	21.96		252	28,183	cytoplasm [GO:0005737]; nucleus [GO:0005634]; proteasome core complex,	
AAA-type ATPase	Gasu_21910	20.4		426	47,530	chloroplast [GO:0009507]; proteasome complex [GO:0000502]; ATP binding	
Succinate--CoA ligase [ADP-forming] subunit	Gasu_22120	20.88	6.2.1.5	436	47,451	mitochondrion [GO:0005739]; ATP binding [GO:0005524]; magnesium ion binding	
Glucose inhibited division protein A	Gasu_22140	17.78		718	81,193	flavin adenine dinucleotide binding [GO:0050660]; tRNA wobble uridine	
KH domain-containing protein	Gasu_22220	53.16		481	53,176	RNA binding [GO:0003723]	
Peroxiredoxin (Alkyl hydroperoxide reductase	Gasu_23160	30.85	1.11.1.15	223	24,283	peroxidase activity [GO:0004601]	
Bifunctional enzyme involved in thiolation and	Gasu_23290	28.73	3.6.1.27	1085	123,843	4 iron, 4 sulfur cluster binding [GO:0051539]; ATP binding [GO:0005524]; metal ion	

Glycerol-3-phosphate dehydrogenase [NAD(+)]	Gasu_23710	116.8	1.1.1.8	415	44,955	glycerol-3-phosphate dehydrogenase complex [GO:0009331]; glycerol-3-	
Repressor/activator protein 1 homolog	Gasu_23740	17.24		417	47,741	chromosome, telomeric region [GO:0000781]; nucleus [GO:0005634]; telomere	
C3H1-type domain-containing protein	Gasu_23770	50.21		791	87,831	integral component of membrane [GO:0016021]; metal ion binding [GO:0046872]	
Prolyl-tRNA synthetase (EC 6.1.1.15)	Gasu_24190	27.71	6.1.1.15	706	79,513	cytoplasm [GO:0005737]; integral component of membrane [GO:0016021]; ATP	
Pre-mRNA-processing factor 8	Gasu_24380	17.5		2364	275,689	spliceosomal complex [GO:0005681]; isopeptidase activity [GO:0070122];	
Transducin family protein / WD-40 repeat family	Gasu_24960	24.85		478	55,032		
Cytosolic Fe-S cluster assembly factor NUBP1	Gasu_25140	15.38		318	34,191	cytoplasm [GO:0005737]; 4 iron, 4 sulfur cluster binding [GO:0051539]; ATP	
Transcription-repair coupling factor (Superfamily	Gasu_25190	16.06		833	95,626	ATP binding [GO:0005524]; helicase activity [GO:0004386]; nucleic acid binding	
Uncharacterized protein	Gasu_25390	14.04		705	80,392		
Molecular chaperone DnaJ	Gasu_25590	17.82		476	52,169	ATP binding [GO:0005524]; heat shock protein binding [GO:0031072]; metal ion	
Hemolysin-related protein	Gasu_25820	14.84		610	69,277	integral component of membrane [GO:0016021]; flavin adenine dinucleotide	
Uncharacterized protein	Gasu_25870	45.74		148	17,487		
Phosphatidate cytidylyltransferase (EC 2.7.7.41)	Gasu_26040	58.39	2.7.7.41	499	57,196	integral component of membrane [GO:0016021]; phosphatidate	
Protein transport protein SEC23	Gasu_26230	233.2		768	86,226	COPII vesicle coat [GO:0030127]; endoplasmic reticulum membrane	
Protein RFT1 homolog	Gasu_26290	16.96		498	57,713	integral component of membrane [GO:0016021]; dolichol-linked oligosaccharide	
Alpha-1,4 glucan phosphorylase (EC 2.4.1.1)	Gasu_26320	21.09	2.4.1.1	887	100,979	glycogen phosphorylase activity [GO:0008184]; linear malto-oligosaccharide	
Cytochrome c-type biogenesis protein CcmE	Gasu_26480	27.26		293	33,523	integral component of membrane [GO:0016021]; plasma membrane	
Transcription initiation factor TFIIB	Gasu_26550	70.3		332	36,948	metal ion binding [GO:0046872]; TBP-class protein binding [GO:0017025];	
AAA-type ATPase	Gasu_26790	14.36		848	93,639	chloroplast [GO:0009507]; membrane [GO:0016020]; ATP binding [GO:0005524];	
U3 small nucleolar RNA-associated protein 7	Gasu_26820	19.68		558	64,053	nucleolus [GO:0005730]	
Uncharacterized protein	Gasu_26920	18.92		1751	197,655		
5' nucleotidase family protein	Gasu_27140	18.97		598	70,792	hydrolase activity [GO:0016787]; metal ion binding [GO:0046872]	
Salicylate hydroxylase (EC 1.14.13.1)	Gasu_27320	25.55	1.14.13.1	408	45,824	FAD binding [GO:0071949]; salicylate 1-monooxygenase activity [GO:0018658]	
Beta-galactosidase (EC 3.2.1.23)	Gasu_27500	23.43	3.2.1.23	1038	118,205	beta-galactosidase activity [GO:0004565]; carbohydrate metabolic process	Yes
Assimilatory sulfite reductase (ferredoxin) (EC	Gasu_27710	34.33	1.8.7.1	700	79,674	chloroplast envelope [GO:0009941]; chloroplast nucleoid [GO:0042644]; 4 iron, 4	
N-acetyltransferase	Gasu_27730	16.32		314	36,410	N-acetyltransferase activity [GO:0008080]	
Aquaglyceroporin related protein, MIP family	Gasu_28080	19.76		357	39,630	integral component of membrane [GO:0016021]; channel activity [GO:0015267]	
Chromatin remodeling complex / DNA-dep	Gasu_28100	25.68		924	105,912	ATP binding [GO:0005524]; nucleosome-dependent ATPase activity [GO:0070615]	
MYB domain transcription factor family	Gasu_28180	15.6		676	77,328	nucleus [GO:0005634]; transcription repressor complex [GO:0017053];	
Nuclear pore complex protein	Gasu_28270	176.8		1093	126,581	nuclear membrane [GO:0031965]; nuclear pore outer ring [GO:0031080]; structural	
Hexosyltransferase (EC 2.4.1.-)	Gasu_28490	14.17	2.4.1.-	465	54,963	Golgi membrane [GO:0000139]; integral component of membrane [GO:0016021];	
Serine/threonine protein kinase	Gasu_28780	20.74		497	56,815	ATP binding [GO:0005524]; protein serine/threonine kinase activity [GO:0004674]	
UDP-glucose/GDP-mannose dehydrogenase	Gasu_28830	42.39		438	48,429	NAD binding [GO:0051287]; oxidoreductase activity, acting on the CH-CH group of	
Uncharacterized protein	Gasu_28850	52.45		457	53,553	nucleus [GO:0005634]; transcription, DNA-templated [GO:0006351]	
RNA-binding protein	Gasu_28880	14.99		187	21,569	RNA binding [GO:0003723]	
2-oxoisovalerate dehydrogenase E1	Gasu_28890	71.51	1.2.4.4	366	40,513	3-methyl-2-oxobutanoate dehydrogenase (2-methylpropanoyl-transferring) activity	
Translation initiation factor eIF-4F	Gasu_30270	18.88		1548	172,936	translation initiation factor activity [GO:0003743]	
Microfibrillar-associated protein	Gasu_30380	46.73		442	52,362		
Uncharacterized protein	Gasu_30520	82.27		466	53,713	integral component of membrane [GO:0016021]	
Ubiquitin-protein ligase E3 (EC 6.3.2.19)	Gasu_30800	26.03	6.3.2.19	1392	157,054	ligase activity [GO:0016874]; ubiquitin-protein transferase activity [GO:0004842]	
Uncharacterized protein	Gasu_30940	24.71		1005	113,273		
Transcription initiation factor IIE subunit alpha	Gasu_31230	25.25		364	42,051	translation initiation factor activity [GO:0003743]; transcription initiation from RNA	
Pentatricopeptide (PPR) repeat-containing	Gasu_31260	23.58		715	81,696		
Ribosome production factor 2 homolog	Gasu_31320	17.01		299	34,789	nucleolus [GO:0005730]; rRNA binding [GO:0019843]; maturation of LSU-rRNA	
Peptidyl-prolyl cis-trans isomerase (PPIase) (EC	Gasu_31360	25.52	5.2.1.8	163	18,103	peptidyl-prolyl cis-trans isomerase activity [GO:0003755]; protein folding	
Cyclin-dependent serine/threonine protein	Gasu_31860	31.9	2.7.11.22	401	46,355	ATP binding [GO:0005524]; cyclin-dependent protein serine/threonine kinase	
Smad nuclear interacting protein 1 isoform 1	Gasu_32070	14.7		290	33,795	RNA binding [GO:0003723]	
Vesicle transport V-snare protein	Gasu_32370	21.04		171	19,844	integral component of membrane [GO:0016021]; protein transport [GO:0015031];	
Signal peptidase complex subunit 2 (EC 3.4.-.-)	Gasu_32430	15.64	3.4.-.-	179	20,863	integral component of membrane [GO:0016021]; signal peptidase complex	

