# “Life Finds a Way”: Extreme Genomic Features of the Eukaryotic Extremophile, *Galdieria sulphuraria*

**Jessica Mary Downing**

**PhD**

**University of York**

**Biology**

**December 2022**

# Abstract

Extremophiles, though normally bacteria and archaea, are found in all domains of life. The eukaryotic red alga, *Galdieria sulphuraria,* belongs to the class Cyanidiophyceae, the only class of eukaryotic extremophiles. *G. sulphuraria* is a moderate thermophile, growing at temperatures up to 55 ˚C, considered the upper temperature limit for photosynthesis, an acidophile (pH 0-2), and tolerates toxic levels of heavy metals and reactive oxygen species. *Galdieria* species are unique among the Cyanidiophyceae as they are metabolically flexible, and can grow both phototrophically, and heterotrophically on a variety of complex carbon sources. These traits make *G. sulphuraria* a promising candidate for use in biotechnology.

Using whole genome sequencing, I have constructed three complete, and two draft, nuclear genome assemblies, and four genome annotations of *G. sulphuraria* isolates, demonstrating that *G. sulphuraria* has a compact (~13 Mb – 16 Mb) nuclear genome, with numerous (72-73) tiny chromosomes. With the benefit of completely assembled and annotated genomes, I analysed the collinearity between the completed genome assemblies, which revealed significant structural divergence between *G. sulphuraria* isolates, supporting the notion of cryptic species within the lineage. The annotations enabled the identification of a putative thermophile specific gene, reverse gyrase, and a partial, predicted, meiotic toolkit. Moreover, single nucleotide polymorphism data, indicated that meiotic recombination may be occurring.

Finally, I estimated the *G. sulphuraria* genome wide spontaneous mutation rate, revealing that the *G. sulphuraria* nuclear genome exhibits an extraordinarily fast spontaneous mutation rate of $3.19 \times 10^{-8}$ per base pair per generation, 100-fold higher than other free living eukaryotes, with a higher rate of mutation in duplicated genomic regions. This suggests that adaptive evolution may play a role in extreme tolerance. Taken together, the genomic features which I have demonstrated provide evidence of how *G. sulphuraria* maintains a flexible, adaptive, lifestyle in a rapidly changing extreme environment.

# Table of Contents

# List of Figures

# List of Tables

## 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.

Some of this work was made publicly available in "Downing JM, Lock SCL, Iovinella M, et al. Comparisons between Complete Genomes of the Eukaryotic Extremophile *Galdieria sulphuraria* Reveals Complex Nuclear Chromosomal Structures. bioRxiv; 2022. DOI: 10.1101/2022.10.04.510839." where I am the primary author.

# Acknowledgements

First and foremost, I thank God, the Almighty, for sustaining my strength and granting me numerous blessings that have enabled the completion of this thesis. Studying the Lord's creation has been a true joy. I also thank the Blessed Virgin Mary, whose constant intercession has been a source of great comfort. I therefore dedicate this thesis to Our Lady of Perpetual Help.

I express my utmost gratitude to my primary supervisor, Professor Seth J Davis, who has been extremely supportive of my creative direction and skills development in the production of this thesis, an excellent source of knowledge and guidance, and who always cares for the welfare of his students. I am extremely grateful to Professor James Chong and Dr. Georg Feichtinger for their continued supervision and support. I am grateful to my thesis advisory panel member, Dr. Thierry Tonon. I would also like to thank my funders, the Biotechnology and Biological Sciences Research Council.

I am grateful to Dr. Daniel Jeffares, for his excellent scientific advice. I would like to thank the Bioinformatics Technology Facility and the Viking High Performance Computing team. I extend special thanks to all of my colleagues in the Davis Lab, especially to Dr. Sarah Lock and Dr. Manuella Iovinella, for their support and contributions to the *Galdieria* project, and to Dr. Kayla McCarthy, Dr. James Ronald, and Amanda Davis, for their technical support. I also thank all of my fellow students and colleagues on L2 for support, encouragement, and advice.

I am indebted to my parents, whose personal and financial support has eased the burden of completing a PhD. I also thank my grandfather, Des, for his support and encouragement throughout my studies. I thank my sisters, Sally and Lucy (who always helped me with Excel), and my cousin Ciara, for moral support. I am grateful to my fiancé, Connor, for his prayers and constant support. I extend thanks to my aunts, Helen and Finn. I would like to thank my dear friends Amy and Christine, whom I am grateful to have met during the undertaking of this PhD. I would also like to thank Fr. Richard, Fr. Daniel, and Fr. Stephen for their prayers, support, and guidance. Finally, I thank all of the teachers and educators, who taught, encouraged, and inspired me along the way.

# Chapter 1: Introduction

## Life Finds A Way

In the 3.5 billion years that life has existed on earth, life forms have moved into every ecological niche imaginable. “Life finds a way” cautions the fictional character Ian Malcolm in Michael Crichton’s 1990 science fiction novel, Jurassic Park [1]. While this thesis is not about genetically modified dinosaurs chasing people around an island, there is an element of truth in this novel. Life indeed does find a way, and this is made apparent in reality by the diversity of organisms that live at the extremes, the extremophiles.

Extremophiles are defined as organisms that have optimal growth conditions that are extreme from a human perspective [2]. One could perhaps be forgiven for thinking that life could not possibly survive in hot, acidic pools, or deep below the ocean surface, however communities of organisms have evolved together in these environments. Extreme environments represent strong selective pressures, and the manner in which extremophiles evolve to meet these challenges should reflect this – there is little expectation that these organisms are “the same” as species that inhabit moderate environments in these evolutionary scenarios.

## Life at the Extremes

Extremophiles can be divided into categories based on the optimal conditions they grow in [3]. Thermophiles have optimum growth temperatures above 45 ˚C, while hyperthermophiles, a subcategory of thermophiles, have optimal growth temperatures above 80 ˚C and are exclusively archaea and bacteria. The opposite of thermophiles, psychrophiles, grow at temperatures lower than 15 ˚C [4]. Halophiles thrive in saline conditions, piezophiles can tolerate pressures of up to 130 MPa. Alkaliphiles grow at an optimum pH of more than 9, whereas acidophiles have pH optima lower than pH 3. Other types of extremophiles include but are not limited to; radiophiles, which can tolerate high levels of radiation, and metallophiles, that tolerate high metal concentrations. Often, several extremophilic traits are present in a given organism, for example thermoacidophiles thrive at both acidic pHs and high temperatures. With their diverse traits, extremophiles are outstanding organisms, however they are some of the least studied, owing to their challenging growth

conditions [3]. Despite this, extremophiles have garnered a significant amount of interest in recent years, by virtue of their potential applications in biotechnology.

Industrially, thermophilic extremophiles have gained much attention. Many thermostable enzymes have been discovered in thermophiles and employed in the biotechnology industry. Most famously, the thermostable polymerase Taq Polymerase, isolated from the hyperthermophile *Thermus aquaticus* from Yellowstone National Park [5], revolutionised molecular biology by enabling the polymerase chain reaction to be conducted rapidly and efficiently [6]. At elevated temperatures, the risk of contamination is reduced and the solubility of reaction components is significantly increased, negating the need for specialised closed bioreactors. Consequently, there is huge potential for process improvement, in terms of cost and efficiency, by using thermophiles and their enzymes [4].

Thermostable extremozymes have been investigated for application in biomass conversion. Pretreatment methods for lignocellulosic materials that aim at combining hydrothermal procedures at higher temperatures and pressures with enzymatic degradation processes are under development [7]. By using thermostable extremozymes in this process, cooling steps can be omitted, reducing energy costs and enhancing substrate accessibility thus improving reaction rate [7] .

Thermophilic proteins possess a number of adaptations which allow them to remain stable and avoid denaturation at elevated temperatures, thus conferring an organism thermotolerance. Most commonly, these proteins have an increased number of hydrophobic residues and a larger hydrophobic core, an increased number of disulfide bonds, and increased ionic interactions. These adaptations allow for correct function at extreme temperatures by stabilising the protein and decreasing overall protein flexibility [3]. Conversely, psychrophilic proteins – those that like lower temperatures – face the opposite thermodynamic challenge, which is reduced molecular motion due to decreased entropy and enthalpy [3]. Since these enzymes operate at lower temperatures, they are desirable for industry since they can reduce energy costs by removing the need for expensive heating steps that are normally required. Psychrophilic amylases, lipases, and proteases for application in laundry detergents are arising. The pulp and paper industry, as well as the food industry, are also interested in cold-active enzymes [8].

Halophiles survive in hypersaline environments by maintaining their osmotic balance – this means that they accumulate salts intracellularly at concentrations that are isotonic with their external environment. Their proteins therefore must cope with very high salt concentrations and have adapted by acquiring a large number of negatively charged amino acid residues on the surface to prevent precipitation and have been applied for catalysis of reactions in aqueous/organic and non-aqueous media [9].

Alkaliphilic and acidophilic organisms maintain a neutral intracellular pH, therefore their intracellular proteins do not need to be adapted to extreme pHs, however their secreted enzymes are tolerant to extreme pHs and have industrial potential. Alkaliphilic proteases, amylases, lipases, and other enzymes are desirable for use in detergents. The polymer-hydrolysis industry is in search for acidophilic hydrolases, and several acidophilic enzymes used for starch hydrolysis have been isolated [9].

In addition to their industrial benefits, extremophiles are some of the oldest surviving organisms in existence and are of great biological interest for studying what life may have been like on an early earth, and can be basal in taxonomy studies [10]. The environmental conditions on the early earth were extremely hostile and rapidly changing, with organisms contending with a variety of extreme abiotic stresses including high levels of volcanic activity and a reducing atmosphere, and were not dissimilar from the habitats extremophiles reside in today [11]. These early life forms were bacteria and archaea, with the earliest organisms evolving in anaerobic conditions [12]. It remains that most known extremophiles are bacteria and archaea.

Extremophiles are gaining interest as astrobiological models, as it is hoped that the wide variety of extreme tolerance seen in the extremophiles will enable the survival of these organisms to the extreme conditions seen on other planets, such as extremes of temperature, desiccation, and increased levels of UV radiation [13]. The ability of extremophiles to resist simulated Martian conditions, such as extreme conditions of temperature, desiccation, UV radiation, and low pressure, has been investigated. It was demonstrated that *Sulfolobus solfataricus*, *Geobacillus thermantarcticus*, and *Haloterrigena hispanica* show good resistance to Mars simulated conditions [14]. Projects are underway to send extremophiles into space, and measure their tolerance in real time to conditions outside of the International

Space Station, including radiation and zero gravity [15]. The future applications of extremophiles in space are exciting.

Extremophilic bacteria and archaea have further mechanisms of extreme tolerance as well as adaptations to their protein structure. Horizontal gene transfer, which enables the exchange of DNA between organisms of different species, is a well-known driving force in prokaryotic adaptation and evolution [16]. The thermophilic bacteria *Thermotoga maritima* and *Aquifex aeolicus* have acquired ~24 and 16.2% of their genes through HGT from archaeal thermophiles. Many of these genes grant these organisms extremophilic traits required for survival, such as archaeal reverse gyrases, thought to have been acquired from bacteria, which introduce positive supercoiling to double stranded DNA, thereby increasing the melting temperature [17], and enabling some thermophilic archaea to maintain AT rich genomes [18].

Microbial communities have been reported to evolve faster in extreme environments [19], indicating that adaptive evolution plays a role in extreme tolerance. Conversely, some thermophiles, such as *Thermus thermophilus*, have extraordinarily efficient DNA repair strategies, which combat the high levels of DNA damage expected in extreme environments [20]. These DNA level adaptations, combined with adaptations to protein structure, enable the survival of bacteria and archaea within extreme environments.

## The Cyanidiophyceae

Despite the most primitive extremophiles belonging to the bacterial and archaeal lineages, extremophiles exist in all kingdoms of life. The Cyanidiophyceae comprise the basal clade of eukaryotic extremophiles, and owing to their unique characteristics as the only eukaryotic extremophiles, have been gaining interest over the last several years [21]. They are characterised by thick cell walls, a plastid, 1-3 mitochondria, a vacuole, and are either spherical or oblong in shape. The Cyanidiophyceae are red algae (Rhodophytes) that may have diverged from the other members of the Rhodophyta around 1.3bn years ago. The red algae are a distinct eukaryotic lineage whose members lack chlorophyll b and c, but contain the pigments allophycocyanin, phycocyanin, and phycoerythrin in the form of phycobilisomes on unstacked thylakoids [22]. There has been some debate about the taxonomy of the Rhodophyta and they are considered to belong to

Archaeplastida, which also comprises of green algae plus land plants. Alternatively, it had been suggested that the Rhodophytes are Protists, however the general consensus lies with placing the Rhodophytes in the Archaeplastida [23].

Taxonomical debate is not solely reserved for the higher levels of the hierarchy, and historically, there has also been confusion regarding the taxonomy of the Cyanidiophyceae. Due to the morphological similarities and the fact that these organisms often live in mixed populations, these organisms were first erroneously described as a single species. As these organisms have lost the pigment that makes them red, and instead are green in appearance, in 1896 they were placed among the blue green algae under the name *Chroococcus varius*, which the author later classified as green algae under *Protococcus botyoides f. caldaria* [24], while other authors called the algae *Pleurocapsa caldaria* [25]. A new genus of *Cyanophyta* was proposed and Geitler named the algae *Cyanidium caldarium* [26]. Others proposed that the algae should be transferred to the *Chlorophyta* [27], additionally, owing to morphological similarities, it was suggested that the algae could be transferred to the *Chlorella* genus [28]. The first description of an isolated algal species from the Cyanidiophyceae referred to one of these algae as *Pleurococcus sulphurarius* Galdieri. While this algae phenotypically appears green, its pigments are not the same as those of the Chlorophyta, and it was at this point that the new genus *Galdieria* was instituted, naming the species *Galdieria sulphuraria*. The nomenclature *Cyanidium caldarium* was retained for a strictly autotrophic algae, much smaller in diameter to *G. sulphuraria* (2-6 μm), with no vacuole. A third, club shaped, alga, which divides by longitudinal scission, was isolated. Similarly to *G. sulphuraria*, this alga is thermoacidophilic, however, unlike *G. sulphuraria*, it is strictly autotrophic and does not contain trienoic acids. This alga was named *Cyanidioschyzon merolae*. The presence of α-chorophyll and C-phycocyanin in the chloroplasts allowed authors to provisionally place these algae in the Rhodophyta, and establish a new class, the Cyanidiophyceae [29], and this is the class that is retained today.

At the time, not all authors accepted this new classification, with some retaining *C. caldarium* for both *C. caldarium* and *G. sulphuraria*, considering them different strains of the same species, or different species of the same genus (*Cyanidium*) [30][31]. Conversely, other authors instituted additional *Galdieria* species on the

basis of morphological characteristics, *G. partita* Sentsova, *G. daedala* Sentsova and *G. maxima* Sentsova [32]. Classifying these organisms based on morphology alone was challenging, and the later developments in DNA sequencing technologies brought welcome clarity to the taxonomy of the Cyandiophyceae.

As molecular phylogenetic studies became more accessible, the plastid encoded large ruBisCO subunit (rbcL) gene for *C. caldarium*, *C. merolae*, and several *Galdieria* species/strains was sequenced and the phylogenetic relationship inferred. This defined three genera *Cyanidium*, *Cyanidioschyzon*, and *Galdieria*, clearly placing the Cyanidiophyceae as sister taxa to the Rhodophytes, but rejected that the Russian *Galdieria* lineages (*G. partita* Sentsova, *G. daedala* Sentsova, and *G. maxima* Sentsova) were separate species, but instead different strains of *G. sulphuraria* [33]. More recently, a fourth genus has been added to the class. The species *Cyanidiococcus yangmingshanensis* was designated the same subclass as *Cyanidioschyzon merolae*, yet is morphologically distinct from the rod shaped *Cyanidioschyzon*, instead being subspherical (hence the designation -*coccus*). This work also rejected the nomenclature *G. partita*, *G. daedala*, and *G. maxima* as separate species, since these sequences were dispersed among the *G. sulphuraria* phylogeny [34]. New species of *Galdieria* and *Cyanidium* have also been identified and confirmed through phylogenetic analysis (*G. phlegrea* and *C. chilense*) [35][36].

There is continuing fluidity in the nomenclature. Currently, the nomenclature is as follows; the class Cyanidiophyceae consists of one order, the Cyanidiales, three families, four genera, and six species. The family Cyanidiaceae consists of one known genus and two known species; *Cyanidium caldarium* and *Cyanidium chilensis*. The family *Cyanidioschyzonaceae* consists of two known genera and two known species; *Cyanidioschyzon merolae* and *Cyandiococcus yangmingshanensis*. Finally, the family *Galdieriaceae* consists of one known genus and two known species, *Galdieria sulphuraria* and *Galdieria phlegrea*. While three other *Galdieria* binomials continue to be referenced in the literature (*G. partita*, *G. daedala*, and *G. maxima*) [11], these most likely belong to *sulphuraria*.

Having split from the other Rhodophytes 1.3 – 1.5bn years ago, during the meso-proterozoic eon, the early Cyanidiophyceae were contending with conditions vastly different to what modern-day ecosystems experience. Firstly, the earth during this

period was more tectonically active, and the hot, acidic, environments, in which all but one of the Cyanidiophyceae species currently reside, were widespread [37]. Secondly, while the Great Oxygenation Event, which resulted from the evolution of oxygenic photosynthesis in the cyanobacterial clade, causing fundamental alterations to atmospheric composition, had already taken hold. Atmospheric oxygen levels were low, and transient oxygenation events were continuing to take place. Organisms had to contend with fluctuating levels of oxidative stress and DNA damaging reactive oxygen species. Many obligate anaerobes were unable to cope and thus became extinct, whereas other anaerobes survived in ocean sediment refuges that remained predominantly anoxic. At that time, tolerating these increasing oxygen levels was an extremophilic trait [38].

Of the six Cyanidiophyceae species, only one species is not an extremophile. *C. chilense* is neutrophilic (pH 7.0) and mesophilic (20-25 ˚C). The ancestral state of the *Cyandiophyceae* is extremophilic, although there is continuing debate as to whether the entire Rhodophyte lineage was extremophilic, and only the Cyanidiophyceae retained the ability for extremophily, or if the ancestral Cyanidiophyceae gained extremophily later. The meso-neutrophilic *C. chilense*, colloquially known as “Cave Cyanidium” most likely lost its extremophilic capabilities as it moved into more moderate habitats [36]. Considering that the Cyanidiophyceae is an ancient lineage (1.5 billion years old), six known species is an unexpectedly low amount, and it has therefore been suggested that there is cryptic species diversity – either through species loss or species yet to be isolated/identified. A key feature of the Cyanidiophyceae phylogeny is that branches leading to the different lineages are quite long, whereas terminal branches are short [34][39][40][41]. This also supports the notion that genetic diversity was repeatedly lost, most likely due to population destruction in the volcanic areas in which these species inhabit, resulting in population bottlenecks.

Another possible explanation for low species diversity could be the dispersal mechanism, whereby only a few cells successfully disperse to new sites, again resulting in population bottlenecks in these sites. It has been suggested that cells, spores, and cysts could be transported by wind or birds [42]. The need for species to survive in mesophilic environments during dispersion also likely limits gene flow between physically distant extreme environments. In comparing the available

Cyandiophyceae phylogenies to their current geographical distribution, the dispersal mechanism and route is unclear. Species and isolates do not cluster clearly with location, nor do they cluster with bird migratory pathways or prevailing winds or currents. It may be that strain isolation occurred early during the evolutionary history of the Cyanidiophyceae, during the Proterozoic eon. This would offer some explanation towards the Cyanidiophyceae geopassport since several tectonic cycles occurred during that period. The genetic structure of *Galdieria* populations in Iceland suggests that that all Icelandic *Galdieria* lineages descend from a single, post glacial, colonisation event from north-eastern Asia [43]. Further work is needed to clarify *Cyandiophyceae* population structure and determine the dispersal mechanisms of these species.

### *Galdieria sulphuraria*

Uniquely placed within the *Cyandiophyceae* are the *Galdieria* species. These species grow in a diverse range of environments, and are the only species in the Cyanidiophyceae to grow both heterotrophically and phototrophically [44]. *Galdieria* species typically occupy more extreme habitats within these environments than other Cyanidiophyceae. For example, while most Cyandiophyceae occupy habitats exposed to sulphur fumes, *Galdieria* species also occupy endolithic and interlithic environments, and are exposed to a broader range of environmental fluctuations including desiccation, variation in acidity (up to pH 5-6), temperature (18 – 56 ˚C), and salinity (5-10%). In contrast, other Cyanidiophyceae reside in more protected environments and instead are found in streams and ditches nearby hot springs, which do not exhibit extreme temperature or pH fluctuations [41].

The genus *Galdieria* consists of two species, *Galdieria sulphuraria* and *Galdieria phlegrea*, that were not distinguishable before molecular phylogenetic studies. *G. phlegrea* prefers lower temperatures for growth (25 – 38 ˚C) than *G. sulphuraria* (up to 56 ˚C) [45]. In addition to these organisms being thermoacidophiles, both species have been isolated from the Rio Tinto, Spain, an anthropogenic site originating from a mine, exhibiting concentrations of heavy metals toxic to most species [46][47], and from burning coal spoil heaps in the Czech Republic, where *Galdieria* was isolated from vents exhibiting temperatures between 50 – 55 ˚C [48].

*G. sulphuraria* grows at temperatures considered the upper limit for eukaryotic life (56 ˚C). Its versatility, in both the range of habitats it can occupy and its metabolic flexibility makes it an extremely successful species, and it frequently represents up to 90% of the total biomass and almost all eukaryotic biomass in the environments it occupies [21][49]. Due to this, there has been much interest in investigating the potential industrial uses of *G. sulphuraria*.

The heterotrophic capabilities of *G. sulphuraria* have been of interest for the processing of food and agricultural waste streams, particularly for the degradation of lignocellulosic biomass under non-sterile conditions, which is of specific importance in rural agricultural settings that may not have access to closed bioreactors. *G. sulphuraria* was found to be able to utilise agricultural residues of a protein content ~40% [50]. Moreover, heterotrophic growth makes *G. sulphuraria* a potentially useful factory for the production of the phycobiliprotein C-phycocyanin, which has extensive uses in cosmetics, diagnostics, and foods [51][52]. C-phycocyanin is also produced phototrophically in the cyanobacterium *Arthrospira platensis* [53], termed "*Spirulina plantensis*" in the nutraceuticals industry. Purely phototrophic cultures suffer from low productivities, and it is thought increased cost and environmental efficiency can be achieved by growing *G. sulphuraria* in darkness on waste biomass for the production of C-phycocyanin [54][55]. As a result, there have been investigations to examine the application of *G. sulphuraria* in this area, which have shown that while *G. sulphuraria* C-phycocyanin showed similar antioxidant activities to *Spirulina* C-phycocyanin, the C-phycocyanin of *G. sulphuraria* showed excellent stability in heating and light, therefore C-phycocyanin produced by *G. sulphuraria* is an excellent alternative to using *Spirulina* for C-phycocyanin production [56].

*G. sulphuraria* grows in environments naturally rich in rare, heavy, and precious metals [57]. Consequently, it has been investigated for uses in biosorption, a more environmentally friendly, cost effective, and economical alternative to traditional metal recovery methods [58]. *G. sulphuraria* has demonstrated capability at removing rare earth elements from aqueous solutions [59], and there is additional potential at employing freeze dried *G. sulphuraria* biomass for this purpose [57]. *G. sulphuraria* has the ability to selectively remove gold and palladium ions from metal wastewater [60], and the toxic metals cadmium, copper, lead, and nickel from aquatic environments [61].

The prospective uses of *G. sulphuraria* in bioremoval are not solely limited to metals, and *G. sulphuraria* can remove ammoniacal nitrogen and phosphate from landfill leachate [62]. Overall, *G. sulphuraria* has a wide range of biotechnological applications and these are continuing to be investigate. Advances in genome sequencing technologies and the consequent improved understanding of the underlying biology of the species has furthered the development of *G. sulphuraria* as a model organism for use in biotechnology [63], and increasing the knowledge of *G. sulphuraria* genomics is hoped to improve understanding further.

## The History of Genomics

Genomics is a relatively new discipline. Sequencing technologies and the computational capacity required to analyse sequencing output did not exist until recently.

The three dimensional structure of DNA was famously characterised in 1953, by Watson and Crick [64] working from crystallographic data produced by Franklin and Wilkins [65]. While this paved the way for the concepts of DNA replication and encoding proteins from nucleic acids, methods to determine the sequence of DNA fragments were not established for some time. The first efforts focused on sequencing RNA, obtained from the most freely available populations of relatively pure RNA species, such as single-stranded RNA bacteriophages or microbial ribosomal or transfer RNAs. These could easily be bulk-produced in culture and are not complicated by a complementary strand and RNase enzymes that can cut RNA chains at specific sites were available. This enabled the composition of nucleic acid sequences to be analysed. With further advances, incorporating selective ribonuclease treatments, the first whole nucleic acid sequence was produced by Robert Holley in 1965, who sequenced alanine tRNA from *Saccharomyces cerevisiae* [66]. At the same time, Fred Sanger and his team developed a technique based on the detection of radiolabelled partial-digestion fragments after two-dimensional fractionation [67]. This allowed for additional transfer RNA sequences to be determined, and using this method Walter Fiers produced the first complete protein coding sequencing in 1972, that of the coat protein of bacteriophage MS2, followed four years later by its complete genome [68][69].

At this time, methods were being adapted to enable the sequencing of DNA, aided by the recent purification of bacteriophages with DNA genomes. Wu and Kaiser used DNA polymerase to fill the overhanging 5' end of the *Enterobacteriophage* λ with radioactive nucleotides, supplying each nucleotide one at a time and measuring incorporation [70][71]. This led to the development of primers, specific oligonucleotides that bind to the 5' overhang and "prime" the DNA polymerase. Radioactive nucleotides could then be used to infer sequence anywhere, not just 5' overhangs of bacteriophage genomes [72]. However, base determination was still restricted to short stretches of DNA and involved a lot of analytical chemistry and fractionation procedures. The next crucial practical amendment was the replacement of 2D fractionation with polyacrylamide gel electrophoresis, which separates polynucleotide fragments by size, with much greater resolving power. This technique was used in two protocols in the mid-1970s; Sanger's plus and minus system and Maxam and Gilbert's chemical cleavage technique [73].

The major development that changed sequencing in perpetuity was Sanger's chain-termination techniques, published in 1977 [74]. Now widely known as Sanger sequencing, this method involves the use of chain-terminating and either radioactively or fluorescently labelled dideoxynucleotides (which lack the 3' hydroxyl group required for chain extension) to sequence a DNA strand complementary to the template strand. Fragments were then separated by size and analysed using gel electrophoresis, and later, after improvement of the method, capillary electrophoresis. This was widely accepted as "first generation sequencing" and was a very labour intensive process. The first human genome was sequenced using this method and took over a decade to complete, costing over 2.7 billion USD [75].

Up until 2007, the cost of genome sequencing was falling in accordance with Moore's law, and the cost of sequencing a human genome was $10 million [76]. Due to a reduced requirement for labour and materials, the sequencing of shorter DNA fragments via Sanger sequencing was economically feasible for the inference of phylogenetic relationships, and Sanger sequencing was employed for the sequencing of the rbcL gene for the first Cyandiophyceae phylogenies. However, cost limitations remained for the sequencing of whole eukaryote genomes, which was not economically feasible until next generation sequencing methods became available, and the cost of sequencing plummeted [76][77]

The next improvement to sequencing came concurrently with the development of large scale Sanger sequencing and relied on a luminescent method for measuring pyrophosphate synthesis. As the nucleic acid chain is extended, pyrophosphate is released at each nucleotide addition. The amount of pyrophosphate released varies with each nucleotide. The pyrophosphate is first converted into ATP, which is then supplied to luciferase which emits light proportional to the amount of ATP, and consequently pyrophosphate, supplied. The advantage of this method is that heavily modified nucleotides such as dNTPs did not have to be used, and later sequencing could be observed in real time. However, this method faced a major difficulty with the sequencing of homopolymers [78][79]. Pyrosequencing was licensed to 454 Life Sciences, where it evolved to become the first major successful “Next Generation Sequencing” (NGS) technology, commonly known as 454 Sequencing. The system produced 400-500 bp read lengths and had the ability to sequence 400-600 million bp per run. 454 Life Sciences was purchased by Roche, and as sequencing technologies improved, Roche made the decision to stop supporting 454 from mid-2016 [73].

Between 2004 and 2006, various parallel sequencing methods were introduced. A common feature of these methods is massive sequencing of short (150-800bp) clonally amplified DNA molecules in parallel. Thermofisher’s Ion Torrent sequences single DNA fragments bound to beads via emulsion PCR then semi-conductor sequencing [80]. The Solexa method of sequencing, later acquired by Illumina, binds fragments to flow cell surfaces and they are amplified by “bridge-amplification” PCR and sequenced by an optical method involving fluorescently labelled reversible terminators. The advantage of short read sequencing methods is that large amounts of DNA or RNA can be sequenced extremely accurately, with at least over 70% of bases 99.9% accurate. This enables accurate variant calling, gene discovery, transcriptome assembly, and a number of other analyses to be performed. The major disadvantage of short read sequencing is that in the absence of a good reference genome, constructing chromosome level genome assemblies from short reads presents a major challenge. These methods are collectively known as short read or second generation sequencing.

Third generation technologies are based on different principles and can generate sequences > 10 kb directly from single molecules of native DNA. While earlier forms

of these technologies were less accurate at a nucleotide level, more recent improvements have much higher accuracy. The principle difference from second generation technologies is the sequencing of single molecules in real time [73]. Two primary long read sequencing technologies are currently in popular use; Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT). During PacBio sequencing, the DNA to be sequenced exists as single stranded circular DNA, which is replicated using an immobilised DNA polymerase and fluorescently labelled nucleotides, which when bound to the polymerase produce a light pulse, allowing for sequencing [81]. ONT sequencing instead feeds single stranded DNA through a protein pore, and an ion current is applied. The variation in ion flow through the pore is dependent on the nucleotide base [82]. The main advantages of long read sequencing is the ability to produce de-novo genome assemblies, and also to properly characterise structural variation in genomes. This comes at a cost of accuracy at a base level, although accuracy continues to improve [80]. The sequencing industry is highly mobile, and recently Illumina has announced its long read sequencing platform [83].

Coupled with the massively increased ability to sequence DNA has been an increase in the power and availability of computational resources required to analyse these massive amounts of data, however it would be incorrect to assume that the field of bioinformatics only came as a response to large scale DNA sequencing. The beginnings of bioinformatics occurred even before desktop computers and DNA sequencing. Margaret Dayhoff has been described as “the mother and father of bioinformatics”. Dayhoff combined her expertise in applying computers to biological problems with Robert Ledley, and together they developed COMPROTEIN, a computer program to determine protein primary structure using Edman peptide sequencing data [84]. Entirely encoded in FORTRAN on punch cards, this software is the first occurrence of what we know today as a de novo sequence assembler. Later, the concept of molecular evolution was starting to develop [85]–[87], and multiple sequence alignment algorithms such as CLUSTAL [88], which is still used to this day [89], were developed. Much of this still focused on amino acid sequences, but then came in paradigm shift from protein to DNA in the 1970s, alongside the development of Sanger sequencing, however there remained technical limits.

Prior to the 1970's, the "minicomputer" had the dimensions of a household refrigerator, rendering the acquisition of such a computer inaccessible to most work groups. Even the first desktop computers were not user friendly. In 1977, the first ready to use minicomputers came to market, and the development of specialised software for bioinformatics came rapidly. Richard Stallman published the GNU Manifesto in 1985, promoted the philosophy of free software – the idea that "the users have the freedom to run, copy, distribute, study, change and improve the software" [90]. This school of thought was at the centre of several initiatives in bioinformatics such as the European Molecular Biology Open Software Suite, which began to develop in 1996. It was also during this period that the European Molecular Biology Laboratory (EMBL), GenBank, and DNA Data Bank of Japan (DDBJ) united to standardise data formatting, to define the minimum amount of information for reporting nucleotide sequences, and to facilitate data sharing.

In the early 1990s, the Conseil Européen pour la Recherche Nucléaire (CERN) launched the World Wide Web [91], rendering the internet more accessible, revolutionising all aspects of society, and additionally led to the creation of widely accessible bioinformatics resources [92].

With the advent of next generation sequencing in the early 2000s, and the later reduction in DNA sequencing costs (due to the arrival of several parallel sequencing technologies as earlier described), there was an exponential increase in the amount of sequences in public databases. Major computational resources are now needed to handle this data, as well as store it. While in some cases, a simple laptop computer can suffice, many bioinformatics projects require much more imposing and expensive infrastructures, and many institutions now provide High Performance Computing resources to their scientists.

The "bioinformatics boom" has not been without its challenges. Many new algorithms have been developed to accommodate the flood of genomic data, yet this vast number of, what would seem to the untrained eye, redundant, tools lack standardised comparison methods, and potential software users lack an adequate guide for selecting the right tools for their data [93]. This is especially a problem when working with non-model organisms such as *G. sulphuraria*. Moreover, there remains a gulf between the average "biologist" and the average "developer",

although universities are attempting to bridge this gap by offering more courses on bioinformatics and big data analysis to biology undergraduates.

Genomics is a field that has been decades in the making. Since the sequencing of the first human genome in 2003, sequencing has undergone massive technological advancement that has resulted in a huge increase in data output combined with significant cost reduction – the cost of sequencing a complete human genome currently stands at around $1000 [76][77]. Coupled with the improvement of computational resources, this has enabled mass sequencing for clinical, biotechnological, and ecological purposes. The 100,000 genomes project sequenced 100,000 human genomes with the aim of finding variants linked with disease [94]. In clinical settings, whole genome sequencing has proven to be particularly useful in oncology for correctly identifying cancers and implementing more suitable treatment plans [95]. Sequencing is important in plant biotechnology for the identification of novel genes and improved crop strains, to help tackle the food crisis [96]. The Earth Biogenome Project aims to catalogue and characterise the genomes of all Earth's eukaryotic biodiversity, in order to provide a solid foundation to drive solutions for protecting biodiversity [97]. Genomics even became a household term during the COVID-19 pandemic, with government and health institutions using sequencing to track the mutation and spread of SARS-Cov-2 [98][99]. It is a versatile field, and in this thesis I will describe and examine the genomic data of *G. sulphuraria*.

## The Cyanidiophyceae Genomes

With the massively increased ability to sequence and analyse whole genomes, much work has been conducted on the Cyanidiophyceae genomes. As of October 2022, there were 19 Cyandiophyceae genomes available on GenBank, comprising five of the six known species, with 12 assemblies available for *G. sulphuraria*, 5 assembles for *C. merolae*, and one assembly each for *G. phlegrea*, *C. yangmingshanensis*, and *C. caldarium*. Of these genomes, only one has been completely resolved, which is that of *C. merolae*.

*C. merolae* has a nuclear genome size of 16.5Mb and 20 nuclear chromosomes. Published in 2004, this was the first complete eukaryotic genome assembly. It revealed a GC rich genome (55% GC), and was the smallest known genome for a photosynthetic eukaryote at the time. 30 of the 40 subtelomeres contained

duplicated 20 Kb sequence elements, and a unique telomeric repeat (GGGGGGAAT) was present. This is found on both ends of the chromosomes. 5,331 genes were identified, 86% of which were expressed. The genome has a high paucity of introns, with only 0.5% of the protein coding genes containing introns, with all but one containing only a single intron, which were shown to have strict consensus sequences [100]. For comparison, 5% of all yeast genes and 79% of *Arabidopsis* genes contain introns [101]. Owing to the completeness of this genome, it remains in use as a point of reference and comparison for the analysis of other Cyandiophyceae genomes today.

The *C. merolae* genome contains a relatively limited gene inventory when compared to other free-living algae or eukaryotes not belonging to the Rhodophyta. This compact gene inventory is seen throughout the Rhodophytes, and is indicative of extensive genome reduction in the common ancestor of red algae. Analysis of 14 red algal genera revealed two major phases of genome reduction, with the first phase in the stem lineage resulting in the loss of major functions such as flagellae and basal bodies, the glycosylphosphatidylinositol anchor biosynthesis pathway, and the autophagy regulation pathway, and the second phase in the common ancestor of the extremophilic Cyandiophyceae [102].

The loss of conserved genes in large numbers is usually explained by a considerable change in lifestyle (from free living to pathogenic, for example), or adaptation to a different environment (oligotrophic or extreme) [103]. The loss of single conserved pathways is a widespread phenomenon in other lineages, for example with loss of flagella based motility among Viridiplantae, fungi, and microbial eukaryotes [104]. What makes the initial Rhodophyte genome reduction notable is that a collection of functions were lost that are normally conserved in free-living lineages, therefore implying selection for streamlined genomes [103]. The reason for this initial, extensive, gene loss is unclear, but it has been suggested that it may have occurred in response to the environment becoming nutrient deficient. An oligotrophic environment such as a biofilm would reduce the need for many of the pathways that were lost in the initial phase of gene loss, such as flagellum based motility, lowered reliance on signalling mechanisms and GPI-anchor biosynthesis, and selected for a reduced cytoskeleton [11]. The second wave of gene loss, in the common ancestor of the Cyanidiophyceae, coincides with a shift to extremophily, and also coincided

with the recruitment of bacterial genes through horizontal gene transfer, replacing their corresponding eukaryotic homologs [102].

## Horizontal Gene Transfer

Horizontal gene transfer (HGT) is the movement of genetic information between organisms in the same generation, as opposed to vertical inheritance where organisms receive their genetic information from their parent/s [105]. In bacteria and archaea, this process is widely accepted and understood as an important evolutionary driver [106], [107]. This phenomenon is so widespread among prokaryotes that it has been doubted whether we can be confident that prokaryotic phylogenetic trees properly reflect the biological reality [16], [108]. In eukaryotes, outside of endosymbiotic or pathogenic gene transfers, the situation has been less clear cut.

HGT genes can be identified by considering a number of factors, including the phylogenetic position of the gene – is it more closely related to bacterial/archaeal homologs rather than algal homologs? Gene composition is another important factor, for example genes acquired horizontally from bacteria and archaea are likely to have fewer introns, higher GC content, and differing codon usage when compared with genes that have evolved vertically within the eukaryotic lineage in question [109].

The first whole genome sequencing of a *G. sulphuraria* isolate, strain 074W, originally isolated from Java, Indonesia, unveiled a 13.7 Mb genome, and revealed that *G. sulphuraria* possesses a number of genes that have putatively been acquired through horizontal gene transfer from bacteria and archaea [110]. While gene transfers could be traced to a variety of donor taxa, a significant number of transfer events are traced to extremophilic bacteria, and are likely responsible for some of *G. sulphuraria's* extreme tolerance. Two monovalent cation:proton antiporters, that may be partially responsible for salt stress resistance, were found to have been acquired from bacteria. Additionally, genes encoding dimethylglycine methyltransferase (SDMT), which allow for the production of compatible solute betaine from glycine which accumulates in *G. sulphuraria* under salt stress, may have originated from halophilic cyanobacteria. *G. sulphuraria*'s resistance to toxic heavy metals may also be attributed to HGT, with two genes lacking introns encoding the bacterial arsenical efflux pump found in the genome, as well as mercuric reductase, which can reduce

cytotoxic $Hg^{2+}$ into less toxic metallic mercury, likely acquired from Proteobacteria [110]. Overall, it was reported that 5% of *G. sulphuraria* genes could have been acquired by HGT.

The role of HGT in *G. sulphuraria*'s sister taxon, *G. phlegrea*, was investigated using a 11.4 Mb draft genome. This revealed that *G. phlegrea* had regained the previously lost urea hydrolysis pathway from eubacteria. These genes were unlinked, indicating that they were acquired in several transfer events. Altogether, these findings implied that extensive genome reduction, which is a common outcome in eukaryotes for adaptation to a specialised niche, can be relieved by the gain of previously lost, or novel, functions through HGT [35].

The possibility of HGT in eukaryotes has been hotly debated. Some authors rejected the assertion that *Galdieria* species have undergone extensive HGT, and instead hold that since the last eukaryotic common ancestor likely evolved in somewhat extreme conditions, and that the genes responsible for extremophily were lost as organisms moved into more moderate habitats [111]. This theory, known as differential loss, affirms strict vertical inheritance on all genes except for transfers of pathogenic origin and endosymbiosis. Differential loss proponents have assessed eukaryote genomes for HGT and have concluded that, while transfers from endosymbiotic ancestors of chloroplasts and mitochondria are uncontroversial and are well explained by what we understand of eukaryote biology, direct HGT events from prokaryotes to eukaryotes are vastly overestimated in the available literature [111]. Claims of this type of HGT were put forward in the human genome sequence (concerning transfers from the gut microbiome), but were quickly refuted [112], [113]. Authors claim that should HGT be occurring in eukaryotes, that like prokaryotes, we should be able to detect recent and ancient HGT, and that these recent HGT products should have high sequence identity with their donors, as is seen in prokaryote genomes, where amino acid sequence of recent HGT products can be 100% identical with its donor. Instead, for prokaryote to eukaryote HGT, it has been observed that there are no HGT candidates with over 70% predicted amino acid sequence identity with their prokaryotic donors, and that genes with higher predicted amino acid sequence identity over 70% are probably artifacts or contaminants, rather than recent gene transfers [114].

Nevertheless, the evidence for HGT from prokaryotes in the *G. sulphuraria* genome is compelling, with genes identified in the *G. sulphuraria* genome that are not found elsewhere in the eukaryotic lineage, such as the mercuric reductase and arsenic efflux pump [110]. The possibility of additional horizontally acquired genes cannot be ruled out either, since eukaryotes may employ mechanisms of incorporating gene transfers into the genome that rapidly make genes (and proteins) more eukaryote compatible (in terms of codon adaptive index, GC content, and protein folding), resulting in a reduction in amino sequence identity [115].

The 70% rule, which had imposed an upper cut-off for predicted amino acid sequence identity to prokaryote donors in putatively horizontally acquired genes [114], has been since challenged, as more exceptions are found, including in the *G. sulphuraria* genome. Rossini et al. identified 18 orthogroups have been identified in the *Galdieria* genome that have over 70% predicted amino acid identity with their prokaryotic donors, and cannot be explained by contamination, endosymbiotic gene transfer, or annotation/assembly artifacts. Therefore there is concern that strictly applying the 70% rule could lead to the removal of true positives [115].

It has also been possible to search for evidence of cumulative effects within *Galdieria* HGT candidates, that is the reduction of amino acid sequence identity in less recent HGT candidates. Orthogroups with fewer species are more likely to be recent HGT products, as the gene may have entered the lineage later when species or strains had already been isolated from one another. It was found that orthogroups with fewer species had higher amino acid sequence identity with their potential non eukaryotic donors, than HGT orthogroups with more species. Moreover, when compared with “native” orthogroups of the same size, amino acid sequence identity to non-eukaryotic donors or homologs was significantly higher in the HGT orthogroups [115].

The exact mechanism by which *Galdieria* has taken up DNA horizontally remains unknown, however the consensus is that these organisms have acquired genes horizontally, and that many of these genes are implicated in the extremophilic characteristics of *G. sulphuraria* [116]. Horizontal gene transfer therefore represents a mechanism by which *G. sulphuraria* has adapted to its extreme environment,

however it is not the only mechanism of adaptation apparent in the *G. sulphuraria* genome.

### The Organellar Genomes of *G. sulphuraria*

Sequencing of the *G. sulphuraria* nuclear genome revealed HGT as a partial mechanism for extremophily, but *G. sulphuraria* also has a mitochondrial genome and a plastid genome. The Rhodophytes are one of three ancient lineages of photosynthetic eukaryotes derived from the primary endosymbiosis event that established the plastid, and with the Cyanidiophyceae having diverged from the Rhodophytes over 1 billion years ago, making up the only extremophilic phototrophs, the Cyanidiophyceae plastid genomes were investigated and compared to the mesophilic Rhodophytes in order to understand the effects of extremophily on the plastid [44].

It was found that the *G. sulphuraria* plastid genome is typical in size when compared with other red algal plastid genomes, and does not show any unusual strand specific skews of gene distribution or nucleotide frequency. It is a 167,741 bp circular genome with 224 intronless genes encoding 158 proteins with known functions. The *C. merolae* plastid genome is smaller at 149,987 bp, with 207 genes [44]. Despite *G. sulphuraria's* ancient divergence, there is a high level of collinearity between the *G. sulphuraria* and *C. merolae* plastid genomes, with both genomes sharing several syntenic gene blocks [44]. This indicates that there is selection for the maintenance of gene order in the plastid genome. It has been suggested that conserved gene order in the Cyanidiophyceae plastid genomes is due to extremophily [117], demonstrating an additional mechanism by which these species tolerate extreme environments.

On the other hand, the *G. sulphuraria* mitochondrial genome has been shown to exhibit a variety of interesting features. The mitochondrial genome of *G. sulphuraria* 074W is 21,428 bp in size, with a GC content of 43.9%. It also has a high mitochondrial gene strand skew (0.88) and 2 sets of tandem repeats. The *G. sulphuraria* mitogenome has the smallest size, the fewest genes and introns, the highest gene strand skew, the highest GC content, and the greatest proportion of repeats when compared with all other known red algal genomes [44]. Comparative analyses reveal that a number of genes have been lost from the *G. sulphuraria*

mitochondrial genome, and were likely transferred to the nuclear genome – although this was not without difficulties since in addition to the accelerated evolution of the *G. sulphuraria* mitogenome, *G. sulphuraria* diverged from other Cyanidiophyceae hundreds of millions of years ago and therefore there is a lack of closely related gene homologs for similarity searching. This is a challenge that is pervasive in the study of *G. sulphuraria* gene function across the organellar and nuclear genomes.

Another interesting feature observed in the *G. sulphuraria* mitogenome is extreme GC skew. Chargaff's second parity rule states that complementary nucleotides are at approximately equal frequencies within a single strand of DNA. The *G. sulphuraria* mitogenome is extremely G rich and pyrimidine poor compared with other red algae, and exhibits the highest genome wide GC skew among all eukaryotic mitogenomes, and the highest AT skew among all non-metazoans. This suggests an excess of purines of the forward strand and an excess of pyrimidines on the reverse strand [44]. Conversely, GC and AT skew is usually inversely correlated in other eukaryotes.

The mitochondrial genomes of the Cyanidiophyceae have been divided into two categories – G type, or C-type. G-type, is the smaller "*Galdieria*" type mitogenomes described above. This is only found in *Galdieria* species. The C-type mitogenome is the larger *Cyanidium* type mitogenome, around 10 Kb larger than G-type mitogenomes and found in *Cyanidioschyzon*, *Cyandiococcus*, and *Cyanidium* species. Not only are C-type mitogenomes larger by approximately 10 Kb, they are also more gene dense, containing over double the number of genes and fewer non coding regions. C-type genomes do not have significant AT or GC skews.