15-cis-phytoene synthase (EC 2.5.1.32)	Gasu_32640	159.7	2.5.1.32	453	53,261	farnesyltranstransferase activity [GO:0004311]; geranylgeranyl-diphosphate	
Phosphatidylinositol-bisphosphatase (EC	Gasu_32740	293.8	3.1.3.36	825	94,074	early endosome membrane [GO:0031901]; phagocytic vesicle membrane	
Uncharacterized protein	Gasu_32980	31.72		559	65,295		
Cytochrome c biogenesis protein, putative, Ccb2	Gasu_32990	22.84		285	31,891		
DNA binding protein	Gasu_33200	15.5		411	45,701		
Dihydrolipoamide acetyltransferase component	Gasu_33530	18.82	2.3.1.-	600	64,128	acyltransferase activity [GO:0016746]	
5'-3' exoribonuclease	Gasu_34120	20.65		571	67,807	exonuclease activity [GO:0004527]; nucleic acid binding [GO:0003676]	
40S ribosomal protein S6	Gasu_34220	18.39		237	27,166	ribosome [GO:0005840]; structural constituent of ribosome [GO:0003735];	
NAD-dependent epimerase/dehydratase	Gasu_34330	49.34		396	44,971		
Stress-induced-phosphoprotein 1	Gasu_34500	39.06		571	64,975		
Protein disulfide-isomerase A4 (EC 5.3.4.1)	Gasu_34670	80.04	5.3.4.1	386	43,955	endoplasmic reticulum [GO:0005783]; protein disulfide isomerase activity	Yes
DUF155 domain-containing protein	Gasu_34990	21.59		367	42,097	integral component of membrane [GO:0016021]	
Rhomboid domain-containing protein	Gasu_35020	37.12		196	23,102	integral component of membrane [GO:0016021]; serine-type endopeptidase	
Chaperone protein DnaJ	Gasu_35080	58.5		276	32,682		
UDP-sulfoquinovose synthase (EC 3.13.1.1)	Gasu_35090	46.47	3.13.1.1	601	68,019	UDPsulfoquinovose synthase activity [GO:0046507]	
Dolichyl-diphosphooligosaccharide--protein	Gasu_35460	23.36	2.4.99.18	714	80,918	endoplasmic reticulum [GO:0005783]; integral component of membrane	
Transducin family protein / WD-40 repeat family	Gasu_35820	70.5		474	53,851		
C3H1-type domain-containing protein	Gasu_35940	16.29		819	90,906	metal ion binding [GO:0046872]	
Coatomer subunit gamma	Gasu_36000	34.04		927	104,804	COPI vesicle coat [GO:0030126]; Golgi membrane [GO:0000139]; structural	
N-acetylglucosaminyldiphosphodolichol N-	Gasu_36100	74.26	2.4.1.141	165	18,600	N-acetylglucosaminyldiphosphodolichol N-acetylglucosaminyltransferase activity	
Uncharacterized protein	Gasu_36360	34.13		486	55,075	cytoplasm [GO:0005737]; 4 iron, 4 sulfur cluster binding [GO:0051539]; metal ion	
Uncharacterized protein	Gasu_36500	16.76		484	58,557		
DIOX_N domain-containing protein	Gasu_36940	20.3		338	38,745		
Uncharacterized protein	Gasu_37020	33.41		1365	156,578	integral component of membrane [GO:0016021]	
Carboxyl-terminal processing protease isoform 1	Gasu_37410	14.36	3.4.21.102	633	71,565	serine-type endopeptidase activity [GO:0004252]	
AAA-type ATPase	Gasu_37490	14.37		406	45,820	chloroplast [GO:0009507]; proteasome complex [GO:0000502]; ATP binding	
Uncharacterized protein	Gasu_38100	14.77		722	83,594		
ATP-dependent DNA helicase (EC 3.6.4.12)	Gasu_38190	18.03	3.6.4.12	529	60,483	nucleus [GO:0005634]; ATP binding [GO:0005524]; ATP hydrolysis activity	
Serine/threonine-protein phosphatase (EC	Gasu_38260	14.64	3.1.3.16	306	34,713	protein serine phosphatase activity [GO:0106306]; protein threonine phosphatase	
14-3-3 protein-like protein	Gasu_38270	82.47		264	30,101		
Probable cytosolic iron-sulfur protein assembly	Gasu_39260	15.99		392	44,318	CIA complex [GO:0097361]; iron-sulfur cluster assembly [GO:0016226]	
Molecular chaperone DnaJ	Gasu_39280	17.01		883	97,852	integral component of membrane [GO:0016021]; ATP binding [GO:0005524]; heat	
Lysine decarboxylase (EC 4.1.1.18)	Gasu_39370	23.1	4.1.1.18	507	56,061	lysine decarboxylase activity [GO:0008923]	
Probable ATP-dependent transporter ycf16	Gasu_39400	14.8		288	32,196	chloroplast [GO:0009507]; ATP binding [GO:0005524]; ATPase-coupled	
Box C/D snoRNP component Nop58	Gasu_39490	15.13		515	57,523	nucleolus [GO:0005730]; ribosome biogenesis [GO:0042254]	
Transcription elongation factor SPT5	Gasu_39630	15.56		1029	116,168	nucleus [GO:0005634]; ribosome [GO:0005840]; structural constituent of ribosome	
Prolyl-tRNA synthetase (EC 6.1.1.15)	Gasu_39770	21.5	6.1.1.15	569	65,791	cytoplasm [GO:0005737]; ATP binding [GO:0005524]; proline-tRNA ligase activity	
Guanosine-3',5'-bis(Diphosphate) 3'-	Gasu_39890	22.87	3.1.7.2	576	66,223	guanosine-3',5'-bis(diphosphate) 3'-diphosphatase activity [GO:0008893];	
2-hydroxyglutarate dehydrogenase isoform 1 (2-	Gasu_39930	80.27	1.1.99.2	438	49,099	2-hydroxyglutarate dehydrogenase activity [GO:0047545]	
30S ribosomal protein S3, chloroplastic	Gasu_40370	40.43		345	39,780	ribosome [GO:0005840]; rRNA binding [GO:0019843]; structural constituent of	
Ubiquitin family protein	Gasu_40840	13.88		258	29,006	nucleus [GO:0005634]; cell cycle [GO:0007049]	
Beta-galactosidase (EC 3.2.1.23)	Gasu_40850	76.29	3.2.1.23	1171	135,511	beta-galactosidase activity [GO:0004565]; carbohydrate metabolic process	
Adenylate kinase (EC 2.7.4.3) (ATP-AMP	Gasu_41080	39.04	2.7.4.3	267	29,972	cytosol [GO:0005829]; mitochondrial intermembrane space [GO:0005758];	
MFS transporter, SP family, sugar:H+ symporter	Gasu_41440	14.27		231	26,073	integral component of membrane [GO:0016021]; transmembrane transporter	
V-type H+-transporting ATPase subunit d (EC	Gasu_41690	32.23	3.6.3.14	292	32,587	Golgi apparatus [GO:0005794]; plasma membrane [GO:0005886]; vacuolar	
Uncharacterized protein	Gasu_41740	21.9		929	107,257	cytosol [GO:0005829]; EARP complex [GO:1990745]; endocytic recycling	
DNA ligase	Gasu_42030	31.32		574	64,952	nucleus [GO:0005634]; ligase activity [GO:0016874]; cellular response to DNA	
Protein phosphatase (EC 3.1.3.16)	Gasu_42380	15.2	3.1.3.16	281	30,986	metal ion binding [GO:0046872]; protein serine phosphatase activity	
DNA repair protein RAD51 homolog	Gasu_42740	16.69		365	40,092	chloroplast [GO:0009507]; chromosome [GO:0005694]; nucleus [GO:0005634];	