C-type and G-type species can not only be recognised by their mitogenomes, but also on the basis of morphological characteristics and ecological habitats. G-type cells are typically larger (5-10 μm compared with 1-4 μm), have a vacuole, and a simple spherical shape, whereas C-type species have more diverse morphologies. G-type species have several mitochondria per cell that have a branched structure, whereas reported C-type species contain a single mitochondrion per cell [41].

The ecological habitats of G-type species are much more diverse and include hydrothermal regions, acid mine drainage sites, and endolithic environments, whereas C-type species reside in more ecologically "stable" environments such as

ditches and streams surrounding hot springs that exhibit lower temperature variations. The existence of G-type and C-type mitogenomes is also supported by phylogenetic analysis, which revealed extraordinarily long internal branches of the G-type, further support the implication of cryptic or extinct species diversity within the Cyandiophyceae, but further specifies this to G-type lineages [41].

### *G. sulphuraria* population structure

As previously discussed, the phylogenetic structure of the Cyandiophyceae had been an area of confusion for decades, with species being repeatedly misclassified. Genome sequencing has been of great benefit to solving these classification problems. Phylogenetic analyses using partial ribulose-1,5-bisphosphate carboxylase (rbcL) gene fragments led to the hypothesis of diverging clades within the species [118]. Six *G. sulphuraria* clades were identified using the rbcL phylogeny, and it was demonstrated that *G. phlegrea* sits outside of the *G. sulphuraria* clade [119]. However, because different genes face varying evolutionary pressures, single gene phylogenies may not always be representative of the whole genome. This is especially pertinent in the case of rbcL, since it is plastid encoded and therefore may not represent the evolutionary pathway of the nuclear or mitochondrial genomes, especially if the organism is sexual [46].

Whole genome sequencing allowed for construction of the nuclear, plastid, and mitochondrial genome phylogenies based on the pangenomes of each respective genome, that is the genes that are shared across all *G. sulphuraria* isolates for each genome. The phylogenies for all three genomes supported the 6 + 1 *Galdieria* are incongruent. Incongruence between the plastid, mitochondrial, and nuclear genomes suggests these genomes have evolved differently. Moreover, the majority of single gene phylogenies were found to be distinct [120]. The differing single gene and pangenome tree topologies could be explained by frequent recombination.

## Furthering the understanding of the underlying mechanisms of extremophily

Whole genome sequencing has so far revealed that horizontal gene transfer has facilitated adaptation of *G. sulphuraria* to certain extreme conditions, such as the presence of toxic metals [110]. It has demonstrated that the organellar genomes, particularly the mitochondrial genome, exhibit extreme features including GC skew

and a reduced gene complement [44], which may be implicated in the metabolic flexibility of *Galdieria* species as well as extremophily. Whole genome sequencing has also enabled the phylogenetic relationships of the Cyanidiophyceae to be resolved [120], uncovering the potential for lost or unidentified species within the lineage.

Many of the genes identified in the *G. sulphuraria* genome have unidentified functions, as they are either too diverged from identified proteins in the available databases, or have evolved de novo in *G. sulphuraria*. These so called dark-genes could further explain *G. sulphuraria's* extremophilic traits [11]. This demonstrates that there is still a lot to be learned about the *G. sulphuraria* genome.

While some effort has been placed in identifying how the Cyanidiophyceae, particularly *G. sulphuraria*, manage to survive in extreme environments, this research is still fairly limited when compared to that of prokaryotic microbial communities. Analysis of prokaryotic extremophilic microbial communities has revealed evolutionary patterns that support the ability of these species to survive in extreme habitats [19]. In this thesis, I seek to apply genomics based approaches to the further understanding of the secrets of the extremophilic lifestyle of *G. sulphuraria*.

# Chapter 2: Assembly, Annotation, and Comparison of Complete *G. sulphuraria* Nuclear Genomes

## Introduction

Understanding the genome of *Galdieria sulphuraria* is central to understanding the underlying biology of the species. A number of *Galdieria* genomes are published on GenBank [1][2], which estimate the genome size to be between approximately 12Mb and 15Mb (*Table 1*). These genomes have uncovered some mechanisms for extremophily in *G. sulphuraria*, particularly that horizontal gene transfer has facilitated the evolution of this extreme eukaryote [3][4], and uncovered the global *Galdieria* population structure [120]. However, while these supply useful annotations for phylogenetic analysis and potential gene discovery, they tell us little about the wider genome structure, as many these genomes have not been assembled to completion, nor has the biological impact of these draft assemblies been discussed.

High quality genome data would be particularly useful for understanding *Galdieria* biology. In the absence of a pre-existing reference genome, 2nd generation short read sequencing faces a major challenge in the construction of long contigs [123]. Additionally, having completed genome assemblies enables the undertaking of a variety of robust downstream analyses, including, but not limited to, macro and micro synteny analysis, variant calling, and the calculation of the substitution rate.

From the perspective of population structure, it has been shown that there is a large amount of diversity within the *G. sulphuraria* species, with the *G. sulphuraria* nuclear, mitochondrial, and plastid genomes splitting into 6 clear lineages, indicating that isolates within each lineage may have been separated from isolates of other lineages for thousands if not millions of years [120]. I hypothesise that there must be an impact of this divergence on genome structure, and therefore it is important to evaluate the long read sequencing data of a variety of *G. sulphuraria* isolates that are representative of the six *G. sulphuraria* lineages.

There have been attempts to quantify the number of chromosomes in *G. sulphuraria* using pulse field gel electrophoresis, and 40 chromosomes were estimated [47]. These studies additionally reported a genome size of 9.8 Mb, indicating that *G. sulphuraria* possesses many tiny chromosomes considering the small genome size. Since then, advances in DNA sequencing have enabled the construction of some

draft genome assembles. The 2016 *G. sulphuraria* SAG 107.79 assembly resolved into 117 contigs with a contig N50 of 134Kb [123]. Although this assembly was incomplete with a huge number of sequencing errors, it provided a useful example of how Oxford Nanopore Technologies (ONT) sequencing can be applied to the assembly of whole genomes. In contrast, the NCBI reference genome for *G. sulphuraria* 074, originally isolated from Java, Indonesia, sequenced using Sanger and 454 sequencing, is assembled into 433 contigs [110]. More recently, a number of PacBio genomes have been made available [4][8], yet there has been no wider discussion about the *G. sulphuraria* genome structure, nor the impact of ONT sequencing in this field.

The construction of complete *Galdieria* genomes is critically important for the examination of how chromosomal architecture supports the adaptive, flexible, and versatile life capacity of this species. Here I describe the resequencing and reassembly of *G. sulphuraria* SAG 107.79, as well as the sequencing and assembly of 2 further *G. sulphuraria* strains, ACUF 138 and ACUF 017 into complete genomes using ONT sequencing.

| Strain | Accession | Submitter | Total Sequence Length (Mb) | Submission Date | Contig N50 (Kb) | Number of Contigs | Assembly Method | Coverage | Sequencing Technology |
|---|---|---|---|---|---|---|---|---|---|
| 074W | GCF_000341285.1 | *Galdieria* sulphuraria Genome Project | 13.7 | 25/02/2013 | 117 | 518 | Arachne v. 3.0 | 8x Sanger; 10x 454 | Sanger; 454 |
| 107.79 | GCA_001704855.1 | The University of York | 12.1 | 14/08/2016 | 134 | 117 | Minimap/miniasm/nanopolish | 50x | ONT MinION |
| YNP5578.1 | GCA_006232335.1 | Heinrich Heine Universitaet Duesseldorf | 14.2 | 11/06/2019 | 171 | 115 | Canu v.1.5 | 53x | PacBio RSII |
| SAG21 | GCA_006232365.1 | Heinrich Heine Universitaet Duesseldorf | 14.1 | 11/06/2019 | 158 | 135 | Canu v.1.5 | 39x | PacBio RSII |
| Az2 | GCA_006232395.1 | Heinrich Heine Universitaet Duesseldorf | 13.8 | 11/06/2019 | 165 | 127 | Canu v.1.5 | 46x | PacBio RSII |
| MtSh | GCA_006232405.1 | Heinrich Heine Universitaet Duesseldorf | 14.7 | 11/06/2019 | 187 | 101 | Canu v.1.5 | 102x | PacBio RSII |
| 5572 | GCA_006232475.1 | Heinrich Heine Universitaet Duesseldorf | 14 | 11/06/2019 | 188 | 108 | Canu v.1.5 | 56x | PacBio RSII |
| 002 | GCA_006232505.1 | Heinrich Heine Universitaet Duesseldorf | 13.8 | 11/06/2019 | 189 | 107 | Canu v.1.5 | 67x | PacBio RSII |
| MS1 | GCA_006232515.1 | Heinrich Heine Universitaet Duesseldorf | 14.6 | 11/06/2019 | 172 | 129 | Canu v.1.5 | 63x | PacBio RSII |
| RT22 | GCA_006232545.1 | Heinrich Heine Universitaet Duesseldorf | 15.3 | 11/06/2019 | 182 | 118 | Canu v.1.5 | 91x | PacBio RSII |
| UTEX2919 | GCA_019693475.1 | New York University AbuDhabi, UAE | 14.6 | 19/08/2021 | 3.83 | 9211 | ABySS v. 2.1.5 | 100x | PacBio |

*Table 1: Available G. sulphuraria assemblies on GenBank* [122].

## Methods

### Strain Preparation

*G. sulphuraria* isolates 138, 017, 427, and 074, originally isolated from El Salvador, the Phlegraean Fields, Italy, Gunnhuver, Iceland, and Java, Indonesia respectively, were obtained from the Algal Collection of University of Naples (ACUF) [125]. *G. sulphuraria* 107.79, originally isolated from Sonoma County, California, USA, was obtained from the Culture Collection of Algae (SAG) at Göttingen University [126]. *G. sulphuraria* isolate THAL033, isolated from Geng Zi Peng, Taiwan, was obtained from Tung-Hai Algal Lab Culture Collection [127]. All strains were isolated from a single colony obtained after streaking the culture across agar plates, respectively, and colonies were inoculated in Allen medium pH 1.5 [28] and cultivated at 37°C under continuous fluorescent illumination of 45 µmol photons·$m^{-2}$·$s^{-1}$.

### DNA Extraction

Illumina Sequencing

After a culture was grown to stationary phase 2 ml of a given culture was centrifuged (5 minutes, 13.2 rpm) and the algal pellet was resuspended in 40 µl PBS pH 7.5 and vortexed to mix. The samples were centrifuged and resuspended in PBS a further two times. Tubes were placed in a dry ice ethanol bath for 30 seconds then transferred to a 30˚C water bath, this was repeated 3 times.

From there, 500 µl of DNA extraction buffer 1 was added and incubated for 30 minutes at 55 ˚C, inverting every 10 minutes. Then 150 µl of DNA extraction buffer 2 was added and the samples incubated for a further 10 minutes at 65 ˚C. Finally, 650 µl of phenol:chloroform:isoamyl alcohol 25:24:1 was added. To this, 1 mm silica beads were added (0.5 cm of the tube), the samples were mixed by inversion, and mixed on a bead beating machine for 5 minutes. After centrifugation to clarify this solution, 600 µl of the supernatant was taken into a fresh tube and 480 µl of cold isopropanol was added. The samples were stored at -20 ˚C for a minimum of 2 hours. Next, samples were centrifuged at 15000 g for 30 minutes at 4 ˚C, and the supernatant discarded. 200 µl of ethanol was added then centrifuged at 13.2 rpm for 5 minutes and the supernatant discarded. The tubes were dried at room temperature, then resuspended in 30 µl of Tris-EDTA.

Samples were incubated with 1 µl of RNAse A and 1 µl of Proteinase K for 2 hours at 37 ˚C, then purified using the Qiagen DNeasy Plant Mini Kit. DNA quality and concentration were assessing using a Nanodrop photospectrometer ND-1000 (Thermo Fisher Scientific).

Nanopore Sequencing

After a culture was grown to stationary phase, 2 ml of a given culture was centrifuged for 5 minutes at 13.2rpm 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 1µl of Protienase K, 100µl of Viscozyme™ were added and incubated for an  hour at 37˚C. Then 40µl of PBS pH 7.5 added and vortexed to mix. 500µl of DNA extraction buffer 1.1 was added and incubated at 55˚C for 30 minutes, mixing by inverting every 10 minutes. Then 150µl of DNA extraction buffer 2 (2.1 or 2.2 dependent on strain) was added and incubated for a further 10 minutes at 65˚C. Next 690µl of Phenol:Chloroform:Isoamyl Alcohol 25:24:1 was added and mixed gently though inversion for 5 minutes. This was centrifuged at 13.2rpm for 5 minutes and 600µl of the top layer of the supernatant was then taken and placed into a fresh tube. Here 480µl of isopropanol was added and samples stored at -20˚C for 2 hours. Following this, samples were centrifuged at 15g for 30 minutes at 4˚C, supernatant then discarded. 200µl of 70% ethanol was added and then centrifuged at 13.2rpm for 5 minutes, supernatant discarded. Finally, tubes were air dried and then DNA re-suspended in 40µl of TE buffer.

Clean-up of DNA samples were completed using Zymo DNA Clean & Concentrator™-25 kit, the method was as follows: DNA binding buffer (DNA Binding Buffer: DNA sample) were added to DNA samples in a ratio of 2:1 and mixed briefly by vortexing. The mixture was transferred to a Zymo-Spin™ Column in a Collection Tube and was centrifuged for 30 seconds at 13.2rpm. The flow-through was discarded. 200µl DNA Wash Buffer was added to the column and centrifuged for 30 seconds at 13.2 rpm and the flow-through discarded. This wash step was repeated. It was then again centrifuged at 13.2rpm for 30 seconds and 40µl DNA Elution Buffer at 65˚C was slowly added directly to the column matrix and incubated at room temperature for ten minutes. The column was then transferred to a 1.5ml microcentrifuge tube and centrifuged at for 30 seconds at 13.2rpm to elute the DNA.

This step was repeated once. DNA quality and concentration were assessed using a Nanodrop photospectrometer ND-1000 (Thermo Fisher Scientific).

DNA yields varied from strain to strain. For example, for isolate THAL 033 buffer 1 with 34.5% Methanol then adding buffer 2.2 was best. For isolates 074 and ACUF 017 buffer 1.1 and then buffer 2.2 gave best results. For isolate ACUF 427 buffer 1.1 then buffer 2.1 gave the best yields.

| **DNA Extraction buffers** | | |
|---|---|---|
| **Buffer 1.1** | **Buffer 2.1** | **Buffer 2.2** |
| 200mM Tris-HCl pH8 | 200mM Tris-HCl pH8 | 100mM Tris-HCl pH8 |
| 200mM NaCl | 200mM NaCl | 700mM NaCl |
| 100mM LiCl | 100mM LiCl | 20mM EDTA pH8 |
| 25mM EDTA pH8 | 25mM EDTA pH8 | 2% CTAB |
| 1M Urea | 1M Urea | 0.0125mM PVP-40 |
| 1% SDS | 1% CTAP | |
| 1% NP-40 | 100mM Lithium acetate | |

*Table 2: DNA Extraction Buffers*

**DNA Sequencing**

For Illumina sequencing, library preparations were conducted 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 [128] and assembled using Spades v3.1 [129].

For MinION (Oxford Nanopore Technologies) sequencing, DNA libraries were prepared by shearing DNA in a Covaris, and then generating libraries with Oxford Nanopore Technologies SQK-006 library preparation kit using the standard ONT protocol (Oxford Nanopore Technologies), as per manufacturers' instructions. This was then run on a R7.3 flow cell (ONT) for 36 hours.

**RNA Preparation and Extraction**

*G. sulphuraria* cultures 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 with 10 g/L Sucrose at pH 2, in Allen medium with 0.5% Cellulose (w/v), 0.5% Xylan (w/v) and 0.5% Laminarin (w/v) at pH 2. Samples were collected by centrifugation at 1h, 12h, 96h, 192h and 336h. Pellets were washed 3 times in PBS buffer (pH 7.5) and samples stored at -80 ˚C before RNA extraction.

Cells were ground into a fine powder with a pestle and a drill in the presence of liquid nitrogen. RNA was isolated and cleaned up using the Monarch Total RNA Miniprep Kit (New England BioLabs, T2010S). RNA quality and concentration was assessed using a Nanodrop photospectrometer ND-1000 (Thermo Fisher Scientific). All RNAs were treated with DNaseI and then pooled by strain relative to the concentration of each sample. RNA quality and concentration was then assessed using the Agilent 2100 Bioanalyzer (Agilent Technologies).

RNA library preparation and sequencing were performed at Novogene (UK) Company Limited (Cambridge). Library preparation was performed using NEB Next® Ultra™ 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 250-300bp 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 150bp PE sequencing mode.

**Assembly of Oxford Nanopore Technologies (ONT) Sequencing Reads**

Oxford Nanopore Technologies MinION reads were basecalled with guppy 4.0.11 [130] with options --flowcell FLO-MIN106, --kit SQK-LSK108, --trim_strategy dna and --trim_barcodes.

Three draft assemblies were generated from each set of nanopore reads with the assemblers Canu2.1 [131], Raven v1.5.3 [132], and SMARTdenovo [133]. Canu2.1 was run with options genomeSize=13m and -fast. SMARTdenovo was run in consensus mode, -c 1.

Each draft assembly was polished once with medaka v1.3.3 [134], and polished three times with Pilon v1.23 [135], using the Illumina reads mapped to the assembly using the Burrows-Wheel Aligner v0.7.17 [136].

Assemblies were assessed with Tapestry v1.0.0 [137], aligned to each other by minimap2 v2.20 [138], and edited in Biopython. For ACUF138, contigs with less than 10% unique material were removed from each polished assembly. The final assembly was compiled from contigs from the Canu2.1 and SMARTdenovo assemblies based on the following list of factors: did the reads end at the contig ends? Do the contigs have both telomeres? Are there similar contigs in all three assemblies? Is the nanopore and Illumina read coverage uniform across the contig, without any breakages?

**Genome Annotation**

Transcript assemblies were constructed using Trinity using both de-novo and genome guided modes [139]. RNA sequencing reads to their respective Illumina and ONT assemblies using the splice aware aligner STAR v. 2.7.3 [140] using the defaults. At STAR index generation, --genomeSAindexNbases was determined based on STAR genomeGenerate recommendations, and varied with each genome dependent on genome size, with values from 10-12. The RNA sequencing reads aligned to the Illumina assembly were used for Trinity in genome guided mode. These Illumina assemblies were generated using SPAdes [129], see Iovinella and Lock [120] for more details. Annotations were predicted with funannotate [141], run without coding quarry, using the eukaryota database (fetched 02/03/2021) for functional predictions [142]. InterProScan v 5.46-81.0 [143] was run separately.

**Identification of Reverse Gyrases**

Homologs of *Sulfolobus acidocaldarius* reverse gyrase (GenBank ID L10651.1_prot_AAA72346.1_1) were identified using BLASTp from BLAST+ v.2.13.0 [144], with an expect value cut off of 0.001. The predicted amino acid sequence of each homolog was then searched on the NCBI database using BLASTp [117][118][141], to confirm a predicted topoisomerase type 1Ac domain [146] was present in each homolog. A multiple sequence alignment of the *G. sulphuraria* putative reverse gyrases and reverse gyrases from *S. acidocaldarius, Thermococcus*

*kodakaraensis,* and *Pyrococcus furiosus*, was generated using MUSCLE v. 3.8.1151 [147].

### Macrosynteny Analysis

Genome wide collinearity within and between genomes was detected with MCScanX v. 2020.10.23 [148], by first utilising BLASTp from BLAST+ v. 2.13.0 [144] to detect homologous protein coding genes. MCScanX was employed using the default settings. The expect value cut-off for BLASTp was set to $1e^{-50}$. The resulting colinear blocks were viewed in Circos. For the mapping of *G. sulphuraria* SAG 107.79 to *Cyanidioschyzon merolae* 10D, minimap2 v2.20 [138] with option -x asm20 was employed.

### High Performance Computing

The Viking High Performance Computing Cluster [149] was used for the computational analysis for this project.

## Results

### Oxford Nanopore Technologies (ONT) Sequencing Read Sets

| Strain | Reads | Bases (bp) | Coverage* | Read N50 (bp) |
|---|---|---|---|---|
| 017 | 318,457 | 1,814,334,535 | 140 | 8,010 |
| 033 | 3,033,324 | 12,337,260,831 | 949 | 5,661 |
| 074W | 2,545,765 | 10,711,941,319 | 824 | 5,785 |
| 107.79 | 369,895 | 4,340,924,435 | 334 | 20,391 |
| 138 | 2,766,336 | 19,136,249,663 | 1472 | 10,619 |
| 427 | 228,137 | 1,374,843,850 | 106 | 8,804 |

*Table 3: Overall ONT read information for all sequenced G. sulphuraria isolates. Coverage calculated assuming a 13 Mb genome.*

| Strain | Reads | Bases | Coverage* | Read N50 |
|---|---|---|---|---|
| ACUF 017 | 8,848 | 167,239,756 | 13 | 17,853 |
| THAL 033 | 8,003 | 138,522,527 | 11 | 16,789 |
| 074W | 14,846 | 262,071,929 | 20 | 17,071 |
| SAG 107.79 | 112,948 | 2,863,050,807 | 220 | 25,766 |
| ACUF 138 | 277,086 | 5,914,750,772 | 455 | 20,909 |
| ACUF 427 | 8,004 | 145,069,171 | 11 | 17,337 |

*Table 4: ONT read information for reads >15 Kb for all sequenced G. sulphuraria isolates. Coverage calculated assuming a 13 Mb genome.*

To produce complete, telomere to telomere, assemblies, longer reads are preferred. Typically, a read coverage of 40x is required to produce assemblies of this quality. For 15Kb < reads, only the isolates SAG 107.79 and ACUF 138 have this level of read coverage. This is likely due to DNA being sheared during the extraction procedure for the other four isolates. Therefore, for the purpose of this thesis, I focused on SAG 107.79 and ACUF 138 for telomere-to-telomere assemblies. Aside from 074W, which already has published assemblies and annotations, ACUF 017 had the next best read statistics, with slightly higher coverage than ACUF 427 and THAL033, and was the third assembly I focused on for completion, with ACUF 427 and THAL 033 remaining as draft assemblies.

### *G. sulphuraria* SAG 107.79 Assembly: Early Signs of Something Strange

Since the 2016 *G. sulphuraria* SAG 107.79 assembly [123], further investigations and improvements on the assembly of this genome had been made, although never published. A SMARTdenovo assembly with 74 chromosomes had been constructed. It was reported that *G. sulphuraria* chromosomes contained shared regions, with ONT sequencing reads mapping to multiple chromosomes within the assembly. This added another layer of complexity in assembling the *G. sulphuraria* genome. In the SMARTdenovo assembly, the redundant regions resulting from shared chromosomal regions were omitted as they would not contain novel genes. This, however, did not reflect the biological reality. I re-introduced the redundant regions to this assembly through mapping the chromosomes to one another to find the overlaps, then manually stitching the chromosomes together and checking this against the Illumina assemblies. This produced a 14.2 Mb assembly, with 74 contigs and N50 203 Kb.