DEP domain-containing protein	Gasu_42980	27.33		499	57,965	intracellular signal transduction [GO:0035556]
Chaperone protein / DnaJ-related protein	Gasu_43060	37.38		152	16,740	integral component of membrane [GO:0016021]
TatD DNase family protein isoform 1 (TatD	Gasu_43250	203.3		321	36,270	endodeoxyribonuclease activity, producing 5'-phosphomonoesters [GO:0016888]
RNA polymerase primary sigma factor	Gasu_43280	25.18		670	78,065	DNA binding [GO:0003677]; sigma factor activity [GO:0016987]; DNA-templated
OMPdecase (EC 2.4.2.10) (EC 4.1.1.23)	Gasu_43430	25.93	2.4.2.10;	502	55,943	chloroplast [GO:0009507]; orotate phosphoribosyltransferase activity
Ubiquitin fusion degradation protein	Gasu_43460	17.24		335	36,976	ubiquitin-dependent protein catabolic process [GO:0006511]
DNA-directed RNA polymerases I and III subunit	Gasu_43470	125.8		300	34,245	nucleus [GO:0005634]; DNA binding [GO:0003677]; protein dimerization activity
PDZ GRASP-type domain-containing protein	Gasu_43650	19.01		90	10,459	Golgi apparatus [GO:0005794]; membrane [GO:0016020]
Serine/threonine protein kinase (EC 2.7.11.1)	Gasu_44250	102.2	2.7.11.1	881	100,109	ATP binding [GO:0005524]; protein serine kinase activity [GO:0106310]; protein
Ubiquitin-protein ligase E3 (EC 6.3.2.19)	Gasu_44390	22.91	6.3.2.19	1119	124,765	ligase activity [GO:0016874]; ubiquitin-protein transferase activity [GO:0004842]
Single-strand DNA-binding protein	Gasu_44730	15.06		213	24,666	single-stranded DNA binding [GO:0003697]; DNA replication [GO:0006260]
Uncharacterized protein	Gasu_45340	41.87		425	48,063	integral component of membrane [GO:0016021]
Tryptophanyl-tRNA synthetase (EC 6.1.1.2)	Gasu_45460	20.68	6.1.1.2	349	39,924	ATP binding [GO:0005524]; tryptophan-tRNA ligase activity [GO:0004830];
Pre-mRNA-splicing factor ATP-dependent RNA	Gasu_45470	167.9	3.6.4.13	1118	125,892	nucleus [GO:0005634]; ATP binding [GO:0005524]; hydrolase activity
Transcriptional repressor NF-X1	Gasu_45530	31.69		980	110,356	nucleus [GO:0005634]; DNA-binding transcription factor activity [GO:0003700];
Uncharacterized protein	Gasu_45640	53.06		259	30,714	
Serine/threonine kinase 19 (EC 2.7.11.1)	Gasu_45700	46.53	2.7.11.1	275	32,163	protein serine kinase activity [GO:0106310]; protein threonine kinase activity
Carotenoid cis-trans isomerase, CrtH-like	Gasu_46400	90.97		591	65,242	isomerase activity [GO:0016853]; oxidoreductase activity [GO:0016491]
UDP-glucuronate decarboxylase (EC 4.1.1.35)	Gasu_46450	36.26	4.1.1.35	344	38,989	NAD+ binding [GO:0070403]; UDP-glucuronate decarboxylase activity
Uncharacterized protein	Gasu_46510	16.97		232	26,190	integral component of membrane [GO:0016021]
Xylulokinase (EC 2.7.1.17)	Gasu_46540	38.53	2.7.1.17	542	61,413	xylulokinase activity [GO:0004856]
DNA-directed RNA polymerase subunit beta	Gasu_46620	99.35	2.7.7.6	1158	131,091	DNA binding [GO:0003677]; DNA-directed 5'-3' RNA polymerase activity
Amidophosphoribosyltransferase (ATase) (EC	Gasu_46700	14.49	2.4.2.14	545	60,501	amidophosphoribosyltransferase activity [GO:0004044]; iron-sulfur cluster binding
Alpha-galactosidase (EC 3.2.1.22)	Gasu_48280	59.05	3.2.1.22	894	100,725	chloroplast [GO:0009507]; raffinose alpha-galactosidase activity [GO:0052692];
Glutathione S-transferase (EC 2.5.1.18)	Gasu_48410	27.77	2.5.1.18	518	61,256	integral component of membrane [GO:0016021]; glutathione transferase activity
ER membrane protein complex subunit 2	Gasu_48540	25.11		298	35,290	EMC complex [GO:0072546]; transferase activity [GO:0016740]
Formylglycinamide ribonucleotide	Gasu_48850	28.38	6.3.5.3	1439	161,784	chloroplast stroma [GO:0009570]; mitochondrion [GO:0005739]; ATP binding
X-Pro aminopeptidase (EC 3.4.11.9)	Gasu_49250	53.22	3.4.11.9	504	56,811	manganese ion binding [GO:0030145]; metalloaminopeptidase activity
GC-rich sequence DNA-binding factor	Gasu_49450	16.15		663	78,332	nucleus [GO:0005634]; DNA binding [GO:0003677]; mRNA splicing, via
DNA-directed RNA polymerase subunit beta	Gasu_49590	24.35	2.7.7.6	1184	133,264	DNA binding [GO:0003677]; DNA-directed 5'-3' RNA polymerase activity
Uncharacterized protein	Gasu_49800	38.35		436	50,988	
Conserved oligomeric Golgi complex subunit 5	Gasu_49940	37.21		688	78,689	Golgi transport complex [GO:0017119]; membrane [GO:0016020]; intra-Golgi
Zinc finger (CCCH-type) family protein	Gasu_50210	23.42		286	33,980	integral component of membrane [GO:0016021]
MAGE domain-containing protein	Gasu_50290	20.54		258	28,874	
Uncharacterized protein	Gasu_50380	39.54		253	28,387	
Sodium/hydrogen exchanger	Gasu_50550	17.61		653	72,374	integral component of membrane [GO:0016021]; sodium:proton antiporter activity
Splicing factor U2AF 35 kDa subunit	Gasu_50930	48.37		285	34,184	U2AF complex [GO:0089701]; metal ion binding [GO:0046872]; RNA binding
Uncharacterized protein	Gasu_51290	16.73		358	41,824	
Uncharacterized protein	Gasu_51380	164.7		424	48,563	
HIV-1 Vpr-binding protein isoform 1 (HIV-1 Vpr-	Gasu_51450	23.76		1417	159,825	nucleus [GO:0005634]; protein ubiquitination [GO:0016567]
DNA damage-binding protein 2	Gasu_51510	35.6		571	65,575	Cul4-RING E3 ubiquitin ligase complex [GO:0080008]; nucleus [GO:0005634];
Glutamine amidotransferase (EC 6.3.5.2)	Gasu_51560	22.99	6.3.5.2	517	57,681	ATP binding [GO:0005524]; GMP synthase (glutamine-hydrolyzing) activity
AAA-type ATPase	Gasu_52190	18.53		541	61,186	chloroplast [GO:0009507]; ATP binding [GO:0005524]
Vesicle-fusing ATPase (EC 3.6.4.6)	Gasu_52270	15.32	3.6.4.6	754	84,458	chloroplast [GO:0009507]; ATP binding [GO:0005524]; ATP hydrolysis activity
E3 SUMO-protein ligase RanBP2	Gasu_53080	16.36		193	22,504	ligase activity [GO:0016874]; intracellular transport [GO:0046907]
Protein translocase subunit SecA	Gasu_53400	30.78		927	108,221	membrane [GO:0016020]; ATP binding [GO:0005524]; protein import
tRNA (guanine(26)-N(2))-dimethyltransferase	Gasu_53530	14.28	2.1.1.216	589	66,656	tRNA (guanine-N2-)-methyltransferase activity [GO:0004809]; tRNA binding
Cleavage and polyadenylation specifity factor	Gasu_53550	44.9		717	81,562	nucleus [GO:0005634]; mRNA processing [GO:0006397]; snRNA processing
Monodehydroascorbate reductase (NADH) (EC	Gasu_53600	42.21	1.6.5.4	597	67,888	mitochondrion [GO:0005739]; flavin adenine dinucleotide binding [GO:0050660];

Hydroxymethylglutaryl-CoA lyase (EC 4.1.3.4)	Gasu_53620	64.07	4.1.3.4	336	36,115	chloroplast [GO:0009507]; hydroxymethylglutaryl-CoA lyase activity [GO:0004419]	
N-acetylglucosaminyltransferase	Gasu_53640	31.67		545	63,340	membrane [GO:0016020]; glucuronosyltransferase activity [GO:0015020]	
Serine/threonine protein kinase	Gasu_54050	115.3		1008	113,930	ATP binding [GO:0005524]; protein serine/threonine kinase activity [GO:0004674]	
Aspartyl protease	Gasu_54460	15.09		493	53,194	aspartic-type endopeptidase activity [GO:0004190]	Yes
Delta-aminolevulinic acid dehydratase (EC	Gasu_54890	21.28	4.2.1.24	417	46,961	chloroplast [GO:0009507]; metal ion binding [GO:0046872]; porphobilinogen	
Serine-type peptidase (DEGP1)	Gasu_55030	22.49		393	43,341	serine-type endopeptidase activity [GO:0004252]	
Uncharacterized protein	Gasu_55110	29.96		415	47,697	integral component of membrane [GO:0016021]; glycosyltransferase activity	
Bifunctional 6-phosphofructo-2-kinase / fructose-	Gasu_55220	19.48	2.7.1.105;	473	54,826	6-phosphofructo-2-kinase activity [GO:0003873]; ATP binding [GO:0005524];	
Spc7 domain-containing protein	Gasu_55540	19.44		1075	124,120		
C-CAP/cofactor C-like domain-containing	Gasu_55560	22.06		509	57,556	cell morphogenesis [GO:0000902]	
Glycosyl transferase family 1	Gasu_55570	14.38		503	57,586	integral component of membrane [GO:0016021]; transferase activity [GO:0016740]	
Translocation protein, Sec family	Gasu_55610	30.69		474	52,427	integral component of membrane [GO:0016021]; protein transport [GO:0015031]	
Dynamin family protein	Gasu_55810	53.38		745	85,336	integral component of membrane [GO:0016021]; GTP binding [GO:0005525]	
Uncharacterized protein	Gasu_55840	23.36		456	53,585		
Alpha-L-arabinofuranosidase B	Gasu_55910	17.03		455	51,817	hydrolase activity, hydrolyzing O-glycosyl compounds [GO:0004553]	
Uncharacterized protein	Gasu_56020	62.67		729	83,608		
Uncharacterized protein	Gasu_56380	21.8		340	38,565	lyase activity [GO:0016829]; protein-phycocyanobilin linkage [GO:0017009]	
PsbB mRNA maturation factor Mbb1	Gasu_56860	14.58		518	60,633	mRNA binding [GO:0003729]; mRNA processing [GO:0006397]	
Nuclear cap-binding protein subunit 2 (20 kDa	Gasu_57000	18.1		195	23,012	nuclear cap binding complex [GO:0005846]; nucleus [GO:0005634]; RNA cap	
Uncharacterized protein	Gasu_57140	15.2		170	19,847	spliceosomal complex [GO:0005681]; mRNA splicing, via spliceosome	
tRNA (adenine(58)-N(1))-methyltransferase (EC	Gasu_57240	25.23	2.1.1.220	319	36,295	tRNA (m1A) methyltransferase complex [GO:0031515]; tRNA (adenine-N1-)-	
Uncharacterized protein	Gasu_57390	104.2		754	85,538	protein phosphatase binding [GO:0019903]; regulation of phosphoprotein	
Ribonucleoside-diphosphate reductase subunit	Gasu_57540	22.84	1.17.4.1	464	52,694	ribonucleoside-diphosphate reductase activity, thioredoxin disulfide as acceptor	
MFS transporter, SP family, sugar:H+ symporter	Gasu_57650	25.41		512	57,656	integral component of membrane [GO:0016021]; transmembrane transporter	
Malonyl-CoA decarboxylase (EC 4.1.1.9)	Gasu_57660	15.17	4.1.1.9	563	65,670	malonyl-CoA decarboxylase activity [GO:0050080]; fatty acid biosynthetic process	
Zinc transporter, ZIP family	Gasu_57920	38.86		328	35,495	integral component of membrane [GO:0016021]; zinc ion transmembrane	
Ferrochelatase (EC 4.99.1.1)	Gasu_58560	23.23	4.99.1.1	462	52,680	ferrochelatase activity [GO:0004325]; heme biosynthetic process [GO:0006783]	
Uncharacterized protein	Gasu_58630	16.98		852	97,604	integral component of membrane [GO:0016021]	
Molecular chaperone DnaK	Gasu_58930	244.1		878	97,720	ATP binding [GO:0005524]	
O-phosphoserine phosphohydrolase (EC	Gasu_59090	33.05	3.1.3.3	441	49,804	D-phosphoserine phosphatase activity [GO:0036425]; L-phosphoserine	
TFIIS N-terminal domain-containing protein	Gasu_59330	23.91		410	47,530	nucleus [GO:0005634]	
Uncharacterized protein	Gasu_59690	23.38		190	21,280	ornithine decarboxylase inhibitor activity [GO:0008073]	
Cyclin-dependent serine/threonine protein	Gasu_59750	25.76	2.7.11.1	349	41,126	ATP binding [GO:0005524]; protein serine kinase activity [GO:0106310]; protein	
Peptidase	Gasu_59770	21.43		919	102,908	integral component of membrane [GO:0016021]; metalloexopeptidase activity	
Diphthamide synthase (EC 6.3.1.14)	Gasu_59960	19.03	6.3.1.14	709	80,823	diphthine-ammonia ligase activity [GO:0017178]	
Pre-mRNA-processing factor 4	Gasu_59990	66.99		793	93,665	mRNA cis splicing, via spliceosome [GO:0045292]	
Thylakoid protein	Gasu_60110	16.28		316	36,620	photosystem II assembly [GO:0010207]	
Fused signal recognition particle receptor	Gasu_60350	33.92		626	70,942	chloroplast [GO:0009507]; signal recognition particle receptor complex	
GPN-loop GTPase (EC 3.6.5.-)	Gasu_60420	14.27	3.6.5.-	351	39,266	cytoplasm [GO:0005737]; nucleus [GO:0005634]; GTP binding [GO:0005525];	
Haloacid dehalogenaselike hydrolase	Gasu_60510	39.09		301	34,401	hydrolase activity [GO:0016787]	
Sedoheptulose-1,7-bisphosphatase, chloroplast	Gasu_60820	14.39	3.1.3.37	312	34,245	metal ion binding [GO:0046872]; sedoheptulose-bisphosphatase activity	
Pantetheine-phosphate adenylyltransferase (EC	Gasu_62590	40.91	2.7.7.3	333	37,135	pantetheine-phosphate adenylyltransferase activity [GO:0004595]; biosynthetic	
GTP-binding protein	Gasu_63070	24.01		479	53,803	GTP binding [GO:0005525]; GTPase activity [GO:0003924]	
[pt] maturase	Gasu_63140	17.14		568	68,658		
Uncharacterized protein	Gasu_63440	40.29		461	53,633	integral component of membrane [GO:0016021]	
AAA-type ATPase	Gasu_63910	17.04		626	72,141	chloroplast [GO:0009507]; ATP binding [GO:0005524]	
Methyltransferase GidB (Glucose inhibited	Gasu_64020	16.99		282	32,025	cytoplasm [GO:0005737]; rRNA methyltransferase activity [GO:0008649]	
Uncharacterized protein (Fragment)	Gasu_65160	47.34		95	9,610		
DNA replication licensing factor MCM7 (EC	Gasu_02880	24.56	3.6.4.12	803	90,207	MCM complex [GO:0042555]; nucleus [GO:0005634]; ATP binding [GO:0005524];	