### *G. sulphuraria* 138 Assembly: Manual Assembly Construction Based on Many Assemblers

Polished assemblies produced using Canu2.1, SMARTdenovo and Raven varied in length and quality (*Table 5*). The Canu2.1 assembly contained 209 contigs, many of which were short and did not contain unique genetic material, and these were removed by keeping only contigs with more that 10% unique material, reducing the number of contigs to 98. While the Raven and SMARTdenovo assemblies were similar in size, and not drastically different in the number of contigs, the contig N50 value for the Raven assembly was 20 Kb lower than the SMARTdenovo assembly, and most of the Raven contigs were not resolved telomere to telomere. As a result,

the final assembly was based on the SMARTdenovo assembly, and was supported by contigs from the Canu2.1 assembly, and alignments to the Raven assembly as described below.

| | All Contigs | | | | Contigs with >10% Unique Material | | |
|---|---|---|---|---|---|---|---|
| | **Number of Contigs** | **Assembly Length (bp)** | **Contig N50 (bp)** | **Longest Contig (bp)** | **Number of Contigs** | **Assembly Length (bp)** | **Contig N50 (bp)** |
| **Canu2.1** | 209 | 19,986,028 | 163,204 | 492,458 | 98 | 15,745,391 | 195,223 |
| **SMARTdenovo** | 93 | 16,894,060 | 221,349 | 385,618 | 74 | 15,682,667 | 221,438 |
| **Raven** | 96 | 16,625,650 | 201,073 | 376,192 | 84 | 15,865,013 | 203,757 |

*Table 5: Performance of different genome assemblers at assembling the G. sulphuraria ACUF 138 ONT sequencing reads.*

After mapping the ACUF 138 SMARTdenovo assembly to the SAG 107.79 assembly, 72 SMARTdenovo contigs successfully mapped to contigs from the completed 107_2021 assembly (minimap2 mapping quality = 60), but the contigs were not completely collinear between the two isolates. This was examined further

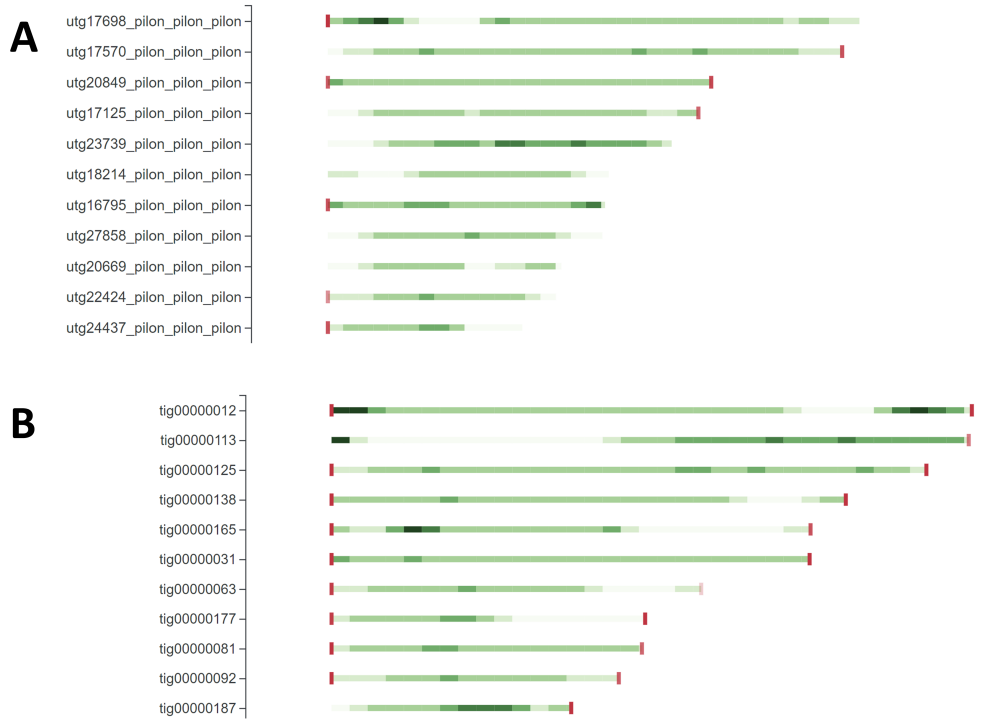



*Figure 1: Tapestry ideograms of G. sulphuraria ACUF 138 SMARTdenovo contigs (A) and the Canu2.1 contigs that replaced them (B). Dark green indicates higher read coverage. Telomeres are shown in red.*

with the completed assemblies. Each SMARTdenovo contig had alignments to Canu2.1 and Raven contigs. SMARTdenovo contigs with both telomeres, and with similar contigs in the Canu2.1 and Raven assemblies, were retained in the final assembly, with the corresponding Canu2.1 and Raven contigs discarded. This was done to ensure that the same contigs were predicted by all assemblers, therefore making it less likely that an individual contig is an assembly artefact, and instead correctly reflects the chromosomal architecture of the isolate. An example of an assembly artefact that was omitted from the final assembly is shown in *Figure 4*, which is discussed in more detail further in this section.

SMARTdenovo contigs with absent telomeres were checked for alignments to Canu2.1 contigs and replaced with the respective Canu2.1 contig if it contained the absent telomere(s), as this increased the number of telomere to telomere chromosomes in the final assembly, but did not introduce incorrectly assembled contigs, as the respective SMARTdenovo and Canu2.1 contigs were completely aligned except for the missing telomeric material.

SMARTdenovo contigs with missing telomeres and no complete Canu2.1 were retained after inspection, as removing them resulted in a severe reduction in the number of genes found after genome annotation (*Table 6*). This resulted in a final assembly with 73 contigs, 50 resolved telomere to telomere, 20 with one telomere, and 3 with no telomeres. 11 contigs were assembled with Canu2.1, the remaining 52 contigs were assembled with SMARTdenovo.

| **Canu2.1 Contig** | **SMARTdenovo Contig** | **Reason** |
|---|---|---|
| tig00000012 | utg17698 | Telomeres |
| tig00000031 | utg20849 | Length |
| tig00000063 | utg27858 | Telomeres |
| tig00000081 | utg16795 | Telomeres |
| tig00000092 | utg22424 | Telomeres |
| tig00000113 | utg23739 | Telomeres and Length |
| tig00000125 | utg17570 | Telomeres |
| tig00000138 | utg17125 | Telomeres |
| tig00000165 | utg18214 | Telomeres |
| tig00000177 | utg24437 | Telomeres |
| tig00000187 | utg20669 | Telomeres |

*Table 6: G. sulphuraria ACUF 138 Canu2.1 and SMARTdenovo sequence IDs for replaced contigs, and the reason for the replacement of the SMARTdenovo contig.*

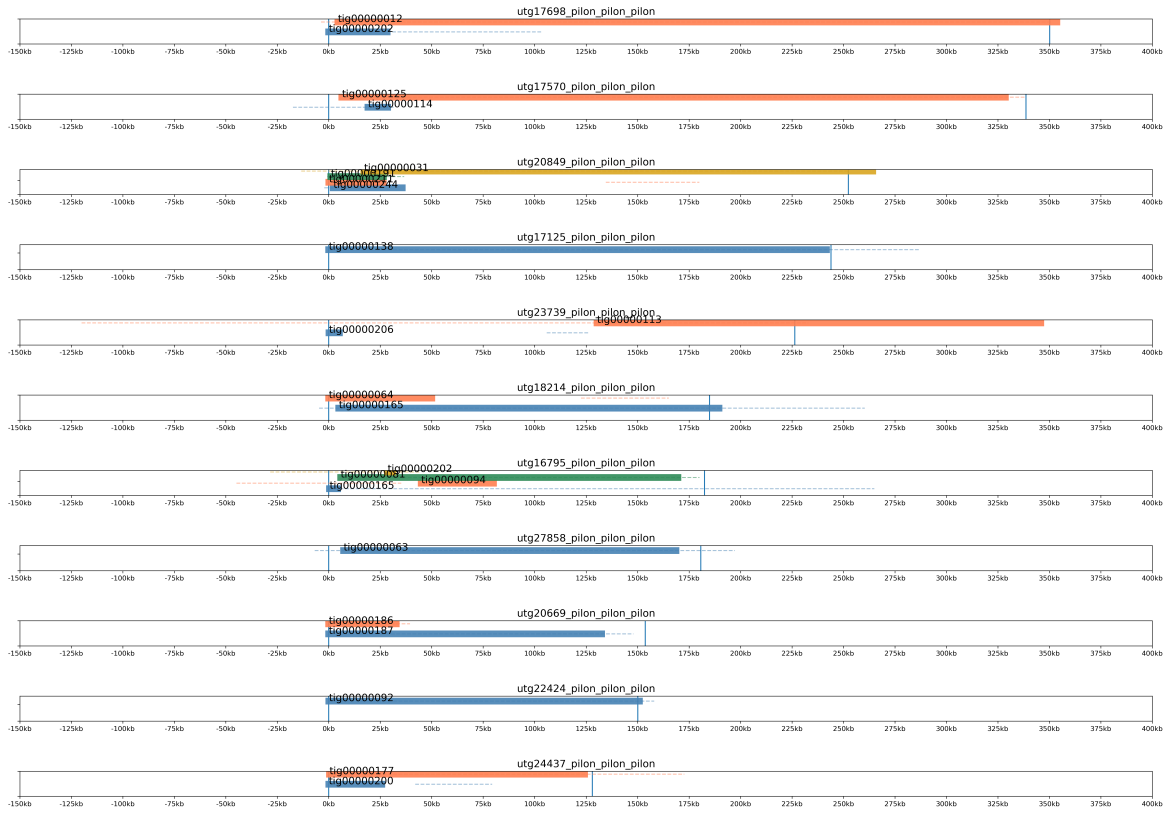



*Figure 2: A visualisation of minimap2 alignments of the G. sulphuraria ACUF 138 Canu2.1 contigs to the SMARTdenovo contigs.*

As demonstrated by the minimap2 alignments (*Figure 2*), the Canu2.1 contigs that replaced SMARTdenovo contigs were all similar size to their corresponding SMARTdenovo contig, with the exception of Canu2.1 tig00000113, which shares sequence with the SMARTdenovo contig utg23739, but does not map to the entire contig and contains some unique sequence. Canu2.1 tig00000113 was retained in the final assembly instead of SMARTdenovo utg23739 due to tig00000113 containing a telomere not present in utg23739.

While the ONT read alignments for each contig were checked for irregularities, it was found that tig00000113 contained clipped reads on the 5' end, which align to no other region of the nuclear genome. I suspected that these reads possibly belonged to an accessory genome, and after conducting a BLAST search, two plastid genes were found on tig00000113 with the loci tig00000113:3443-3946 and tig00000113: 4106-4839, where the clipped reads are mapped (*Figure 3 C + D*). This section of the contig was most likely assembled in error as Canu2.1 failed to assemble these reads to the plastid genome. The alignment of tig00000113 to utg17863 is an

example of chromosome dovetailing in *G. sulphuraria* ACUF 138. There is redundancy between these contigs that may be a result of recombination. SMARTdenovo collapses this region into a single haplotype, however I elected to retain the shared region over two separate contigs since this enables more uniform read coverage and the annotation of different haplotypes, better reflecting the biology of the species.

Canu2.1 appears to manage assembling these repeats into separate contigs, hence the high number of redundant contigs in the raw Canu2.1 assembly, whereas SMARTdenovo treats these regions as haplotigs and collapses them into one contig, however this has resulted in the omission of some unique regions in the SMARTdenovo assembly.

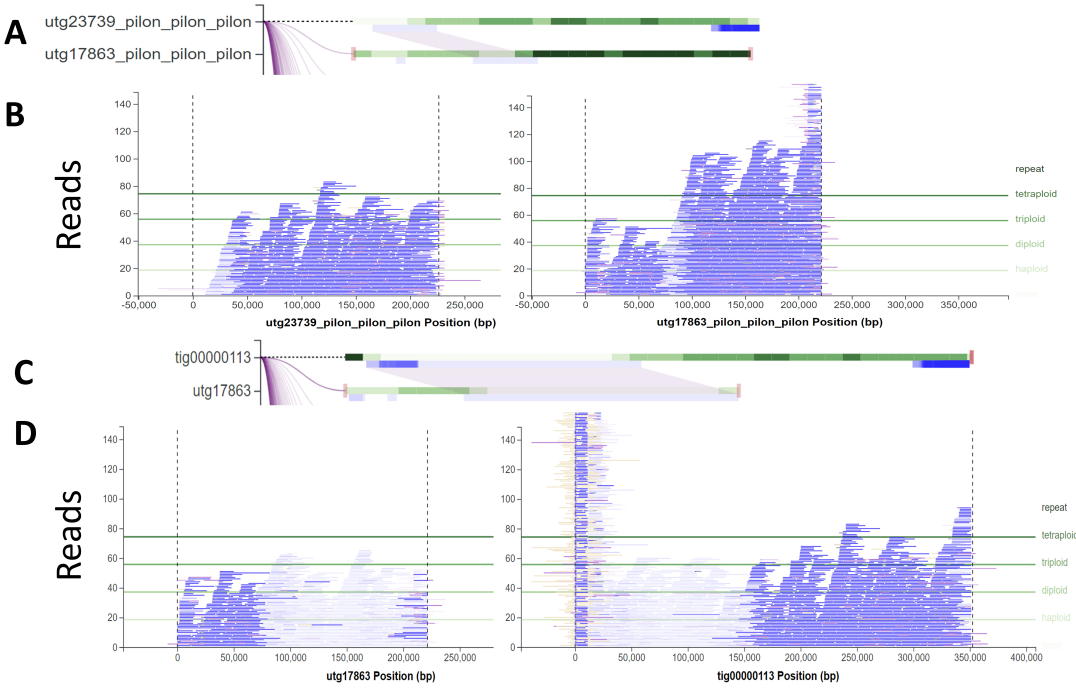



*Figure 3: A demonstration of shared regions between G. sulphuraria ACUF 138 contigs through highlighted alignments between Tapestry ideograms (A + C), and the corresponding ONT read alignments for these contigs (B+D).*

For each assembly method (Canu2.1, SMARTdenovo), the longest contig in the SAG 107.79 assemblies was approximately 500 Kb, whereas only the Canu2.1 assembly of ACUF 138 resolved a contig of similar length. The longest Canu2.1 ACUF 138 contig, tig00000044, contains material that SMARTdenovo resolved into two contigs, utg18934 and utg19167. Both of these SMARTdenovo contigs have both telomeres. Mapping the ONT reads to tig00000044 (*Figure 4B*) shows that only

reads that map to other contigs bridge between where SMARTdenovo utg18934 ends at 385,618bp, and that ONT read coverage falls in this region. Mapping of the Illumina reads to this region shows a complete reduction in read coverage (*Figure 4C*). Thus it was subsequently decided that Canu2.1 tig00000044 was erroneously assembled, and the SMARTdenovo contigs utg18934 and utg19167 were retained instead.

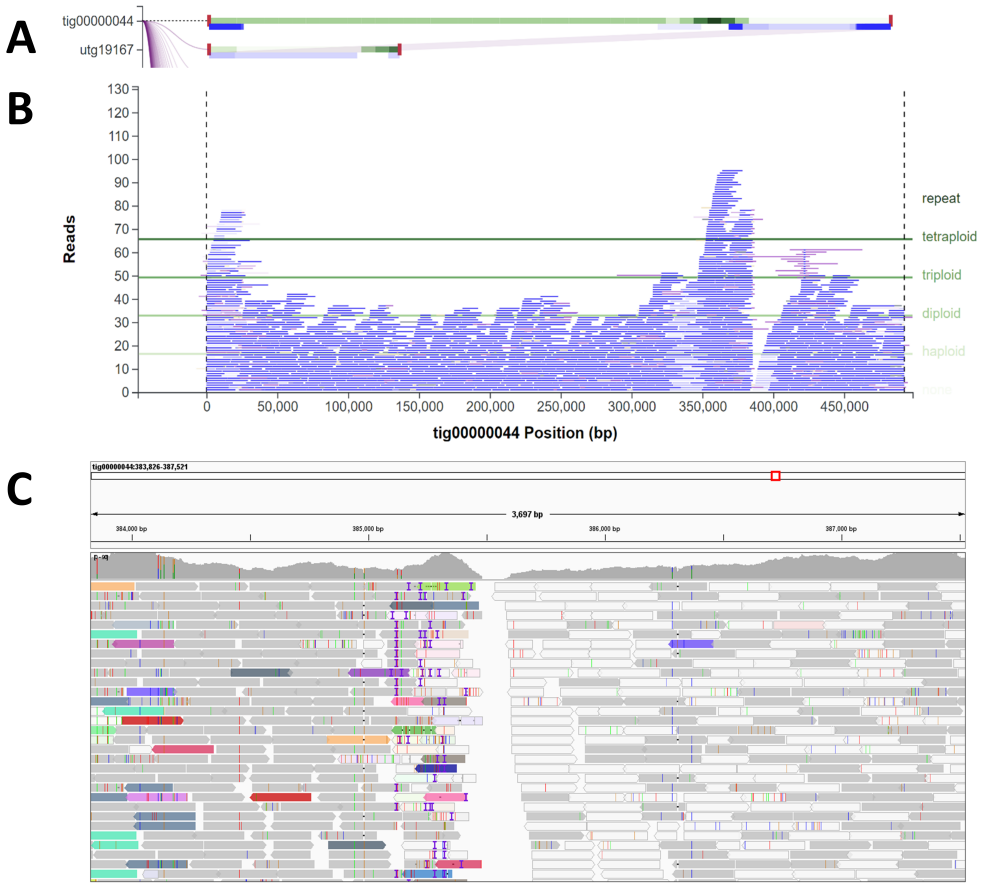



*Figure 4: A) ACUF 138 Canu2.1 tig00000044 Tapestry ideogram aligned to SMARTdenovo utg19167. B) ONT read alignments to tig00000044, showing a severe reduction in read coverage at the contig break point. C) Illumina read alignments to tig00000044, showing no read coverage over the contig break point.*

In order to assess completeness of the assemblies, ensuring the right balance between achieving telomere to telomere chromosomes and the number of genes, each assembly was fully annotated. By all measures, the final assembly outperforms the other assemblies. While the polished SMARTdenovo assembly contains the most genes, this is largely due to sequence redundancy. The filtered SMARTdenovo assembly, created by removing all contigs with <10% unique material, created a

good assembly to work as a base, but had fewer genes and fewer complete telomeres than the final assembly. The draft hybrid assembly, which utilised Canu2.1 tig00000044 as the longest contig, and contained no chromosomes missing any telomeres, suffered a loss of nearly 1000 genes compared to the final assembly, as a result of being 2,181,083 bp shorter.

| Assembly | Polished SMARTdenovo Assembly | Filtered SMARTdenovo Assembly | Draft Hybrid Assembly | Final Hybrid Assembly |
|---|---|---|---|---|
| **Assembly Size** | 16,894,060 bp | 15,682,667 bp | 13,769,060 bp | 15,950,143 bp |
| **Largest Scaffold** | 385,618 bp | 385,618 bp | 492,458 bp | 385,618 bp |
| **Average Scaffold** | 181,657 bp | 211,928 bp | 218,557 bp | 218,495 bp |
| **Num Scaffolds** | 93 | 74 | 63 | 73 |
| **Scaffold N50** | 221,349 bp | 221,438 bp | 221,438 bp | 221,800 bp |
| **Percent GC** | 38.69% | 39.08% | 39.06% | 39.04% |
| **Num Genes** | 6,731 | 6,264 | 5,465 | 6,404 |
| **Num Proteins** | 6,561 | 6,216 | 5,421 | 6,357 |
| **Num tRNA** | 170 | 48 | 44 | 47 |

*Table 7: Assembly and annotation statistics for G. sulphuraria ACUF 138 draft and final assemblies.*

The *G. sulphuraria* ACUF 138 assembly and its methods of construction as described here, demonstrate that constructing a hybrid assembly using both SMARTdenovo and Canu2.1 contigs can produce an improved chromosome level assembly.

**Improving the *G. sulphuraria* SAG 107.79 Assembly**

Based on the principle that Canu2.1 contigs can support a SMARTdenovo assembly in order to achieve a more complete genome, as demonstrated with *G. sulphuraria* ACUF 138, I revisited the *G. sulphuraria* SAG 107 assembly. This time, the 107_2021 assembly was used to train medaka to polish the raw Canu2.1 and SMARTdenovo assemblies. After polishing, and once redundant contigs had been removed, this left a SMARTdenovo assembly of 72 contigs, 14.1 Mb in length, N50 206 Kb and a Canu2.1 assembly of 75 contigs, 13.6 Mb in length, N50 187,599.

After inspection of the alignments of the Canu2.1 and SMARTdenovo assemblies, six SMARTdenovo contigs were replaced by Canu2.1 contigs. All Canu2.1 contigs replaced SMARTdenovo contigs in order to resolve the second telomere, except for tig00000113, which replaced the SMARTdenovo contig for extra length, and only has 1 resolved telomere.

| **Canu2.1 Contig** | **SMARTdenovo Contig** | **Reason** |
|---|---|---|
| Tig00000014 | Utg2533 | Telomeres |
| Tig00000010 | Utg892 | Telomeres |
| Tig00000090 | Utg1219 | Telomeres |
| Tig00000069 | Utg1747 | Telomeres |
| Tig00000019 | Utg875 | Telomeres |
| Tig00000113 | Utg4247 | Length |

*Table 8: G. sulphuraria SAG 107.79 Canu2.1 and SMARTdenovo sequence IDs for replaced contigs, and the reason for the replacement of the SMARTdenovo contig.*

### *G. sulphuraria* ACUF 017 Assembly

Despite the ONT 15 Kb < read coverage being lower than ideal for *G. sulphuraria* ACUF 017, the construction of a complete assembly was attempted using the same methods and principles applied to that of *G. sulphuraria* ACUF 138 and *G. sulphuraria* SAG 107.79. The replaced contigs are shown in *Table 9*.

| Canu2.1 Contig | SMARTdenovo contig | Reason |
|---|---|---|
| Tig00000323 + tig000000324 | Utg26 | Telomeres |
| Tig000000234 + tig00000062 | Utg399 | Telomeres |
| Tig00000309 + Tig00000310 | Utg205 | Telomeres |
| Tig00000224 | Utg364 | Telomeres |
| Tig00000229 | Utg201 | Length |
| Tig00000015 | Utg2 | Telomeres + Length |
| Tig00000066 | Utg134 | Telomeres |
| Tig00000273 | Utg1452 | Length |
| Tig00000111 | Utg487 | Length |
| Tig00000251 | Utg388 | Telomere |
| Tig00000106 | Utg529 | Telomeres |
| Tig00000133 | Utg178 | Telomeres + Length |
| Tig00000138 | Utg1078 | Telomeres |
| Tig00000121 | Utg206 | Length |
| Tig00000327 | Utg429 | Length |
| Tig00000295 | Utg229 | Length |

*Table 9: G. sulphuraria ACUF 017 Canu2.1 and SMARTdenovo sequence IDs for replaced contigs, and the reason for the replacement of the SMARTdenovo contig. The instances where there are two Canu2.1 contigs added together indicated that the final contig was formed by manually stitching the Canu2.1 contigs together while viewing read alignments.*

**Draft Assemblies of *G. sulphuraria* ACUF 427 and THAL 033**

The *G. sulphuraria* ACUF 427 ONT sequencing reads were assembled with SMARTdenovo, and then polished as per the methods. Contigs with less than 10% unique material were removed, and this assembly was annotated. A draft SMARTdenovo assembly of *G. sulphuraria* THAL 033 was constructed, which had

145 contigs, a length of 15.8 Mb, an N50 of 196.1 Kb, and a longest contig of 354.3 Kb.

| ***G. sulphuraria* ACUF 427 Draft Assembly Statistics** | |
|---|---|
| **Number of Contigs** | 78 |
| **Total Length** | 1273875 bp |
| **Longest Contig** | 397773 bp |
| **Contig N50** | 188434 bp |
| **Mean Contig Length** | 163293.27 bp |
| **Number of Genes** | 5406 |
| **% Complete BUSCOs** | 90.1% |
| **% GC** | 37.97% |

*Table 10: G. sulphuraria ACUF 427 nuclear genome assembly and annotation statistics*

**Three Complete *G. sulphuraria* Genome Assemblies and Annotations**

The completed genome assemblies for *G. sulphuraria* SAG 107.79, ACUF 138 and ACUF 017 reveal a highly compact genome with an unusually large number of chromosomes compared with the genome size. The genome size ranged from 13 Mb to 16 Mb and the number of contigs was 72-73, strain dependent. The genome size is consistent with previously reported *G. sulphuraria* genome sequences. Two strains, *G. sulphuraria* SAG 107.79 and *G. sulphuraria* ACUF 138, have complete BUSCOs > 90%, consistent with the assemblies that include the majority of genes. Of the two, *G. sulphuraria* SAG 107.79 is the most complete genome by measure of the number of telomere-to-telomere chromosomes. Although *G. sulphuraria* ACUF 138 has a higher scaffold N50 value and a slightly higher % complete BUSCOs, these differences can be attributed to the larger genome size of *G. sulphuraria* ACUF 138, at 15.95 Mb compared to 14.28 Mb in *G. sulphuraria* SAG 107.79.