Supplementary Table 6: Summary of the sequencing kit number and flow cell using Oxford Nanopore Technologies' (ONT) MinION system

Strain	Library prep method	MinION flow cell	Run date
107	LSK108	FLO-MIN106 R9	20170118
017	LSK108 with native barcoding expansion EXP-NBD103	FLO-MIN106 R9.4	20170808
427	LSK108 with native barcoding expansion EXP-NBD103	FLO-MIN106 R9.4	20170808
033	LSK108	FLO-MIN106 R9.4	20180411
074	LSK108	FLO-MIN106 R9.4	20180626
138	LSK109	FLO-MIN106 R9.4.1	20190815

Supplementary Table 7: Distribution of CAZymes in G. sulphuraria genomes and in other red alage genomes.

		G. sulphuraria								
CAZyme	**Family**	***017***	***033***	***074***	***107***	***138***	***427***	***Average***	***C. merolae***	***P. purpureum***
	1	2	2	1	2	2	0	1.5	2	1
	2	5	4	6	4	5	5	4.8333	7	5
	4	11	11	11	10	11	11	10.833	6	7
	5	1	1	1	1	1	1	1	1	2
	7	0	0	1	0	0	1	0.3333		1
	8	5	5	4	4	5	5	4.6667	4	5
	10	1	1	1	1	1	1	1		2
GT	**13**	1	0	1	1	1	2	1		2
	14	5	5	5	5	5	4	4.8333		3
	17	1	1	1	1	1	0	0.8333		1
	19	2	2	2	2	2	2	2	1	1
	20	4	4	4	4	4	4	4	4	6
	22	3	2	2	2	3	3	2.5	1	
	23									2
	24	0	1	1	1	1	1	0.8333	1	

	25	1	1	2	1	1	1	1.1667		1
	28	3	3	3	3	3	3	3	1	6
	30	1	1	1	1	1	1	1		
	31	7	6	7	7	8	7	7		
	32	2	2	2	1	1	1	1.5	1	4
	33	1	1	1	1	1	1	1	1	
	34	0	1	1	0	0	0	0.3333		
	35	2	2	2	2	2	2	2	1	1
	39	1	2	2	1	2	2	1.6667	6	4
	41	2	2	2	2	2	2	2		1
	45	1	1	1	1	1	1	1		1
	47	0	1	1	1	1	1	0.8333		1
	49	0	1	1	0	0	0	0.3333	1	1
	57	2	1	2	1	1	2	1.5	2	
	58	1	0	1	1	1	1	0.8333	1	
	59	1	0	1	0	1	1	0.6667		
	61	1	1	1	1	1	1	1		1
	64	2	2	2	2	2	2	2	2	
	66	1	1	1	1	1	1	1	1	
	69	2	1	1	3	4	1	2		
	71								2	
	77	1	1	1	1	2	2	1.3333	1	4
	78									1
	83	1	1	1	1	1	1	1	1	
	90	1	1	1	1	1	1	1		1
	nc								7	
GH	**2**	1	1	1	2	1	1	1.17		1
	5									1
	13	6	6	6	6	5	6	5.83	6	10
	14	3	3	3	3	3	3	3	1	2
	15	2	3	3	3	3	3	2.83		
	20									1
	27	1	1	1	0	1	1	0.83		
	30	1	2	2	1	2	2	1.67		
	31	4	5	5	4	5	5	4.67	5	1
	35	7	7	7	7	5	3	6	1	2
	36	1	2	2	1	2	2	1.67	1	3
	37	1	1	1	1	1	1	1	2	1
	38	2	2	2	2	2	2	2		
	47	1	1	1	1	1	1	1	1	
	63	1	1	1	1	1	1	1	1	1
	77	1	1	1	1	1	1	1	2	3

	85		1	1	1	1	1	1	1		1	1
	2		7	7	7	7	7	6	6.83			
AA	**3**		1	1	1	1	1	1	1		1	
	6		3	3	3	1	3	3	2.67		1	
	9											1
	20		3	3	2	2	3	3	2.67		7	6
	25											1
CBM	**33**											1
	41										1	
	48		4	4	5	4	5	3	4.17		5	10
	57											1
CE	**11**		1	1	1	1	1	1	1		1	1
	15											1

Supplementary Table 8: Orthologue analysis of G. sulphuraria genomes and other red algal genomes

	017	**033**	**074**	**107**	**138**	**427**	**C. merolae**	**P_purpureum**
Number of genes	6306	5092	5690	5959	5728	5531	4803	9898
Number of genes in orthogroups	6097	5057	5482	5722	5404	5335	3780	6975
Number of unassigned genes	209	35	208	237	324	196	1023	2923
Percentage of genes in orthogroups	96.7	99.3	96.3	96	94.3	96.5	78.7	70.5
Percentage of unassigned genes	3.3	0.7	3.7	4	5.7	3.5	21.3	29.5
Number of orthogroups containing species	4865	4482	4869	4722	4783	4818	3309	4051
Percentage of orthogroups containing species	77.3	71.2	77.3	75	76	76.5	52.6	64.3
Number of species-specific orthogroups	11	1	11	23	18	5	50	520
Number of genes in species-specific orthogroups	24	7	36	145	86	15	142	2473
Percentage of genes in species-specific orthogroups	0.4	0.1	0.6	2.4	1.5	0.3	3	25

Supplementary Table 9: Summary of unfiltered LCMS analysis.

Majority protein IDs	Fasta headers	Razor + unique peptides	Sequence coverage [%]	Unique + razor sequence coverage [%]	Unique sequence coverage [%]	Mol. weight [kDa]	Seq length	iBAQ	MS/MS Count	*Lysate ave peptide hit*
M2XJ88A3:AA3:A23	>tr|M2XJ88|M2XJ88_GALSU Alpha-glucosidase OS=Galdieria sulphuraria OX=130081 GN=Gasu_25520 PE=4 SV=1	29	38.2	38.2	36.6	57.83	508	1099200000	772	7.67
M2Y2Y7;M2Y2J8	>tr|M2Y2Y7|M2Y2Y7_GALSU Alpha-glucosidase isoform 1 OS=Galdieria sulphuraria OX=130081 GN=Gasu_25530 PE=4 SV=1;>tr|M2Y2J8|M2Y2J8_GALSU Alpha-glucosidase isoform 2 OS=Galdieria sulphuraria OX=130081 GN=Gasu_25530 PE=4 SV=1	26	44.9	43.4	43.4	58.26	512	185670000	496	1.33
M2XIP7	>tr|M2XIP7|M2XIP7_GALSU Beta-galactosidase OS=Galdieria sulphuraria OX=130081 GN=Gasu_27500 PE=3 SV=1	17	20.3	20.3	19.6	118.2	1038	2081700	83	NA
M2XRS9	>tr|M2XRS9|M2XRS9_GALSU Alpha-galactosidase OS=Galdieria sulphuraria OX=130081 GN=Gasu_60270 PE=3 SV=1	14	32.9	32.9	32.9	60.64	535	20942000	186	1.33
M2W2P6	>tr|M2W2P6|M2W2P6_GALSU Beta-galactosidase OS=Galdieria sulphuraria OX=130081 GN=Gasu_27490 PE=3 SV=1	14	19.1	19.1	19.1	108.4	952	881960	70	NA
M2WAH4	>tr|M2WAH4|M2WAH4_GALSU Beta-lg-H3/fasciclin OS=Galdieria sulphuraria OX=130081 GN=Gasu_02360 PE=4 SV=1	7	41.1	41.1	41.1	18.64	175	1683200000	233	3.67
M2XHE4	>tr|M2XHE4|M2XHE4_GALSU Uncharacterized protein OS=Galdieria sulphuraria OX=130081 GN=Gasu_31410 PE=4 SV=1	6	27.5	27.5	27.5	41.09	378	1002000000	187	NA
A5JW32	>tr|A5JW32|A5JW32_GALSU Peroxidase OS=Galdieria sulphuraria OX=130081 GN=Prx01 PE=2 SV=1	6	19.5	19.5	19.5	34.52	323	45402000	65	2.67
M2W452	>tr|M2W452|M2W452_GALSU Uncharacterized protein OS=Galdieria sulphuraria OX=130081 GN=Gasu_21980 PE=4 SV=1	5	11.5	11.5	11.5	43.44	393	39163000	54	3.67
M2X275	>tr|M2X275|M2X275_GALSU Enolase OS=Galdieria sulphuraria OX=130081 GN=Gasu_21490 PE=3 SV=1	5	18	18	18	47.3	438	160620	8	27.67
M2XTK8	>tr|M2XTK8|M2XTK8_GALSU Beta-lg-H3/fasciclin OS=Galdieria sulphuraria OX=130081 GN=Gasu_54440 PE=4 SV=1	4	26.7	26.7	26.7	17.97	165	800680000	161	1.33
M2XQN9;A5JW35	>tr|M2XQN9|M2XQN9_GALSU Peroxidase (Fragment) OS=Galdieria sulphuraria OX=130081 GN=Gasu_64070 PE=3 SV=1;>tr|A5JW35|A5JW35_GALSU Peroxidase OS=Galdieria sulphuraria OX=130081 GN=Prx04 PE=2 SV=1	4	20.1	20.1	20.1	25.4	244	135690000	104	2
A5JW34	>tr|A5JW34|A5JW34_GALSU Peroxidase OS=Galdieria sulphuraria OX=130081 GN=Prx03 PE=2 SV=1	4	18.9	18.9	18.9	31.75	297	18144000	48	1.67
M2WSY0	>tr|M2WSY0|M2WSY0_GALSU Aspartyl protease OS=Galdieria sulphuraria OX=130081 GN=Gasu_54460 PE=3 SV=1	4	12.8	12.8	12.8	53.19	493	9775000	70	4.33
M2WWC0	>tr|M2WWC0|M2WWC0_GALSU Uncharacterized protein OS=Galdieria sulphuraria OX=130081 GN=Gasu_41530 PE=4 SV=1	4	19.3	19.3	19.3	26.05	228	2467500	27	1.67
M2XCY8	>tr|M2XCY8|M2XCY8_GALSU VanW family protein OS=Galdieria sulphuraria OX=130081 GN=Gasu_46440 PE=4 SV=1	4	23.6	23.6	23.6	23.94	225	675000	6	2.67
M2XNY3	>tr|M2XNY3|M2XNY3_GALSU Uncharacterized protein OS=Galdieria sulphuraria OX=130081 GN=Gasu_09340 PE=4 SV=1	3	14.6	14.6	14.6	18.47	164	7985100	14	NA
M2XHJ6	>tr|M2XHJ6|M2XHJ6_GALSU Purple acid phosphatase OS=Galdieria sulphuraria OX=130081 GN=Gasu_29970 PE=3 SV=1	3	9.3	9.3	9.3	61.15	538	129950	4	NA
M2W430	>tr|M2W430|M2W430_GALSU Uncharacterized protein OS=Galdieria sulphuraria OX=130081 GN=Gasu_21790 PE=4 SV=1	2	6.5	6.5	6.5	43.46	401	123370000	62	3
M2W312	>tr|M2W312|M2W312_GALSU Uncharacterized protein OS=Galdieria sulphuraria OX=130081 GN=Gasu_24700 PE=4 SV=1	2	10.2	10.2	10.2	26.56	244	36583000	24	NA
A5JW33	>tr|A5JW33|A5JW33_GALSU Peroxidase OS=Galdieria sulphuraria OX=130081 GN=Prx02 PE=2 SV=1	2	7.8	7.8	7.8	36.62	345	7565400	17	3
M2X634;M2VTN6;M2X4Q0;M2VV38	>tr|M2X634|M2X634_GALSU Ubiquitin OS=Galdieria sulphuraria OX=130081 GN=Gasu_10140 PE=4 SV=1;>tr|M2VTN6|M2VTN6_GALSU Ubiquitin OS=Galdieria sulphuraria OX=130081 GN=Gasu_58040 PE=4 SV=1;>tr|M2X4Q0|M2X4Q0_GALSU Ubiquitin OS=Galdieria sulphuraria	2	19.5	19.5	19.5	14.65	128	2458200	19	NA,8.67,NA,NA
M2X2Y9	>tr|M2X2Y9|M2X2Y9_GALSU Elongation factor 1-alpha OS=Galdieria sulphuraria OX=130081 GN=Gasu_19840 PE=3 SV=1	2	4.2	4.2	4.2	49.97	452	231460	17	25
M2XEA9	>tr|M2XEA9|M2XEA9_GALSU Aldo/keto reductase OS=Galdieria sulphuraria OX=130081 GN=Gasu_41590 PE=4 SV=1	2	14	14	14	18.21	157	198750	7	5.3