The *G. sulphuraria* ACUF 017 assembly has fewer telomere-to-telomere chromosomes, a lower N50 value and complete BUSCOs < 90%. This is due to lower quality raw sequencing data for this sample, with low coverage of ONT reads longer than 15 Kb that would be required to bridge over the numerous subtelomeric regions in the *G. sulphuraria* genome assembly.

| | *G. sulphuraria* Assemblies | | | Closest Complete Genome | *G. sulphuraria* NCBI |
|---|---|---|---|---|---|
| | **ACUF 138** | **SAG 107.79** | **ACUF 017** | ***Cyanidioschyzon merolae* 10D** | **074W** |
| **Assembly Size (Mb)** | 15.95 | 14.28 | 13.14 | 16.52 | 13.78 |
| **Largest Scaffold (bp)** | 385618 | 500147 | 349111 | 1621617 | N/A |
| **Average Scaffold (bp)** | 218495 | 198347 | 182544 | 826015 | 31824 |
| **Num Scaffolds** | 73 | 72 | 72 | 20 | 433 |
| **Scaffold N50 (bp)** | 221800 | 209122 | 196143 | 850100 | 172100 |
| **% GC** | 39.04 | 40.19 | 39.06 | 55.00 | 36.89 |
| **Num Genes** | 6404 | 5975 | 5567 | 5331 | 6723 |
| **Num Proteins** | 6357 | 5920 | 5523 | 5010 | 6622 |
| **Num tRNA** | 47 | 55 | 44 | 30 | 127 |
| **% Genes of Unknown Function** | 21.9 | 19.62 | 20.94 | N/A | N/A |
| **Gene Density** | 2491 | 2390 | 2361 | 3099 | 2050 |
| **Introns Per Gene** | 2.83 | 2.93 | 2.92 | 0.005 | 1.26 |
| **Mean Gene length** | 1592 | 1598 | 1571 | 1552 | N/A |
| **% Coding** | 54.99 | 57.66 | 57.34 | 44.9 | N/A |
| **% Complete BUSCOs** | 90.5 | 90.1 | 88.5 | N/A | N/A |
| **Contigs Telomere-Telomere** | 50 | 62 | 18 | N/A | N/A |
| **Contigs with 1 Telomere** | 20 | 10 | 38 | N/A | N/A |
| **Contigs with No Telomeres** | 3 | 0 | 16 | N/A | N/A |
| **ONT Read Coverage > 15Kb*** | 455 | 220 | 13 | N/A | N/A |

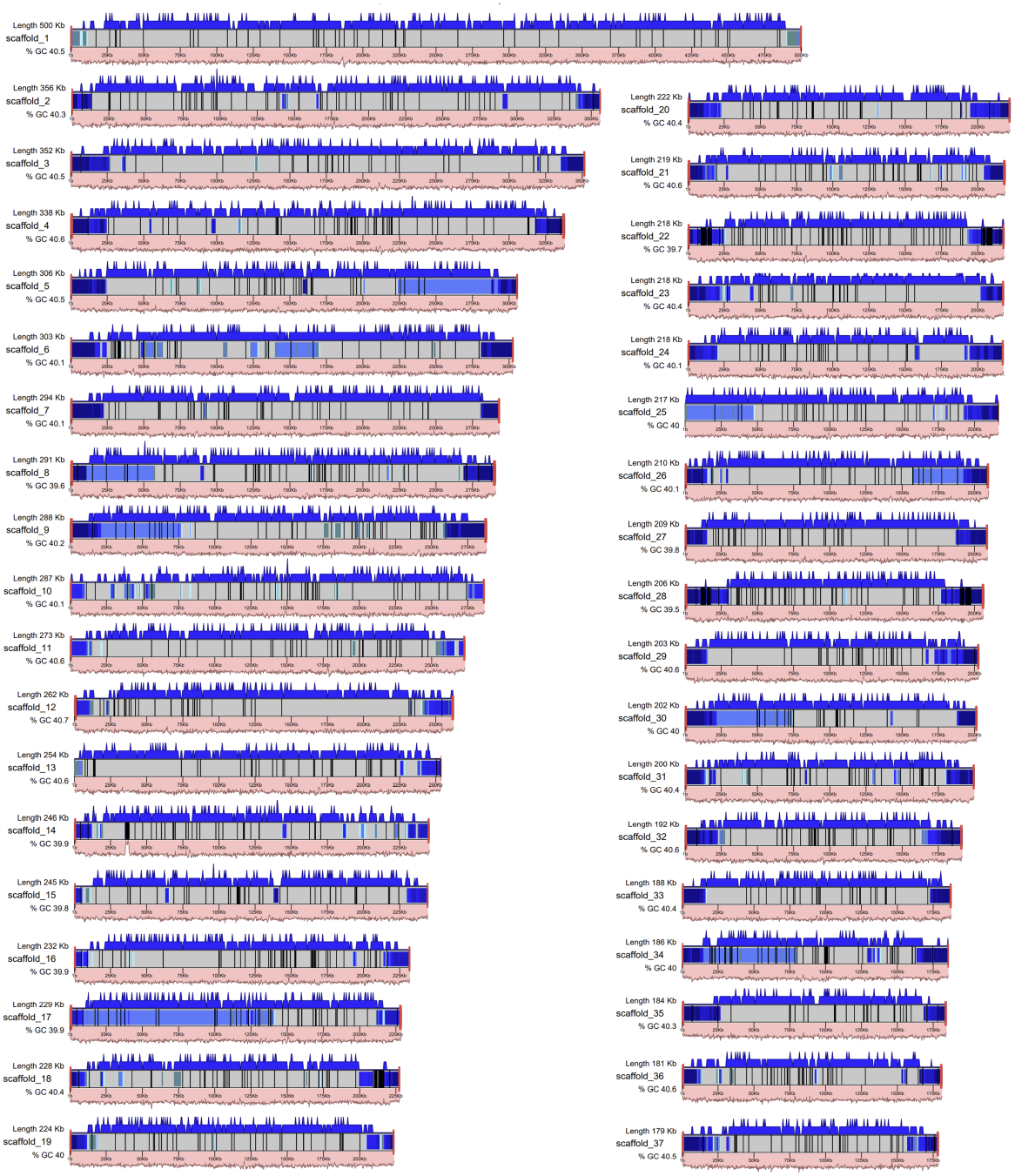
*Table 11: Assembly and annotation statistics for the three final G. sulphuraria assemblies, alongside C. merolae, the closest complete genome, and G. sulphuraria 074, the NCBI reference genome.*

Comparisons of the *G. sulphuraria* SAG 107.79 and *G. sulphuraria* ACUF 138 assemblies reveal important structural differences. *G. sulphuraria* ACUF 138 chromosome length is more uniform than that of *G. sulphuraria* SAG 107.79. The longest contig in *G. sulphuraria* ACUF 138 is 385.6 Kb, and the shortest is 131.4 Kb. *G. sulphuraria* SAG 107.79 has a longer longest contig, SAG 107.79:scaffold_1, at 500.1 Kb, and a shorter shortest contig at 62.6 Kb, although this contig is missing a

telomere. The shortest telomere to telomere contig for *G. sulphuraria* SAG 107.79 is 84.8 Kb. SAG 107.79:scaffold_1 is longer than the next longest chromosome by 144.0 Kb.

As shown in *Table 11*, not every contig is telomere to telomere. To demonstrate the completion of these chromosomes, subtelomeric material was identified through all by all chromosome mapping and these are shown in the dark blue regions in *Figure 5* and *Figure 6*. A good example shown below is 138_scaffold_66, which is missing a telomere, but has strong evidence of subtelomeres plus reduced gene density at the chromosome end, indicating that by missing the telomeric sequence, no unique genes are absent.

# The complete *G. sulphuraria* SAG 107.79 nuclear genome



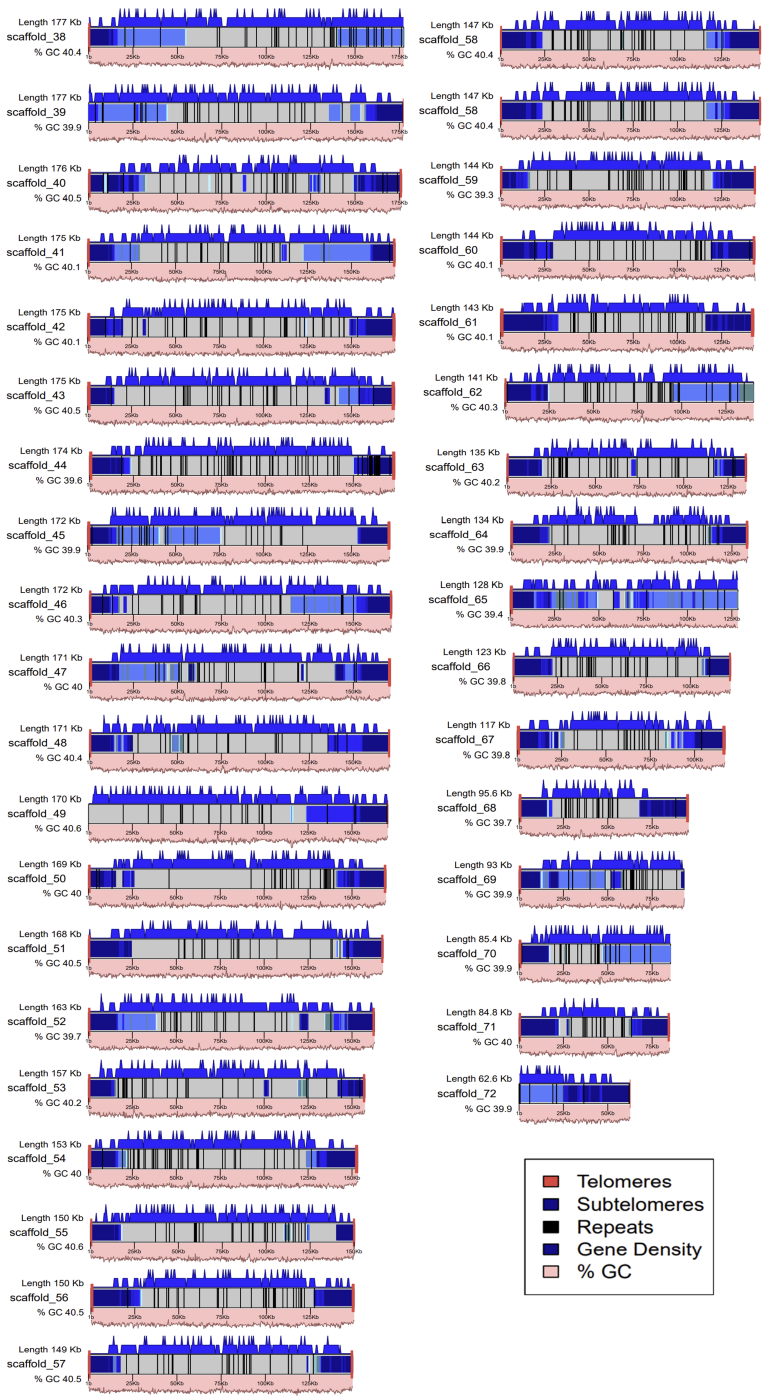
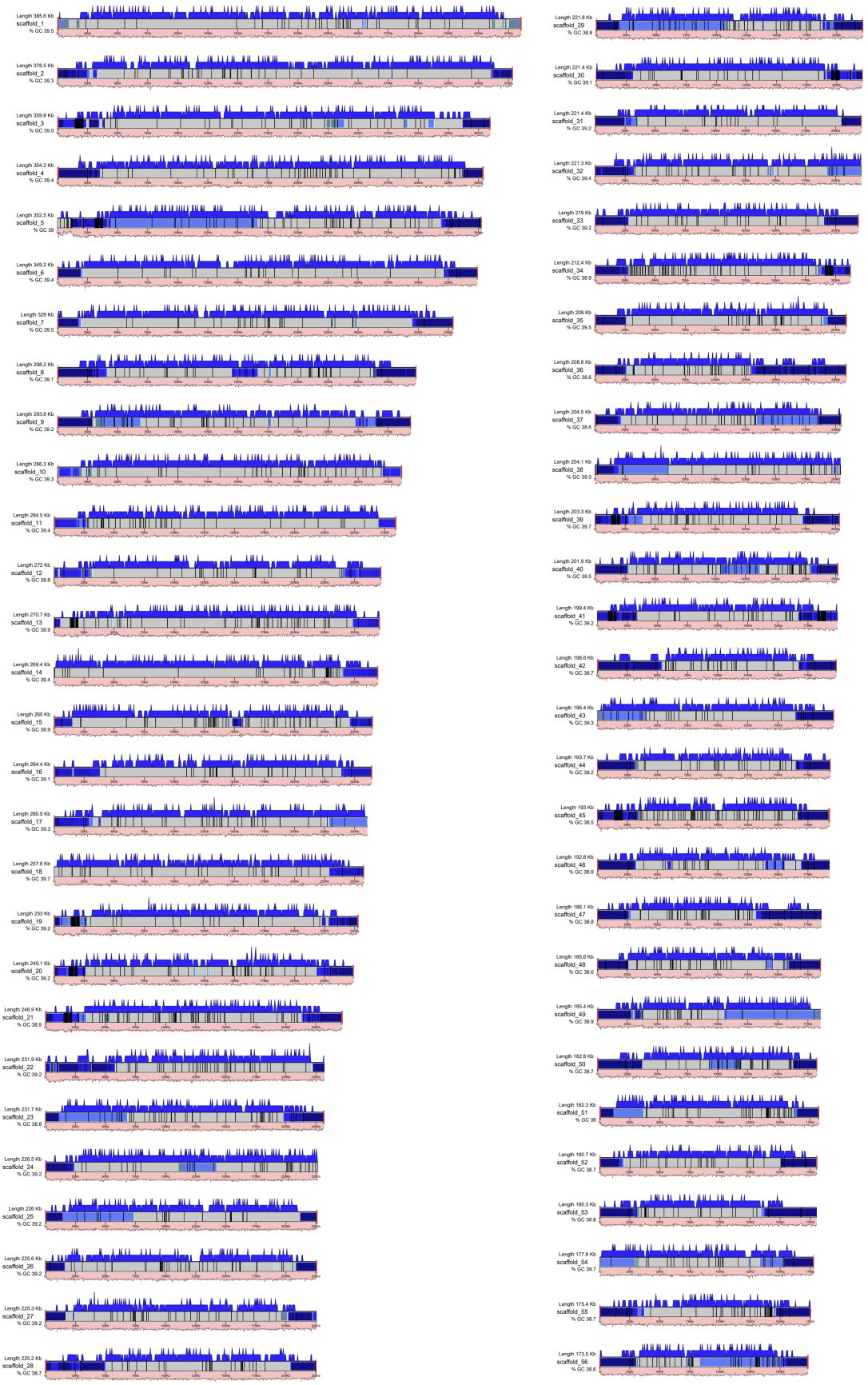



*Figure 5: Karyogram of G. sulphuraria SAG 107.79 nuclear assembly, generated in RStudio* [150] *with KaryoplotR* [151]*. Subtelomeric regions of increasing depth of coverage are shown in blue (dark blue indicating higher depth). Repeats were identified with RepeatMasker* [152]*. % GC content was calculated over 250 bp windows and gene density was calculated over 500 bp windows.*

# The complete *G. sulphuraria* ACUF 138 nuclear genome



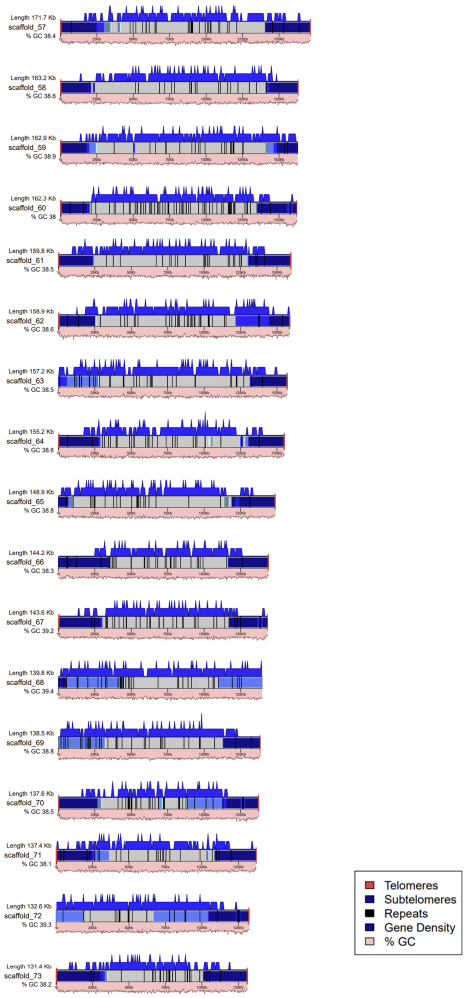



*Figure 6: Karyogram of G. sulphuraria ACUF 138 nuclear assembly, generated in RStudio* [150] *with KaryoplotR* [151]*. Subtelomeric regions of increasing depth of coverage are shown in blue (dark blue indicating higher depth). Repeats were identified with RepeatMasker* [152]*. % GC content was calculated over 250 bp windows and gene density was calculated over 500 bp windows.*

The complete *G. sulphuraria* ACUF 017, and draft *G. sulphuraria* ACUF 427, nuclear genomes

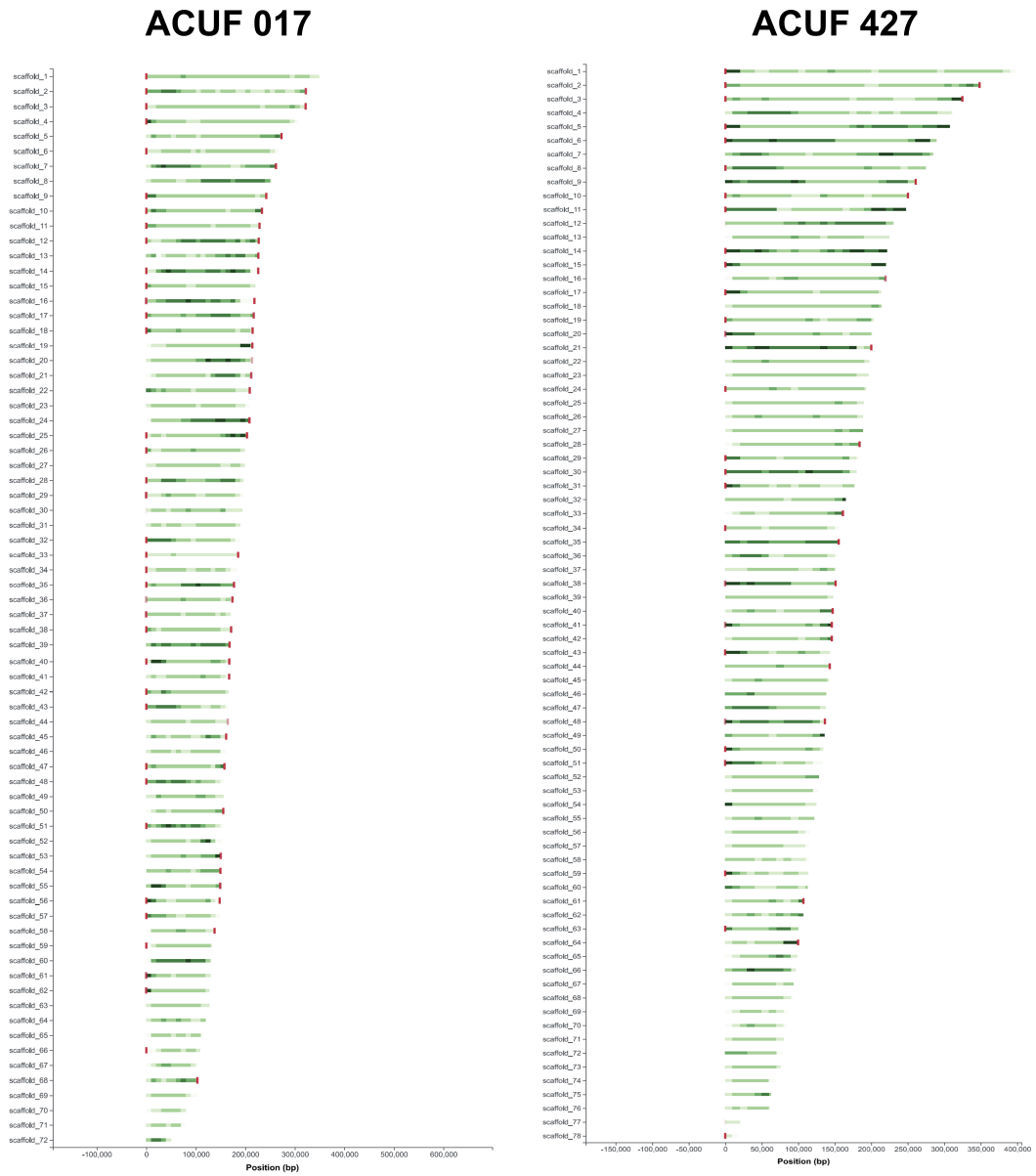



*Figure 7: G. sulphuraria ACUF 017 and ACUF 427 nuclear genome assemblies. Karyograms produced in Tapestry* [137]*. Red signifies telomeres, darker green regions signify increased coverage depth.*

### *G. sulphuraria* Genome Annotations Contain Putative Reverse Gyrases

A single homolog of *Sulfolobus acidocaldarius* reverse gyrase was identified in *G. sulphuraria* SAG 107.79, ACUF 138, and ACUF 017 (*Table 12*). Analysis of the domain architecture of these predicted amino acid sequences with NCBI BLASTp indicates that these predicted proteins contain ATP binding motifs essential for

reverse gyrase activity, and that they are homologous to reverse gyrases of bacterial thermophiles.

| ***G. sulphuraria* Protein ID** | **Expect Value to *S. acidocaldarius*** |
|---|---|
| Gs107_001412-T1 | $3.07 \times 10^{-43}$ |
| Gs138_002746-T1 | $1.27 \times 10^{-22}$ |
| Gs017_002555-T1 | $5.13 \times 10^{-41}$ |

*Table 12: BLASTp G. sulphuraria homologs, and BLAST expect values, to S. acidocaldarius reverse gyrase.*

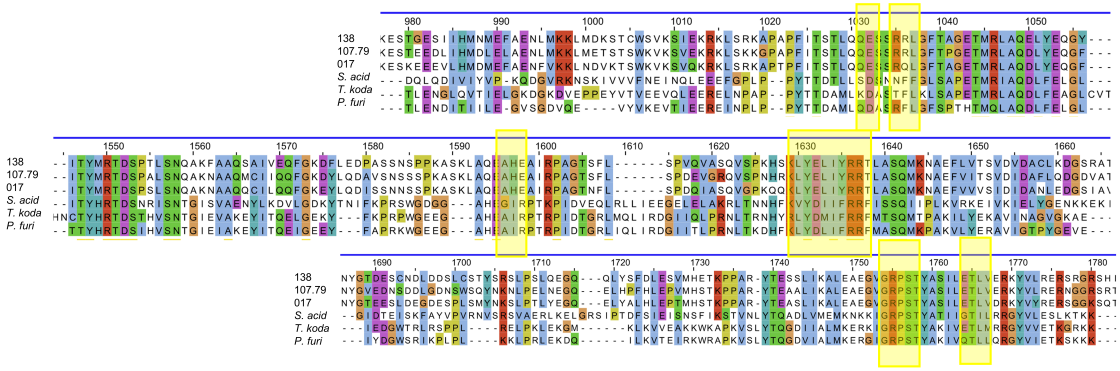



*Figure 8: Predicted amino acid sequence multiple sequence alignment sample, generated using MUSCLE* [147]*, of G. sulphuraria ACUF 138, SAG 107.79, and ACUF 017 putative reverse gyrases, and S. acidocaldarius, T. kodakaraensis, and P. furiosus reverse gyrases. ATP binding sites are highlighted in yellow.*

Multiple sequence alignments to archaeal reverse gyrases demonstrated that the predicted ATP binding sites are highly conserved between archaeal and *G. sulphuraria* reverse gyrases (*Figure 8*).