M2W5E2	>tr|M2W5E2|M2W5E2_GALSU Actin OS=Galdieria sulphuraria OX=130081 GN=Gasu_17630 PE=3 SV=1	2	9.1	9.1	9.1	41.72	375	100220	7	17.33
M2XBV6	>tr|M2XBV6|M2XBV6_GALSU Serine/threonine protein kinase OS=Galdieria sulphuraria OX=130081 GN=Gasu_50860 PE=4 SV=1	2	2.6	2.6	2.6	101.1	911	51616	4	NA

Supplementary Table 10: BLASTp top 10 hits from different G. sulphuraria genes.

Description	Max Score	Query Cover	E value	% ident	Accession
17800					
peroxidase [Galdieria sulphuraria]	659	100%	0	100	XP_005707539.1
peroxidase [Galdieria sulphuraria]	423	70%	7.00E-146	85.15	XP_005703979.1
peroxidase [Galdieria sulphuraria]	369	73%	5.00E-124	70.29	XP_005707538.1
peroxidase [Galdieria sulphuraria]	347	79%	4.00E-116	65.04	XP_005703978.1
peroxidase [Galdieria sulphuraria]	301	50%	9.00E-99	85.89	XP_005702458.1
hypothetical protein N665_0630s0027 [Sinapis alba]	126	54%	2.00E-30	39.34	KAF8086277.1
hypothetical protein F2Q68_00033027 [Brassica cretica]	124	54%	9.00E-30	38.25	KAF2543572.1
PREDICTED: L-ascorbate peroxidase 2, cytosolic [Brassica oleracea var. oleracea]	124	54%	1.00E-29	38.25	XP_013618190.1
hypothetical protein Bca52824_085602 [Brassica carinata]	124	54%	2.00E-29	38.25	KAG2245974.1
L-ascorbate peroxidase 2, cytosolic-like [Rhodamnia argentea]	123	52%	2.00E-29	39.33	XP_030522061.1
21790					
-	-	-	-	-	-
21980					
hypothetical protein [Pseudomonas syringae]	241	34%	2.00E-75	85.07	WP_181426761.1
TPA: SGNH/GDSL hydrolase family protein [Actinobacteria bacterium]	100	92%	4.00E-19	26.33	HGW04074.1
SGNH/GDSL hydrolase family protein [Acidimicrobiia bacterium]	86.3	89%	2.00E-14	25.27	MBV9412948.1
SGNH/GDSL hydrolase family protein [Acidimicrobiia bacterium]	80.1	89%	2.00E-12	24.66	MBV8161695.1
hypothetical protein [Acidimicrobiia bacterium]	74.7	83%	2.00E-10	26.01	MPY95726.1
GDSL-type esterase/lipase family protein [Arenicella xantha]	70.5	90%	4.00E-09	25.65	WP_170131916.1
hypothetical protein EPO04_01555 [Patescibacteria group bacterium]	53.1	91%	0.002	20.11	TAK89772.1
hypothetical protein [Phenylobacterium sp. 20VBR1]	49.7	26%	0.026	33.91	WP_215341195.1
24700					
hypothetical protein Gasu_41530 [Galdieria sulphuraria]	69.7	74%	3.00E-10	27.75	XP_005704825.1