### The Genomes Exhibit Chromosome Copy Number Variation and Gene Duplication

These completed genome assemblies have allowed for further analysis into genome structure. Within each genome, there are several duplicated blocks of colinear genes on different chromosomes, indicating the presence of transposable elements (*Figure 9*). These colinear blocks are more prevalent on chromosome ends.


*Figure 9: Duplicated regions within the G. sulphuraria SAG 107.79 and ACUF 138 genomes, with links representing colinear blocks of protein coding genes.*

ONT sequencing read coverage is non-uniform across the genome, with examples from SAG 107.79 shown in Figure 10. Assuming that read coverage is proportional

to the amount of DNA in the sample (since ONT sequencing is a single molecule based technique), this indicates that there is a variation in chromosome copy number within *G. sulphuraria* populations, and that there may be polyploidy.

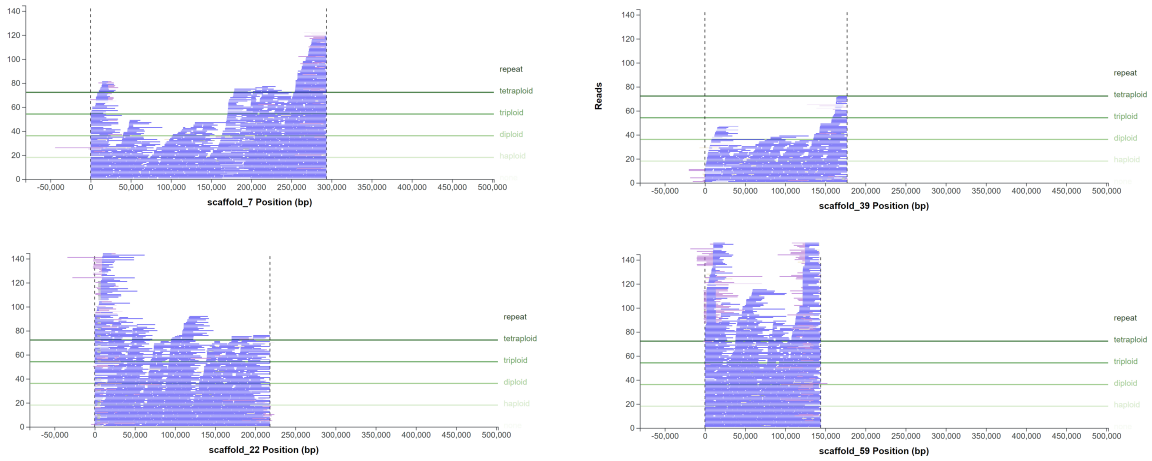



*Figure 10: Read coverage plots generated in Tapestry* [137] *for four SAG 107.79 scaffolds. y-axis indicates the relative number of reads.*

**The Completed *G. sulphuraria* Genomes Exhibit Many Structural Differences**

Macrosynteny analysis between *G. sulphuraria* SAG 107.79, ACUF 138, and ACUF 017, has large structural variations in addition to that of the length of the longest chromosome. The longest chromosomes of ACUF 138 and ACUF 017 are not colinear, despite being much more similar in size than that of SAG 107.79. Instead, the longest chromosome of ACUF 017, Gs017_1, is colinear to Gs107_6, and Gs107_3 – these three chromosomes are all similar in length (Gs017_1: 349111, Gs138_6:349231 Gs107_3:351951). The ACUF 138 longest chromosome is divided into multiple colinear blocks across several ACUF 017 chromosomes. The SAG 107.79 longest chromosome, being significantly longer than any chromosome from ACUF 138 or ACUF 017, is split into at least two blocks on each ACUF 138 and ACUF 017. A selection of these colinear and rearranged blocks are shown in *Figure 11*.

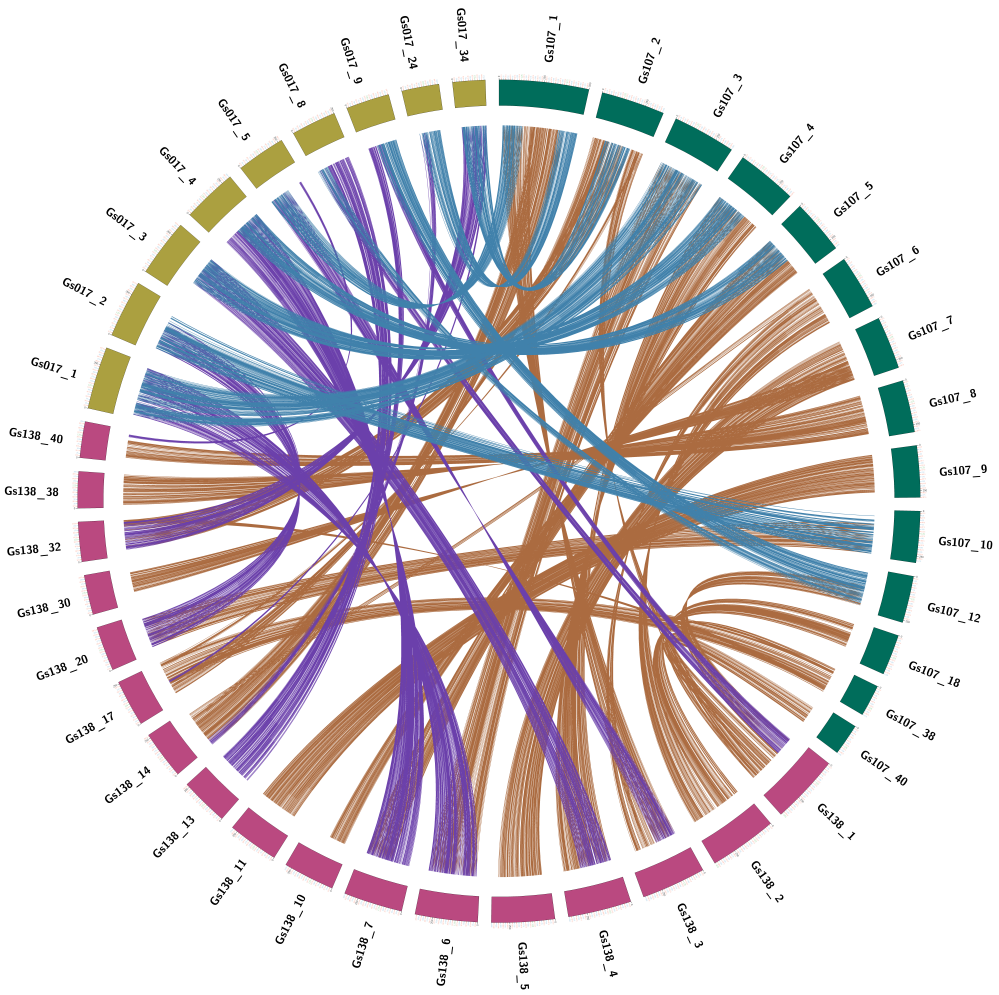



*Figure 11: Visualisation of the macrosynteny between a selection of chromosomes from G. sulphuraria SAG 107.79, ACUF 138, and ACUF 017. The largest chromosomes and their respective syntenic chromosomes were chosen for this visualisation. Collinear blocks of protein coding genes were identified with MCScanX* [148].

No syntenic blocks were detected between the *G. sulphuraria* SAG 107.79 and *C. merolae* 10D genomes. Whole genome alignment demonstrates that these genomes are extremely diverged, with just 3 sites, within 2 chromosomes, across the entire *C. merolae* 10D genome, mapping to *G. sulphuraria* SAG 107.79 (*Figure 12*). These sites map to the subtelomeres of complete telomere to telomere *G. sulphuraria* SAG 107.79 chromosomes, and are non-protein coding regions.

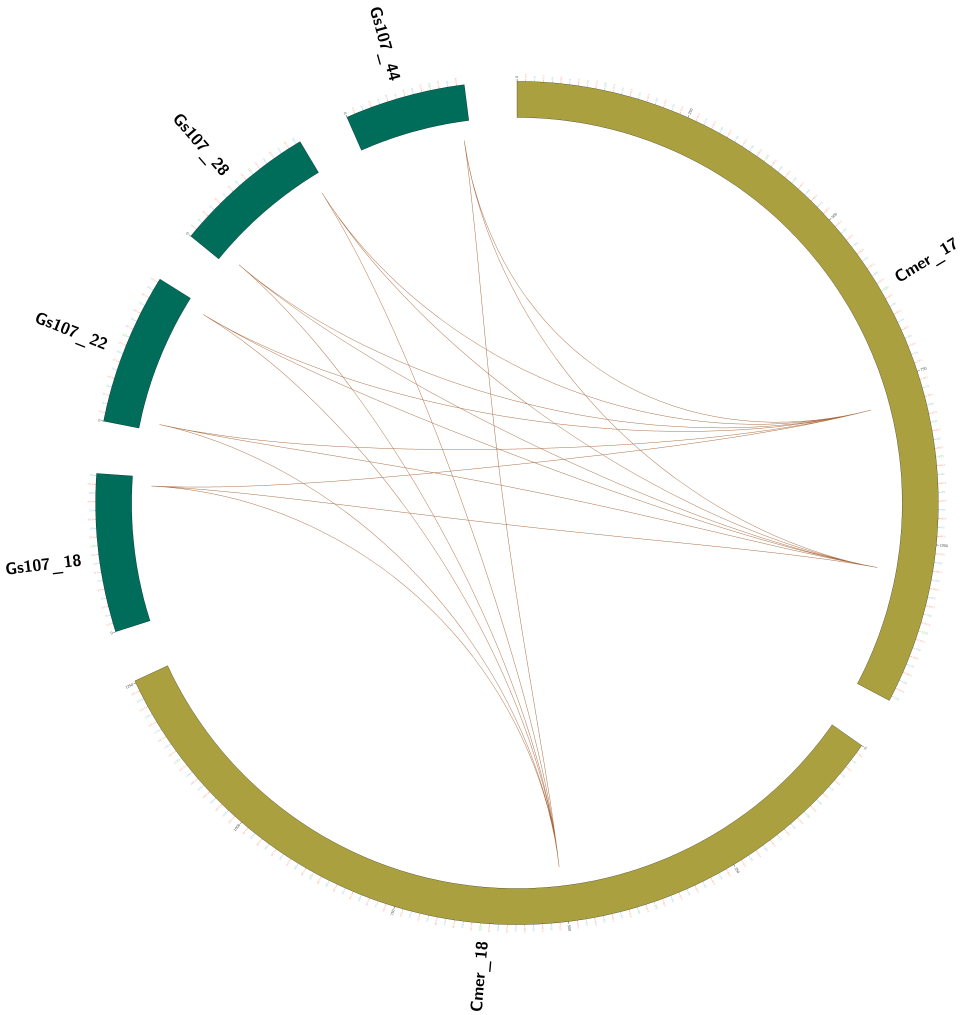



*Figure 12: Alignments of 3 C. merolae regions to G. sulphuraria SAG 107.79.*

Chromosomal structural diversity is seen across all *G. sulphuraria* sequencing data, with the shortest longest chromosome at 333346 bp, and the longest over twice as long at 707270 bp. For isolates ACUF 017 and 002, the longest chromosome length is reported to be identical. Given the diversity of genome structure between the rest of the isolates, this is highly unusual. These two isolates are from the same culture collection, and were isolated from the same fumarole in the Phlegraean Fields, Italy.

| Strain | Species | Length | Isolate Location | Author |
|---|---|---|---|---|
| SAG 107.79 | *G. sulphuraria* | 500417 | Yellowstone National Park, USA | This study |
| ACUF 017 | *G. sulphuraria* | 349111 | La Solfatara, IT | This study |
| ACUF 138 | *G. sulphuraria* | 385618 | El Salvador | This study |
| Soos | *G. phlegrea* | 451165 | Soos National Park, CZ | Rossoni et al. |
| 002 | *G. sulphuraria* | 349111 | La Solfatara, IT | Rossoni et al. |
| MS1 | *G. sulphuraria* | 333346 | Contaminant, USA | Rossoni et al. |
| MtSh | *G. sulphuraria* | 497389 | Mt. Shasta, USA | Rossoni et al. |
| RT22 | *G. sulphuraria* | 459287 | Rio Tinto, ES | Rossoni et al. |
| SAG21 | *G. sulphuraria* | 535216 | Yangmingshan, TA | Rossoni et al. |
| 5572 | *G. sulphuraria* | 488768 | Norris Basin, Yellowstone National Park, USA | Rossoni et al. |
| YNP5578 | *G. sulphuraria* | 707270 | Nymph Creek, Yellowstone National Park, USA | Rossoni et al. |
| GpartN1 | *G. partita** | 364794 | Kodakara Island, JP | Hirooka et al. |

*Table 13: Length of the longest chromosome of previously published G. sulphuraria assemblies by Rossoni et al.* [115]*, and Hirooka et al.* [124]*. *G. partita persists in the nomenclature however it is a G. sulphuraria isolate.*

The collinearity between ACUF 017 and ACUF 002 was therefore examined, and found to be higher than that between SAG 107.79 and ACUF 138, with 80.41% and 72.0% of genes found to be collinear, respectively, and the longest chromosomes of

ACUF 017 and ACUF 002 are collinear. A subsection of these alignments is shown in *Figure 13*.

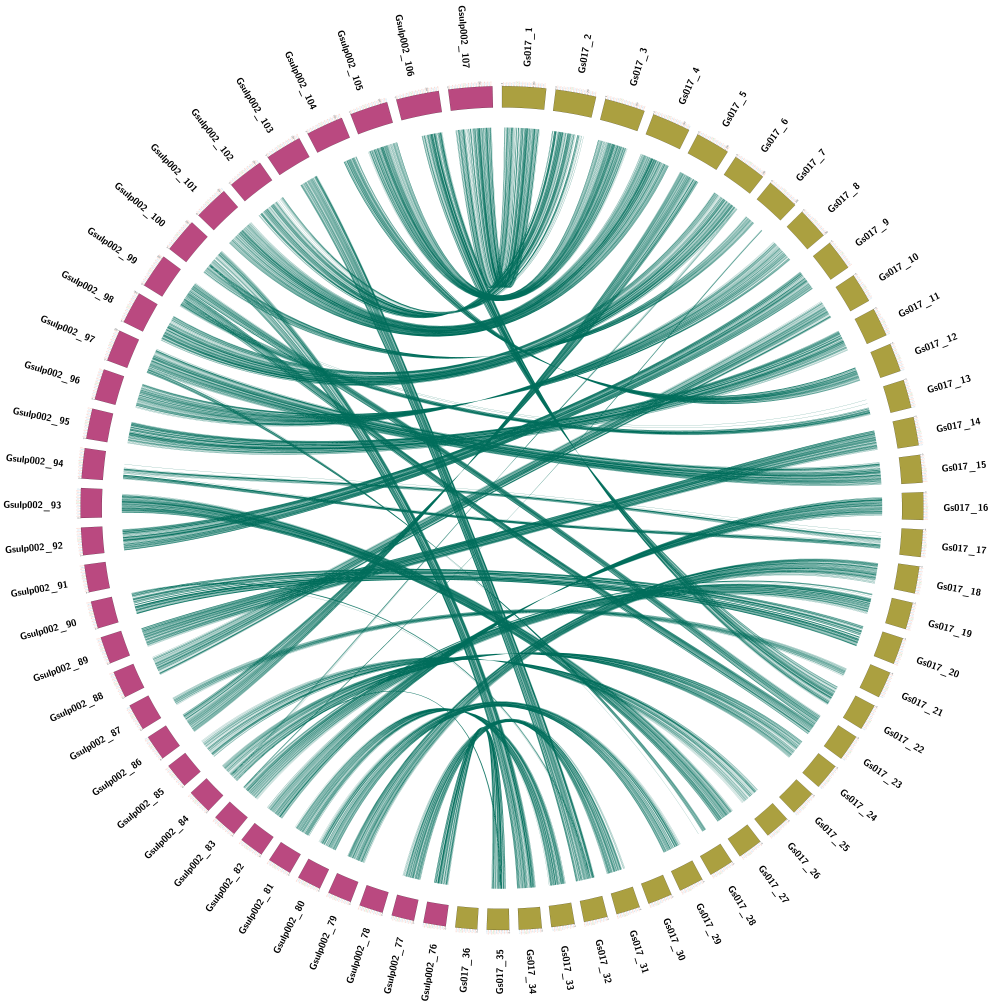



*Figure 13: Macrosynteny of the largest chromosomes of G. sulphuraria ACUF 017 (this study) and G. sulphuraria ACUF 002 (Rossoni et al.* [115]*). Collinear blocks of protein coding genes were identified with MCScanX* [148].

## Discussion

Here I have described three complete genomes of the polyextremophile *G. sulphuraria*, revealing a highly unusual genome structure – with an estimated 72 chromosomes for a genome size of 13.14 Mb (*G. sulphuraria* SAG 107.79). While it is not uncommon to find species with numbers of chromosomes in excess of 70, considering the relatively small genome size of *G. sulphuraria*, the number of nuclear chromosomes is especially high. Other small eukaryote genomes such as that of *Galdieria*'s relative *Cyanidioschyzon merolae* 10D (16.52 Mb) [9], and the green alga *Ostreococcus tauri* (12.56 Mb) [50] have much lower numbers of chromosomes – 20

for both species. This number of chromosomes is significantly higher than previously estimated by Moreira et al. [19], however this experiment underestimated the size of the *G. sulphuraria* genome by at least 3 Mb. This is likely because *G. sulphuraria* contains a lot of small chromosomes that would be indistinguishable on a pulse field gel. Here I show that *G. sulphuraria* has multiple chromosomes within the 100 Kb to 200 Kb size range, had all of these been attributed to a single or a few gel bands, this would cover the underestimation in genome size and number of chromosomes. Additionally, the *G. sulphuraria* plastid genome is also within this size range, at 168 Kb [44], which could have also been misinterpreted on a pulse field gel as a chromosome.

Weber conducted an additional pulse field gel, also underestimating the genome size and number of chromosomes (12 Mb, and at least 42 chromosomes) [101]. What is especially notable about the gel by Weber is that he reported a 1 Mb chromosome in *G. sulphuraria*, which has never been demonstrated in a *Galdieria* nuclear genome assembly from long-read sequencing data. The plastid genomes of plants can form large concatemers to aid genome stability [153][154][155], which could explain the 1 Mb gel band as a *G. sulphuraria* plastid concatemer. Otherwise, it is unclear what Weber found. He also reported that some gel bands were made up of several chromosomes, further supporting the above notion that the Moreira gel was an underestimation. Why *G. sulphuraria* has this many chromosomes is unclear, but since *G. sulphuraria* is extremely dominant in its environment [10], it does raise questions as to whether having many small chromosomes (as opposed to fewer, larger, chromosomes) could provide an evolutionary advantage for eukaryotes in extreme environments, by allowing for increased adaptation.

A recently published assembly from Hirooka et al. [124] reported 80 telomere to telomere chromosomes and a genome size of 17 Mb for a disputed *Galdieria* line *Galdieria partita* (as mentioned in the introductory chapter, this line of *Galdieria* nests within several *sulphuraria* isolates within the *Galdieria* phylogeny, and is therefore unlikely to be its own species separate from *sulphuraria*, although the use of *partita* persists in the nomenclature). This genome is larger than any previously reported *Galdieria* genome. The analysis published on the duplicated regions within this genome demonstrates that this assembly contains approximately 4 pairs of haplotigs that were retained in the final assembly, explaining the above average genome size.

This analysis agrees with my analysis of the *G. sulphuraria* SAG 107.79 and ACUF 138 genomes demonstrating that these genomes contain multiple duplicated regions, many of which are localised to the subtelomeres, indicating the presence of transposable elements. Transposition is known to represent a powerful mechanism of evolution in eukaryotes [156], yet what is interesting is that transposition results in genome expansion, which seems unusual considering the small genome size of *G. sulphuraria*. Therefore, in order to maintain a compact genome, *G. sulphuraria* must possess a well-balanced mechanism of transposon removal. Maintaining some genome duplication could be advantageous to a poly-extremophile, by allowing for sequence redundancy and the toleration of a high rate of mutation in the duplicated regions, enabling potentially advantageous mutations to take place, while not being majorly disadvantaged by deleterious mutations since a functioning copy of the gene exists elsewhere in the genome. This is a point of investigation I will discuss further in “

The *G. partita* assembly from Hirooka et al. [124] retained uncollapsed haplotypes, meaning that pairs of homologous chromosomes were retained in the final assembly. Before filtering, this was the case for the *G. sulphuraria* SAG 107.79, ACUF 017, and ACUF 138 assemblies, as was especially prevalent in the Canu2.1 assemblies. This is because *G. sulphuraria* is diploid and that the genome assemblers were not able to resolve all haplotypes into single chromosomes. Read mapping demonstrates that there are numerous heterozygous sites across the *G. sulphuraria* genome, and this is visible in Figure 4*C*.

The % GC content reported in these genomes is low for a thermophile (~39%). One would expect thermophilic microorganisms to exhibit a higher GC content, due to the increased stability of guanine/cytosine base pairing as a result of the additional hydrogen bond, however this is not always the case. The genome of the thermoacidophilic archaea *Sulfolobus acidocaldarius* is AT rich, with a GC content of 37% [157]. Archaeal genomes have reverse gyrases, which are ATP-dependent type 1A topoisomerases, that function to positively supercoil DNA, therefore increasing DNA stability and preventing DNA damage at high temperatures [158]. I found a single copy of topoisomerase type 1A in each complete genome. Crucially, the predicted amino acid sequences contained a topoisomerase 1Ac domain, which contains highly conserved ATP/ADP binding sites (*Figure 8*). Most topoisomerase type 1A enzymes are ATP independent, and they generally function to relax supercoiled DNA [159]. Only the positive supercoil inducing reverse gyrases are ATP-dependent type 1A topoisomerases, therefore these *G. sulphuraria* putative type 1A topoisomerases may function to induce positive supercoiling in the DNA, thus increasing temperature stability and mitigating the relatively low GC content.

Typically, type 1 topoisomerases are divided into two classes, prokaryotic (type 1A) or eukaryotic (type 1B). The predicted domains of the putative *G. sulphuraria* reverse gyrase are type 1A domains, and the homologous proteins are prokaryotic. The opposite is true for the type 2 and type 3 topoisomerases in the genome, which have homologs in the eukaryotic domain of life. These putative reverse gyrases may therefore be a product of horizontal gene transfer. Alternatively, they may have been

lost in the eukaryotic lineage as a consequence of movement into mesophilic habitats.

While many *G. sulphuraria* chromosomes were collinear between SAG 107.79, ACUF 017, and ACUF 138, there were a number of chromosomes that had undergone massive structural rearrangements between strains. This could be partially explained by the presence of transposable elements, specifically when these chromosomal differences pertain to the subtelomeres, however I wonder if there is an additional driving factor behind this, specifically for chromosomes that have “split in two” (or fused). Perhaps these structural changes are as a result of recombination, specifically non homologous recombination as a result of incorrect alignment of homologous chromosome pairs during meiosis. This is the topic of the next chapter.

Increased collinearity between genomes of isolates from the same *G. sulphuraria* lineage (as determined in Iovinella and Lock [120]) is expected, and this is shown in the comparison between isolates ACUF 002 and ACUF 017. This comparison has shown that two different sequencing methods (Rossoni et al. used PacBio sequencing) can arrive at the same conclusion in the form of the longest chromosomes of these isolates. Also, since these isolates are not close to 100% collinear, or identical, they are not a contamination, which was initially a concern as it is so unexpected to achieve sequencing reads, let alone assemblies, of the exact same length. This has enabled the demonstration that *G. sulphuraria* genome structural diversity is greater between isolates of different lineages, but also that genome structure is not identical for isolates that may be extremely recently diverged.

There is a possibility that some of the structural differences between the ACUF 002 and ACUF 017 assemblies could be artefacts of assembly, especially since the raw sequencing data for ACUF 017 was of less than ideal coverage, however a few examples indicate that some structural differences are biological. Gs017_17 is split into different chromosomes in ACUF 002, yet Gs017_17 has both telomeres and is assembled as a complete chromosome. Moreover, there are a number of chromosomes in the ACUF 002 assembly that appear to be absent in ACUF 017. Gsulp002_91 and Gsulp002_74 share no collinear regions with ACUF 017,

indicating that either these chromosomes have been recently lost in ACUF 017, or recently acquired in ACUF 002. These isolates were isolated from the same fumarole in the Phlegraean fields, it is significant that they do not show 100% collinearity between the two isolates. This indicates that *G. sulphuraria* could be frequently recombining and undergoing genomic structural changes.

The structural diversity between *G. sulphuraria* strains is not without impact. Due to the karyotypical differences between strains, the reference genome used for genomic analyses must be carefully considered. For example, it would be imprudent to use ACUF 017 as a reference genome for analysing a *G. sulphuraria* dataset collected mostly from Yellowstone National Park, as sequencing reads would map incorrectly to certain genomic regions, and because of differing gene localisation between American and Italian lines, any deductions from the resulting data may be less accurate than if a more structurally similar reference genome was used.

These complete genomes of *G. sulphuraria* have revealed secrets on some of the potential molecular mechanisms of extreme adaptation in this eukaryote, demonstrating that the mechanisms for extremophilicity in this eukaryote are multi-faceted and not only limited to previously established mechanisms such as horizontal gene transfer [18]. With the added benefit of these genomes, investigations can continue into furthering the understanding of the inner workings of extremophiles, as will be detailed in the latter sections of this thesis.

# Chapter 3: Understanding The Meiotic Capacity of *G. sulphuraria*: A Genomics Based Approach

## Introduction

Considering the large structural variations between and within *G. sulphuraria* genomes, I considered the possible mechanisms behind these changes. Historically, unicellular organisms have been perceived reproduce mostly, if not entirely, asexually, while multicellular organisms have been considered obligately sexual, with few clonally reproducing. However, increasing evidence suggests that this assumption is based on erroneous comparisons between multicellular and unicellular organisms, after all, a multicellular organisms is a collection of clonally propagating cells [160]. Evidence of sex has been found in all major eukaryotic groups in the form

of the “meiotic toolkit” [161], a collection of genes responsible for the process of meiotic recombination, described in *Table 14*. Authors have suggested ubiquity of this toolkit indicates that the last eukaryotic common ancestor was capable of sex, and that pure asexuality may be as a result of a loss of sex [161][162]. Some authors have suggested that “presumed asexuality may be due to a lack of study” [163], and that the presence of even a few meiotic toolkit genes is enough to indicate that meiosis is occurring [164]. There are many examples of species with recently described sexual cycles. Meiosis and gametes were only detected in the well-studied protozoan *Trypanosoma brucei* recently [165], [166], and the sexual cycle of the fungus *Aspergillus fumigatus* was described in 2009 [167], 150 years after the species was originally classified. Both of these organisms are human pathogens, causing sleeping sickness and the life threatening condition invasive aspergillosis respectively, yet their sexual cycles were not described for many years. For non-pathogenic, poorly understood species, such as *G. sulphuraria*, the question of sex could easily by bypassed.

| **Gene** | **Function** |
|---|---|
| HAP2 | Involved in gamete mating-type determination |
| SPO11 | DNA double strand breaks |
| REC8 | Meiosis-specific cohesin variant (paralog of RAD51) |
| HOP1 | Homologous chromosomes alignment |
| PCH2 | Pachytene checkpoint |

| | |
|---|---|
| DMC1 | Homologous recombination (paralog of RAD51A) |
| MND1 | Cofactor in homologous recombination |
| HOP2 | Homology search and recombination |
| MER3 | Crossover resolution-Pathway I |
| MSH4 | Crossover resolution-Pathway I (mutS family) |
| MSH5 | Crossover resolution Pathway I (mutS family) |
| ZIP4 | Synaptonemal complex (also SPO22) |

*Table 14: Meiotic Toolkit genes and their established function within meiosis and recombination. Taken from* [164]

There is no single widely accepted theory for the origin, or purpose, of sex. The debate around the origin of sex surrounds a couple of questions. 1) Is sex the ancestral state of eukaryotes? 2) What was the evolutionary benefit to sex that it was so widely maintained?

While the widespread presence of the meiotic toolkit across all eukaryotic lineages have led many authors to arrive at the conclusion the ancestral state of eukaryotes is sexual [163][164], this is not the only hypothesis that has been presented on the matter. Maciver concludes that the presence of the meiotic toolkit is not enough to conclude that an organism conducts sex, nor does the expression of meiotic genes preclude sex, since meiosis specific genes are used in other processes, such as DNA repair. Maciver instead suggested that polyploidy obviated the need for a sexual lifestyle, which had been suggested to have been required as to avoid mutational meltdown (the accumulation of deleterious mutations that eventually cause extinction), and that the last eukaryotic common ancestor lived a simple, asexual lifestyle [168].

The ubiquity of sex in eukaryotic lineages is conspicuous as there is a high cost to sex. Recombination can separate beneficial gene combinations, there are a host of potentially detrimental errors and mismatches associated with it, and there is a time cost. Moreover, the molecular machinery required for meiosis is extensive (over 50 proteins in some cases), and as a consequence there is a great risk of failure or sterility if there is a mutation in even a single gene. Mixis, which is the fusion of typically haploid gametes and/or nuclei, is also costly. The fusion of gametes (syngamy) and nuclei (karyogamy), require little energy but take a lot of time. Mitosis takes 15 minutes to 3-4 hours depending on species, cell size, and temperature, whereas meiosis can take from 10 to 100 hours depending on the amount of nuclear DNA. Other costs to mixis include mate searching, sexual selection, competition for mating partners, and physical contact damage [169]. For mixis to occur in *G. sulphuraria*, it is likely that the cell wall must be completely, or partially, broken down to allow for cell fusion. Given that *G. sulphuraria* cells are characterised by a thick cell wall, this process would be an additional energetic cost to the organism.

To answer the paradox of sex, one must ask what benefit does sex offer that it is so widely maintained [170]. A plethora of theories have attempted to answer this question. One theory suggests that meiotic sex provides benefits to organisms by creating recombination, and new gene combinations in offspring, thereby increasing genetic variation in populations. It is now widely recognised that theories based on the benefits of genetic variation are problematic for various reasons, as recombination also comes with the cost of losing beneficial gene associations, sex does not always result in recombination, and genetic variation is a group advantage, and does not bring immediate benefits to individuals within a population [171]. Another theory for the maintenance of sex is that meiosis is a phylogenetically conserved feature that cannot be eliminated because meiosis-mixis cycles are ancestrally fixed, however many eukaryotes are facultatively asexual, only carrying out a sexual cycle under certain conditions, so it is curious how these cycles were maintained in these organisms [172]. Meiosis has also been suggested to be maintained as it is a restoration tool for maintaining the integrity of nuclear DNA, however DNA repair can occur outside of meiosis .

The fitness-associated recombination model states that organisms invest more into recombination into sexual reproduction in environments in which the fitness of the

organism is low. This is supported in many facultatively sexual eukaryotes where it has been shown that sexual processes can be triggered by environmental stress [173]. This model is especially pertinent in extremophiles, which are more likely to have less than optimal fitness in the stressful, rapidly changing environments they occupy [19]. In *G. sulphuraria*, the RADiation sensitive 52 (RAD52) homolog, normally implicated in the double strand break repair process, is induced under salt stress [174]. This fitness associated repair model is difficult, however, to apply to obligately sexual organisms, and there is no complete model for the control mechanisms of fitness associated recombination [170].

Combinational theories have been proposed that sex is a comprehensive DNA restoration mechanism that combines the DNA repair function of meiosis and selectively eliminates defect mutants during the haploid phase [175]. These theories have expanded to suggest that meiotic recombination arose as a DNA repair mechanism in response to an increase of damaging reactive oxygen species (ROS) in the environment as a result of the dawn of photosynthesis, the great oxygenation event, and subsequently the oxidization of $Fe^{2+}$ to $Fe^{3+}$, which is known to generate ROS via the Fenton Reaction [38]. SPO11 is an archaeal topoisomerase VI homolog that has lost its ligase ability and introduces the essential double strand breaks at the beginning of DNA repair [160], has also been demonstrated to have ROS scavenging abilities [176]. It has been shown that the multicellular green alga *Volvox carteri* cannot initiate its sexual cycle in the presence of antioxidants, and an iron chelator inhibits sexual induction in this species [177], [178].

*G. sulphuraria* is known to tolerate high levels of ROS, and resides in environments not dissimilar to the types of environments that the earliest sexual organisms inhabited. Certain *G. sulphuraria* habitats have high levels of environmental iron, which cycles between $Fe^{2+}$ and $Fe^{3+}$. Understanding the potential sexual processes of *G. sulphuraria* could aid the development of the oxidative damage hypothesis for the evolution of sex.

A diverse range of mating systems are understood across the tree of life [166][179][180][181]. Two aspects of sexual reproduction will be discussed here, firstly the morphology of the gametes, and secondly the genetic determination of sex. The ancestral state of sex is considered to have been isogamy, which is a

reproductive system where all gametes are morphologically similar, particularly in size [182]. Although isogamous gametes are morphologically similar, they are not genetically identical, and are almost always associated with mating types. Mating types are the gametic genotypes that determine the molecular mechanisms that ensure compatibility between fusing gametes. Anisogamy involves clearly diverged male and female gametes, with male gametes being smaller and female gametes being larger. Fusion can only occur between the larger and smaller gametes. Anisogamy is almost universal in complex multicellular eukaryotes, however in unicellular organisms the asymmetry of gametes is much less prevalent, and isogamy is the norm [183]. In red algae, however, no isogamous systems have been identified, but this may be due to a lack of study and assumed asexuality.

The sex of an individual can be determined genetically in either the haploid (gamete) phase, or the diploid phase. Mammalian systems are diploid phase sex determination systems, meaning that sex is determined genetically in the diploid phase rather than in the haploid gametes. To clarify this, mammalian males or females are only created upon fertilisation and the formation of a diploid zygote from haploid gametes, which determines the sex depending on whether it has the XX (female) or XY (male) chromosome pair, and only mammalian females can produce large gametes, and males small gametes. In flowering plants, sex is also determined in the diploid phase [180]. The less well understood haploid phase sex determination system, is relatively common among eukaryotes and is reported to have arisen independently in different eukaryotic groups during evolution. In haploid phase sex determination systems, diploids are capable of producing both types of gametes (assuming there are only two mating types, since some isogamous systems have hundreds of mating types). The chromosomes responsible for sex determination in these systems are known as U and V chromosomes (U=female, V=male, although this designation is not always retained) [184]. UV systems have been detected within the Rhodophytes [185]. Sex determining chromosomes contain the genetic material required for the molecular mechanisms for the fusion of gametes of opposite mating types or morphologies, and the specific regions on these chromosomes containing these genes, sex determining regions (SDRs), are often non-recombining, to prevent the transfer of these genes onto the opposite haplotype, which would result in sterility [186]. Elucidating where *G. sulphuraria* sits in this range of different sexual

systems would be of great interest to the community, as it would represent one of the only extremophilic sexual systems.

A number of methods exist for examining signatures of sexual capacity. These can be loosely divided into organismal signs, and molecular/bioinformatic signs [187]. Organismal signs would include observing sexual processes or structures, which could potentially be observed in *G. sulphuraria* by triggering sexual processes with specific stresses. Another approach would include producing nonmonoclonal cultures and searching for crosses [163], however this would not be a trivial task for *G. sulphuraria* as there is no well-established method of genetic transformation for the species.

As previously mentioned, using comparative genomics to determine the presence of the meiotic toolkit can establish that an organism has retained the capacity for sex [161], but this does not demonstrate their expression and function in sexual reproduction, since these genes could function solely in DNA repair [187]. Transposable elements are maintained in sexually reproducing organisms [188], however these can also be spread through horizontal gene transfer, a process normally considered to be reserved for prokaryotes, but is theorised to be a driving factor in the adaptation of *G. sulphuraria* to its extreme environment [35][110]. Incongruence between mitochondrial, plastid, and nuclear phylogenies is also expected in a sexually reproducing organism [187]. Finally, linkage disequilibrium, the non-random association of alleles at different loci, is an important indicator of the genetic forces that structure a genome and can help allude to the presence or absence of sex [189].

During meiotic recombination, sections of DNA are transferred between sister chromosomes to increase the genetic diversity of the resulting offspring. Traces of this process can be detected through observing single nucleotide polymorphisms (SNPs) and whether these have been transferred to a sister chromosome [190]. Linkage disequilibrium refers to the non-random association of genetic loci. When meiotic recombination is taking place, two SNPs that are closer together are more likely to remain on the same chromosome, as there is less chance of the recombination site falling in between the two sites. This association would be non-random, therefore the loci would be in linkage disequilibrium. For two SNPs further

apart on a chromosome, their distribution between sister chromosomes is more likely to be random, as there is more chance the recombination site will fall between the two SNPs. Lewontin normalised linkage disequilibrium to a correlation coefficient $R^2$. For a pair of genetic markers, if $R^2 = 1$, when there is linkage disequilibrium. When $R^2 = 0$ the two markers are unlinked and distributed randomly, therefore there is no linkage disequilibrium. A decay of $R^2$ over increasing genetic distance would indicate that meiotic recombination has been taking place in a population [189].

It has been reported that the plastid, mitochondrial, and nuclear genomes of *G. sulphuraria* are incongruent. Not only were these pangenome phylogenies incongruent, but the majority of genes also presented distinct single gene phylogenies [120]. While different genes face different selective pressures and may therefore present different phylogenetic trees, an additional explanation for incongruence between the accessory and nuclear genomes is frequent recombination. Additionally, the extensive structural rearrangements seen in the *G. sulphuraria* genomes described in "*Chapter 2: Assembly, Annotation, and Comparison of Complete G. sulphuraria Nuclear Genomes*" could be explained by recombination. Considering these observations, and with the benefit of a complete reference genome as described in the previous chapter, and genome sequencing for a collection of *G. sulphuraria* isolates, I used the available genomic data to examine the molecular signatures for sex in *G. sulphuraria*.

## Methods

### Isolate Collection

*Galdieria* strains were obtained from the Algal Collection of University of Naples [125], the Culture Collection of Autotrophic Organisms [191], the Culture Collection of Algae at Göttingen University [126], the Tung-Hai Algal Lab Culture Collection [127]. All strains were isolated from a single colony obtained after streaking the culture across agar plates, respectively, and colonies were inoculated in Allen medium pH 1.5 [28] and cultivated at 37°C under continuous fluorescent illumination of 45 µmol photons·$m^{-2}$·$s^{-1}$. Detailed information on all *Galdieria* isolates is available in

**Appendix**, *Table 20*.

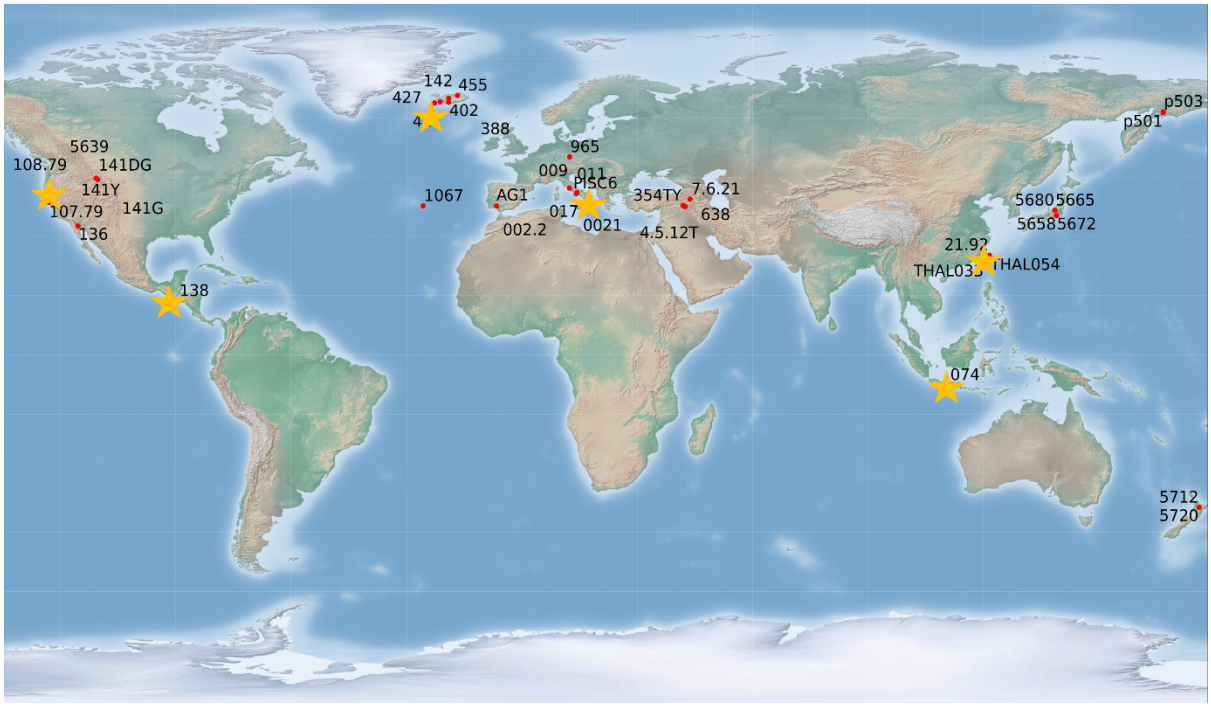



*Figure 14: G. sulphuraria Isolates Sampling Location*

### DNA Extraction and Sequencing

DNA was extracted from the *G. sulphuraria* isolates as described in Chapter 2, Methods, DNA Extraction.

### Meiotic Toolkit

Meiotic toolkit amino acid sequences (*Table 15*) were retrieved from GenBank [121], [122] and queried in a tBLASTn [144], [145] search against the *G. sulphuraria* SAG 107.79, and ACUF 138 CDS sequences.

| Gene | Species | GenBank Accession |
|---|---|---|
| HAP2 | *Arabidopsis thaliana* | AAY51998.1 |
| SPO11 | *Arabidopsis thaliana* | CAB81545.1 |
| RAD51 | *Cyandioschyzon merolae* | XP_005538367.1 |
| HOP1 | *Arabidopsis thaliana* | NP_172691.1 |
| PCH2 | *Saccharomyces cerevisiae* | NP_009745.2 |
| RAD51a | *Zea mays* | AAD32029.1 |
| DMC1 | *Oryza sativa* | BAB85214.1 |
| MND1 | *Arabidopsis thaliana* | ABB73190.1 |
| HOP2 | *Arabidopsis thaliana* | NC_003070.9 |
| MER3 | *Arabidopsis thaliana* | AAX14498.1 |
| MSH4 | *Arabidopsis thaliana* | AAT70180.1 |
| MSH5 | *Arabidopsis thaliana* | NP_188683.3 |
| ZIP4 | *Arabidopsis thaliana* | ABO71664.1 |

*Table 15: Meiotic toolkit homologs used to identify meiotic toolkit genes in G. sulphuraria. These homologs are known to function in meiosis in their species.*

The unique amino acid sequences returned from the BLAST search were extracted from the genomes and were inspected by NCBI's CD-Search [178][179] in order to confirm the presence of the domain architecture expected for each potential meiotic toolkit homolog.

**Variant Calling and Linkage Disequilibrium**

Illumina reads were aligned to the completed *G. sulphuraria* SAG 107.79 genome using the Burrow-Wheeler Aligner v.0.7.17 [136]. Aligned reads were processed using SAMtools v.1.10 [194] and the Picard Toolkit v 2.21.6 MarkDuplicates and AddOrReplaceReadGroups [195]. Variants were called using the Genome Analysis Toolkit v.4.1.0.0 (GATK) [196]. Repetitive, difficult to map regions were excluded from the analysis. These were identified as regions with high depth of coverage as determined by mosdepth v.0.2.8 [197] after all by all chromosome alignments using minimap2 v2.20[138] in mode -ax ava-ont. Hard filtering was applied using BCFtools v.1.10.2 [198] excluding sites with an RMS Mapping Quality (MQ), a Phred-Score

(FS), Quality by Depth (QD) of < 32. The linkage disequilibrium correlation co-efficient ($R^2$) values were calculated using PLINK v.2.00 [199], [200]. After calculating the mean $R^2$ value over 1Kb pairwise distance windows, Spearman's correlation coefficient was calculated using RStudio. VCFtools was used to analyse VCF files, and scikit-allel was employed for principle component analysis [201].

## Results

### The Meiotic Toolkit

Elements of the meiotic toolkit are present in the *G. sulphuraria* SAG 107.79 and ACUF 138 genomes (*Table 16*). Using tBLASTn searching, hits were identified for all meiotic toolkit homologs except Zip4 in all genomes. However, upon inspection of domain architecture, some of the corresponding amino acid sequences were not predicted to function in the meiotic toolkit. In all genomes, homologs of HAP2, SPOII, MND1, and MER3 were clearly identifiable, with the domain architecture indicating predictive function in meiosis, DNA recombination, or DNA repair. This is particularly pertinent for the HAP2 homolog, (Gs107_004995 in SAG 107.79), which was found to contain the hapless 2 domain that is required for gamete fusion. Two copies of SPOII were found in SAG 107.79. The single MND1 isoform identified contains a domain with a computationally predicted function in cell division and chromosome partitioning during meiosis. Two MER3 isoforms were identified, with one isoform containing an additional BRR2 domain, which has a predicted function in recombination and DNA repair. 2 RAD51 like proteins were identified. In *G. sulphuraria* SAG 107.79, the DMC1 homolog appears to be a mis-annotation with the gene being incorrectly predicted. The hypothetical protein contains 2 major domains, including one DMC1 domain. When the amino acid sequence is analysed with BLASTp, the DMC1 domain is 95%< identical to several complete *G. sulphuraria* DMC1 sequences on GenBank, while on the N-terminal end there is a mannosyl-transferase domain. There is a methionine at the start of the DMC1 domain which would indicate that this is a complete protein in its own right and the "fusion" with the mannosyl-transferase domain is an annotation artefact, and therefore *G. sulphuraria* SAG 107.79 contains a functional DMC1.