25520					
alpha-glucosidase isoform 1 [Galdieria sulphuraria]	517	96%	9.00E-177	49.8	XP_005706699.1
alpha-glucosidase isoform 2 [Galdieria sulphuraria]	506	96%	4.00E-172	48.18	XP_005706698.1
hypothetical protein [bacterium]	275	89%	7.00E-83	33.69	NBU21166.1
glucoamylase [Jimgerdemannia flammicorona]	271	91%	3.00E-81	34.11	RUS31521.1
glucoamylase [Jimgerdemannia flammicorona]	270	91%	5.00E-81	34.11	RUP42781.1
glycoside hydrolase family 15 protein [Bdellovibrionales bacterium]	268	93%	3.00E-80	31.76	MBS1985031.1
putative glucoamylase [Terfezia claveryi]	267	88%	1.00E-79	33.84	KAF8434004.1
glycoside hydrolase family 15 protein [Deltaproteobacteria bacterium]	265	88%	7.00E-79	33.41	MBI3557677.1
hypothetical protein [Deltaproteobacteria bacterium]	260	87%	3.00E-77	33.04	MBM4315856.1
hypothetical protein [Deltaproteobacteria bacterium]	258	87%	1.00E-76	32.74	MBM4303691.1
25530					
alpha-glucosidase isoform 2 [Galdieria sulphuraria]	1064	100%	0	96.79	XP_005706698.1
alpha-glucosidase [Galdieria sulphuraria]	517	97%	1.00E-176	50.4	XP_005706697.1
glycoside hydrolase family 15 protein [Bdellovibrionales bacterium]	267	96%	8.00E-80	33.87	MBS1985031.1
hypothetical protein [bacterium]	258	95%	3.00E-76	31.1	NBV51219.1
hypothetical protein [bacterium]	258	90%	3.00E-76	34.04	NBU21166.1
glycoside hydrolase family 15 protein [Morchella conica CCBAS932]	253	91%	3.00E-74	33.47	RPB07199.1
glucoamylase precursor [Aureobasidium melanogenum]	256	87%	2.00E-73	33.55	KAG9570551.1
glucoamylase precursor [Aureobasidium melanogenum]	256	87%	2.00E-73	33.55	KAH0404230.1
glucoamylase [Jimgerdemannia flammicorona]	250	90%	2.00E-73	32.54	RUS31521.1
glucoamylase precursor [Aureobasidium melanogenum]	254	87%	6.00E-73	33.55	KAG9531886.1
27490					
beta-galactosidase [Galdieria sulphuraria]	850	96%	0	47.14	XP_005708381.1
beta-galactosidase [Paenibacillus aceris]	793	93%	0	43.95	WP_205300924.1
hypothetical protein [Paenibacillus aceris]	793	93%	0	43.95	NHW34940.1
beta galactosidase [Jimgerdemannia flammicorona]	788	98%	0	43.49	RUP52310.1
beta-galactosidase [Ktedonobacteraceae bacterium]	788	93%	0	45.77	MBV9614761.1

beta-galactosidase [Ktedonobacteraceae bacterium]	788	93%	0	45.77	MBV9710357.1
beta galactosidase [Jimgerdemannia flammicorona]	786	98%	0	43.06	RUS34276.1
beta-galactosidase [Ktedonobacteraceae bacterium]	780	93%	0	44.59	MBA2397077.1
beta-galactosidase [Paenibacillus planticolens]	769	96%	0	42.36	WP_171686514.1
beta-galactosidase [Amycolatopsis vastitatis]	765	96%	0	42.83	WP_093951092.1
27500					
beta-galactosidase [Galdieria sulphuraria]	902	92%	0	48.22	XP_005708381.1
beta-galactosidase [Galdieria sulphuraria]	858	91%	0	46.58	XP_005706486.1
beta-galactosidase [Ktedonobacteraceae bacterium]	842	91%	0	45.99	MBV9614761.1
beta-galactosidase [Ktedonobacteraceae bacterium]	842	91%	0	45.99	MBV9710357.1
beta-galactosidase [Ktedonobacteraceae bacterium]	834	93%	0	44.25	MBA2397077.1
beta-galactosidase [Rhodanobacter glycinis]	832	92%	0	44.57	WP_092701590.1
beta-galactosidase [Amycolatopsis bartoniae]	820	91%	0	46.01	WP_145933416.1
beta-galactosidase [Nonomuraea rubra]	676	90%	0	40.58	WP_185108619.1
beta-galactosidase [Actinophytocola xanthii]	672	88%	0	39.53	OLF17174.1
beta-galactosidase [Actinophytocola xanthii]	671	88%	0	39.53	WP_198942858.1
29970					
hypothetical protein, conserved [Cyanidioschyzon merolae strain 10D]	712	94%	0	63.23	XP_005538413.1
metallo-dependent acid phosphatase [Galdieria sulphuraria]	434	83%	2.00E-143	45.81	XP_005704704.1
hypothetical protein F1559_004671 [Cyanidiococcus yangmingshanensis]	196	22%	1.00E-56	73.11	KAF6004604.1
hypothetical protein F1559_004672 [Cyanidiococcus yangmingshanensis]	191	21%	6.00E-55	72.57	KAF6004605.1
acid phosphatase type 7 [Bufo bufo]	144	71%	2.00E-33	31.19	XP_040262447.1
acid phosphatase type 7 [Xenopus laevis]	140	79%	2.00E-32	29.15	XP_018085798.1
uncharacterized protein PTSG_07301 [Salpingoeca rosetta]	141	80%	3.00E-32	26.49	XP_004990799.1
hypothetical protein CAEBREN_31395 [Caenorhabditis brenneri]	139	79%	6.00E-32	30.51	EGT41961.1
hypothetical protein GDO78_013319 [Eleutherodactylus coqui]	138	71%	9.00E-32	30.86	KAG9478263.1
acid phosphatase type 7 [Pantherophis guttatus]	138	78%	2.00E-31	27.73	XP_034294135.1

31410					
TPA: MAG TPA: Minor structural protein [Myoviridae sp.]	82.8	73%	6.00E-13	26.6	DAL87870.1
hypothetical protein [Acetobacter senegalensis]	63.9	66%	6.00E-07	23.88	WP_058987854.1
hypothetical protein [Thermoplasmata archaeon]	60.5	83%	7.00E-06	27.41	MBX8640793.1
hypothetical protein [Thermoplasmata archaeon]	60.1	83%	8.00E-06	27.41	MBX8642792.1
41530					
hypothetical protein Gasu_24700 [Galdieria sulphuraria]	72.8	97%	3.00E-11	27.27	XP_005706606.1

Supplementary Table 11: Compositions and concentrations of buffers used in refolding assay.

	Concentration mM											
	PCB		**MIB**			**MMT**			**CB**		**HPCP**	
pH	**Disodium Hydrogen Phosphate**	**Citric Acid**	**Sodium malonate dibasic monohydrate**	**Imidazole**	**Boric acid**	**DL-malic acid**	**MES monohydrate**	**Tris base**	**Disodium Hydrogen Phosphate**	**Citric Acid**	**KCl**	**HCl**
2	2.07	47.93	-	-	-	-	-	-	-	-	49	1
2.5	7.95	42.05	-	-	-	-	-	-	-	-	-	-
3	16.98	33.03	-	-	-	-	-	-	9.27	40.73	-	-
3.5	23.28	26.72	-	-	-	-	-	-	11.41	38.59	-	-
4	27.79	22.21	12.5	18.75	18.75	10	20	20	17.21	32.79	-	-
4.5	31.24	18.76	12.5	18.75	18.75	10	20	20	24	26	-	-
5	33.94	16.06	12.5	18.75	18.75	10	20	20	30.2	19.8	-	-
5.5	36.82	13.18	12.5	18.75	18.75	10	20	20	-	-	-	-
6	39.27	10.73	12.5	18.75	18.75	10	20	20	-	-	-	-
6.5	41.53	8.47	12.5	18.75	18.75	10	20	20	-	-	-	-
7	45.18	4.82	12.5	18.75	18.75	10	20	20	-	-	-	-
7.5	48.02	1.98	12.5	18.75	18.75	10	20	20	-	-	-	-