Several MutS proteins were identified, however it was unclear which of the MSH4 and MSH5 homologs were the true MSH4 and MSH5 proteins, as opposed to MSH1-7. The homologs listed are the longest tBLASTn hits with the lowest expect value for

each homolog, that additionally have predicted MutS domains. Although tBLASTn returned hits for HOP1, HOP2, and PCH2, investigation of the domain architecture of the amino acid sequences revealed that these proteins were more likely to enact different functions. No homologs for ZIP4 were found.

| **Protein** | **SAG 107.79 Sequence IDs** | **ACUF 138 Sequence IDs** |
|---|---|---|
| HAP2 | Gs107_004995-T1 | Gs138_001433-T1 |
| SPOII Isoform 1 | Gs107_005153-T1 | Gs138_001206-T1 |
| SPOII Isoform 2 | Gs107_000064-T1 | Gs138_001580-T1 |
| MND1 | Gs107_004584-T1 | Gs138_005505-T1 |
| MER3 Isoform 1 | Gs107_005125-T1 | Gs138_005415-T1 |
| MER3 Isoform 2 | Gs107_001677-T1 | Gs138_001609-T1 |
| MER3 Isoform 3 | Gs107_004374-T1 | Gs138_003896-T1 |
| MSH4 | Gs107_003745-T1 | Gs138_005378-T1 |
| MSH5 | Gs107_002816-T1 | Gs138_002840-T1 |
| RAD51-like Isoform 1 | Gs107_002841-T1 | Gs138_001844-T1 |
| RAD51-like Isoform 2 | Gs107_002101-T1 | N/A |
| DMC1 | Gs107_005615-T1 (599-949) | Gs138_002632-T1, Gs138_004325-T1 |
| RECA domain containing protein | Gs107_005838-T1 | Gs138_003562-T1 |

*Table 16: Meiotic toolkit homologs in G. sulphuraria SAG 107.79 and ACUF 138.*

## Variant Calling Statistics

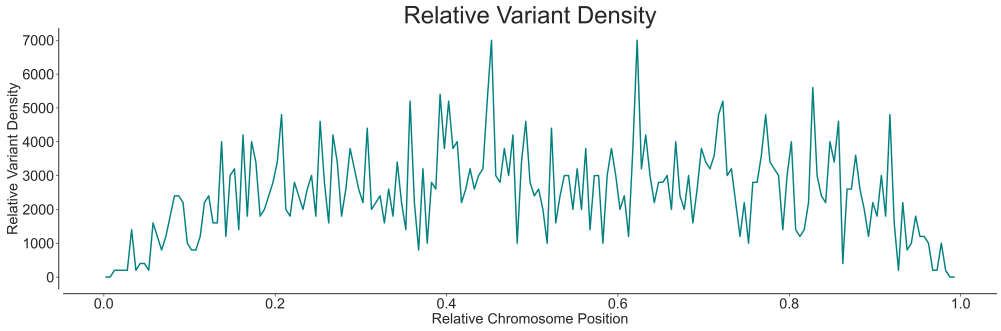



*Figure 15: Relative variant density over all chromosomes for variants called over all 49 samples.*

As the *G. sulphuraria* mitochondrial genome is reported to exhibit unusual features [41][44], and the nuclear genome has been shown to be AT rich, which is unexpected for a thermophile, the spectrum of variants was analysed (*Table 17*). Across 49 samples, 2011 SNPs and 775 INDELs were identified. Variant density was distributed evenly across the chromosomes, with the decay in variant density at chromosome ends accounted for as the telomeric and subtelomeric sites were removed from this analysis due to the repetitive nature of these sites rendering them difficult to map.

| **Substitution Type %** | | **Variant Nucleotide** | | | | **Mean Ref. Nucleotide** | **Standard Deviation** |
|---|---|---|---|---|---|---|---|
| | | **A** | **T** | **G** | **C** | | |
| **Reference Nucleotide** | **A** | | 7.84 | 15.10 | 5.93 | 9.62 | 3.95 |
| | **T** | 7.61 | | 6.32 | 14.16 | 9.36 | 3.43 |
| | **G** | 12.10 | 5.35 | | 4.88 | 7.44 | 3.30 |
| | **C** | 5.35 | 12.02 | 3.36 | | 6.91 | 3.70 |
| **Mean Variant Nucleotide** | | 8.35 | 8.40 | 8.26 | 8.32 | | |
| **Standard Deviation** | | 2.81 | 2.75 | 4.99 | 4.15 | | |

*Table 17: Substitution type matrix denoting the % of substitutions of each type (A -> T, A -> C etc.) in the 49 sample variant call. Assuming an equal proportion of substitutions, the expected value is 8.4%.*

Analysis of substitution types revealed that there were no significant biases towards specific substitutions, however A and T are more likely to be substituted. Although there were differences in the proportion of different substitutions, these were always paired with a similar level of substitution in the opposite direction. For example, 15.1% of variants were A > G, which is elevated above the expected 8.4%

(assuming an equal amount of all substitution types), but 12.1% of variants were G > A, also elevated over the expected 8.4%.

**Principle Component Analysis**

Lock and Iovinella resolved the nuclear phylogeny of *G. sulphuraria* using the same sequencing reads used for this experiment [120]. By examining the samples using Principle Component Analysis (PCA), the genetic diversity of these samples can be assessed, and the robustness of the variant calling pipeline, as the resulting clusters "agreeing" with the currently available phylogeny would demonstrate that these variants are most likely true variants and not artefacts.

The first four principle components shown explain 39.7% of the variance. The samples separate into 6 clusters, which are somewhat similar to the 6 *G. sulphuraria* lineages and the single *G. phlegrea* lineage described by Lock and Iovinella, with the dark green cluster in both PCA plots comprising of *G. phlegrea*. One major difference is that in the Lock and Iovinella phylogeny, the North American, Russian, New Zealand, Azores, and Japanese strains all form one clear lineage, whereas here they are split into two separate clusters. Moreover, while the Javanese isolate 074 remains separate in the PCA analysis (similar to the phylogeny), the El Salvadorian isolate ACUF 138 does not remain in a single cluster between PC1 and PC3. In PC4, the *G. phlegrea* samples do not form a separate cluster from the USA lineage.

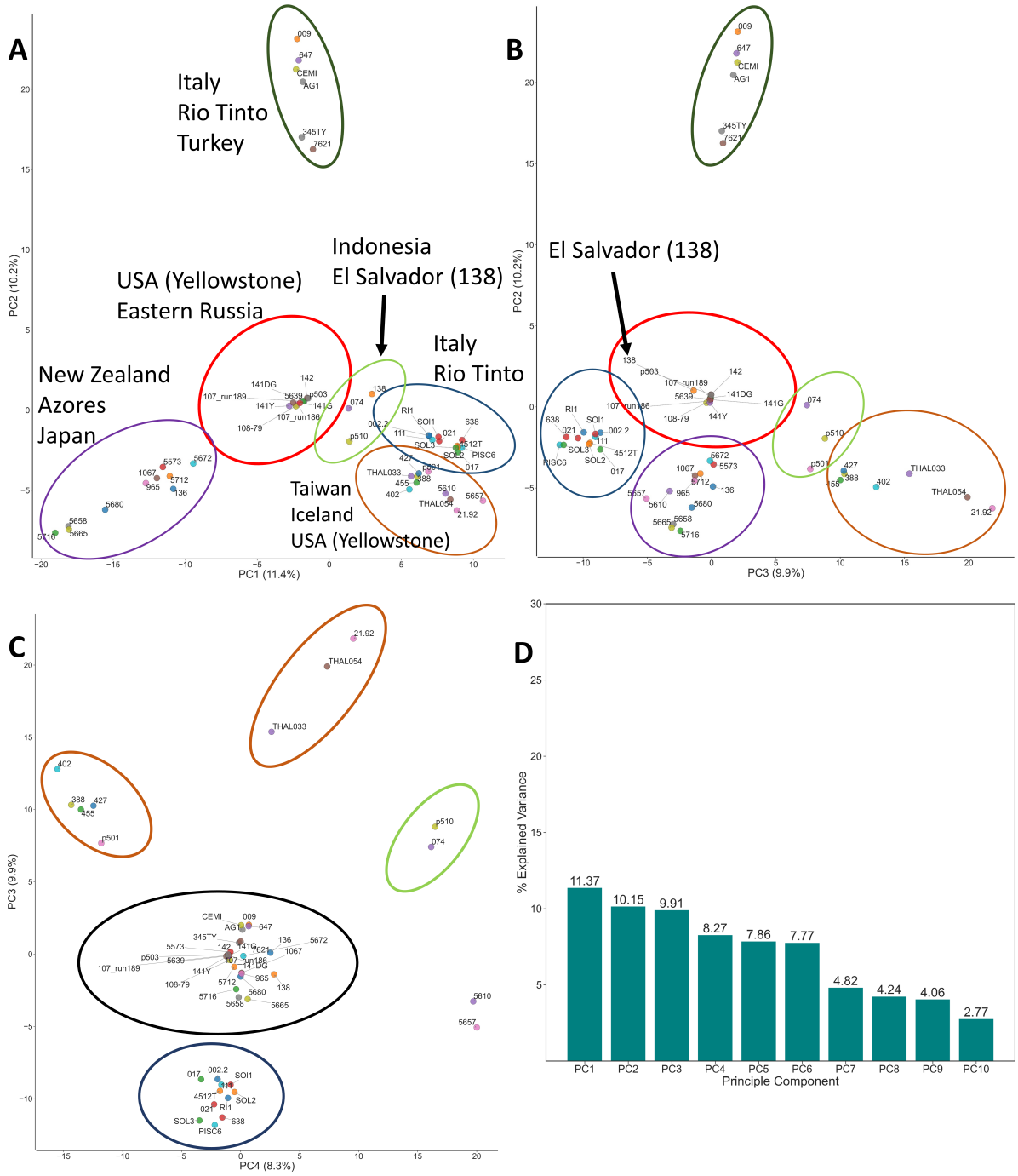



*Figure 16: A-C) Principle Component Analysis scatter plots showing Principle Components 1-4. D) Bar chart showing the % Explained Variance for each Principle Component*

**Linkage Disequilibrium**

The linkage disequilibrium coefficient, $R^2$, was calculated for pairs of SNPs no more that 100Kb apart across all 49 samples (*Figure 17*). For SNPs on the same chromosome, $R^2$ decayed significantly with pairwise distance, whereas for SNPs on different chromosomes, there was no decay.

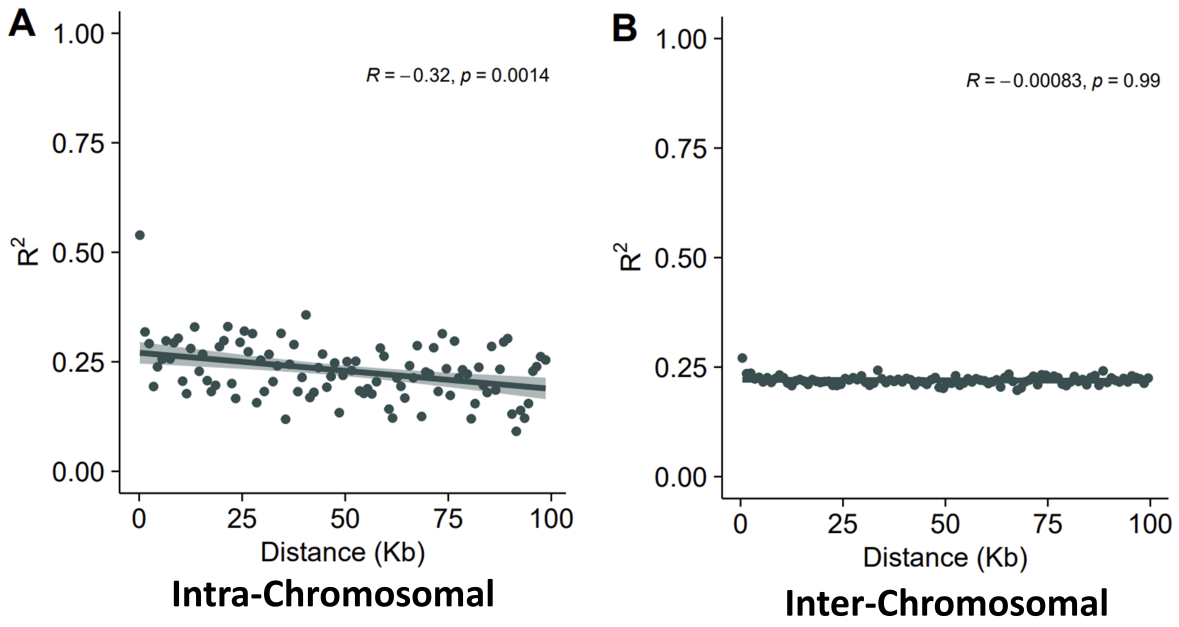



*Figure 17: Mean Linkage Disequilibrium Co-efficient ($R^2$) calculated over 1Kb windows, against mean pairwise distance for A) variants on the same chromosome, and B) variants on different chromosomes. R denotes the Spearman's correlation co-efficient.*

Based on the clusters shown in *Figure 16*, three subsets of samples from the same lineage were analysed separately. For detailed information on each isolate, see Appendix: *Table 21*

| **USA** | **Taiwan** | **Italy** |
|---|---|---|
| 142 | 388 | 002.2 |
| 108-79 | 5657 | 009 |
| 141DG | 402 | 111 |
| 141Y | 427 | 017 |
| 141G | 455 | RI1 |
| 5639 | 21.92 | SOL1 |
| p503 | 5610 | SOL2 |
| 107_run186 | THAL033 | SOL3 |
| 107_run189 | THAL054 | 4512T |
| | P501 | 021 |
| | | 638 |

*Table 18: Samples assigned to each lineage for analysis of individual clusters.*

Though not statistically significant, only the Taiwanese lineage yielded results indicative of recombination. For this lineage, the genome wide $R^2$ value was double the genome wide $R^2$ value when calculated across all 49 samples ($R^2 \sim 0.5$ in Taiwanese lineage, $R^2 < 0.25$ for all lines), and decayed with pairwise distance for the intra-chromosomal variants.

The USA lineage had an extremely low number of SNPs, even with gentle filtering options (excluding sites with an RMS Mapping Quality (MQ), a Phred-Score (FS), and Quality by Depth (QD) of < 25 for the USA lineage, and < 30 for the Taiwan and Italy lineages). This is due to the reference genome used (*G. sulphuraria* SAG 107.79), which is part of the USA lineage, and because the samples within this lineage are highly genetically similar, which is shown in the principle component analysis.

| **Lineage** | **Samples** | **SNPs** | **INDELs** |
|---|---|---|---|
| **USA** | 9 | 137 | 23 |
| **Taiwan** | 10 | 1330 | 537 |
| **Italy** | 11 | 616 | 196 |

*Table 19: Number of samples, SNPs, and INDELs.*

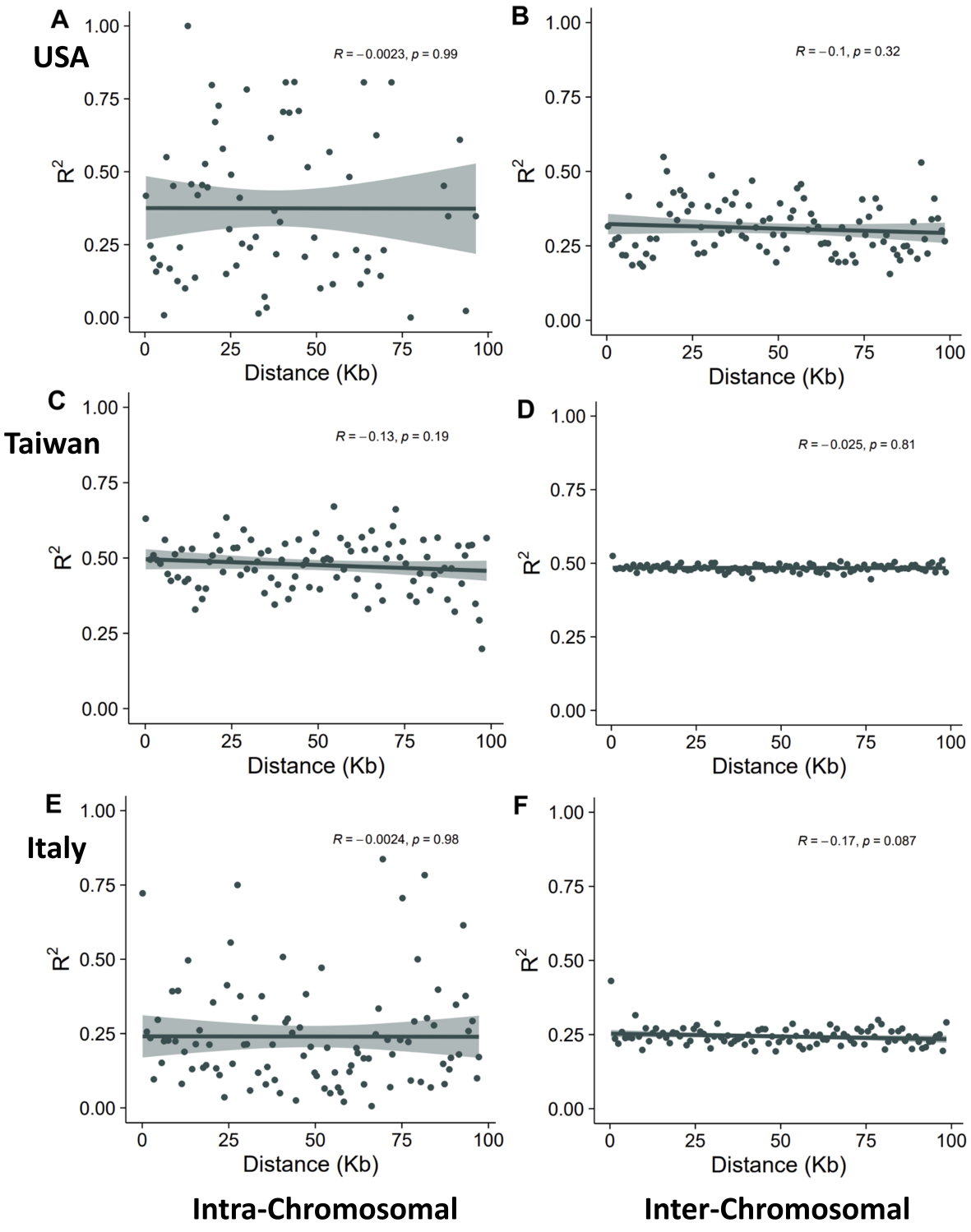



*Figure 18: Mean linkage disequilibrium co-efficient ($R^2$) for variants from the USA, Taiwan, and Italy lineages, calculated over 1 Kb windows. R denotes the Spearman's correlation co-efficient.*

## Discussion

### The *G. sulphuraria* SAG 107.79 and ACUF 138 Genomes Demonstrate the Capacity for Meiosis

Elements of the meiotic toolkit are found in both the SAG 107.79 and ACUF 138 *G. sulphuraria* genomes, however they both lack several genes implicated in meiosis. Namely, these are HOP1, HOP2, ZIP4, and PCH2. Some species show losses of a large component of the meiotic toolkit, while remaining sexually competent [164]. *Drosophila melanogaster* is fully sexual while lacking HOP1, DMC1, HOP2, MND1, MER3, MSH4 and MSH5. This case is explained by the replacement of the meiotic machinery in *Drosophila* by a distant homolog of DMC1 known as *spn-D*, a recombinase that performs the same functions of DMC1 in the fly [202]. The absence of HOP1 (and MER3, MSH4, and MSH5) are explained by the absence of crossover resolution pathway 1 in the fly, but crossovers can be resolved via alternative pathways [203]. Other instances of this phenomena exist, including in the yeast *Schizosaccharomyces pombe*, which lacks the synaptonemal complex and crossover resolution pathway 1, yet still performs meiosis [164]. Therefore, lacking these meiotic toolkit genes does not mean that a species is meiotically incompetent.

The machinery *G. sulphuraria* appears to lack is usually implicated in homologous chromosome searching and alignment (HOP1 and HOP2), and the assembly and checking of the synaptonemal complex (ZIP4 and PCH2). Should *G. sulphuraria* conduct meiosis, which given the presence of several meiosis toolkit genes it is likely to, how does it correctly conduct homologous chromosome alignment and crossover? The chromosomes of *G. sulphuraria* are small and numerous. Any machinery involved the pairing of homologous chromosomes and formation of the synaptonemal complex would likely need to be adapted to dealing with *G. sulphuraria's* small and numerous chromosomes, which could explain the lack of homologs for this type of machinery in the *G. sulphuraria* genome. I hypothesise that *G. sulphuraria* may instead conduct an alternative pathway for alignment and crossover, with enzymes better adapted to its unusual genome structure. The SPO11 domain of SPO11 Isoform 1 (Gs107_005153) is homologous to the archaeal topoisomerase VI domain that is implicated in double strand break repair, and has demonstrated ROS scavenging capabilities. Given the extreme habitat of *G. sulphuraria*, the *G. sulphuraria* SPO11 homolog is an interesting candidate for the

further testing of the oxidative damage hypothesis for meiosis, since, in the hypothetical absence of other nucleic ROS scavenging mechanisms, the G, sulphuraria SPO11 homolog must have higher activity to deal with the high levels of ROS *G. sulphuraria* is exposed to in its environment. Additionally, harsh environments can directly fragment DNA, and the SPO11 isoforms in the *G. sulphuraria* genomes may be involved in the repair for these fragmentation events. The evolutionary distance of *G. sulphuraria* from its neighbours, exemplified further by the low sequence similarity between *G. sulphuraria* meiotic toolkit genes and the toolkit genes of its nearest well understood sexual species. Homology based characterisation of *G. sulphuraria* genes continues to present challenges.

In "*Chapter 2: Assembly, Annotation, and Comparison of Complete G. sulphuraria Nuclear Genomes*", I have shown that there is widespread non-homologous recombination in the *G. sulphuraria* genome, with chromosomes having undergone many structural changes throughout the generations. Perhaps the apparent lack of homologs of genes that are involved in the correct assembly of the synaptonemal complex is intentional, enabling the production of an increased variety of new chromosomes at meiosis, increasing variation within the population. While hypotheses citing increased variation as a reason for the maintenance of meiosis are considered problematic, as much of this variation is potentially detrimental, particularly when separating physically associated genes, these assumptions are based on non-extremophilic organisms. The Cyanidiophyceae represent the only family of extremophilic eukaryotes and face distinct evolutionary pressures. *G. sulphuraria* may not have reached a steady state in its evolution as a consequence of the rapidly changing extreme environment in which it resides, and therefore the increased variation hypothesis may still apply in this case.

Although *G. sulphuraria* contains the meiotic toolkit genes, this is not enough to conclude that it definitely conducts meiosis, and additional evidence, such as clear linkage disequilibrium is required to further signal that meiosis is taking place.

**Linkage Disequilibrium**

While separate analysis of *G. sulphuraria* lineages was inconclusive regarding the presence of linkage disequilibrium, when calculated over all isolates there is statistically significant linkage decay, indicating that homologous recombination has

been occurring. This decay, however, is slight and does not exhibit a classical linkage decay curve that one would usually expect in a frequently recombining organism. There are a number of possible reasons for this. Firstly, the *G. sulphuraria* chromosomes are short and show large variations in length (~ 100 – 500 Kb), and there could be steric hinderances preventing recombination occurring between pairs of shorter chromosomes, however calculating this over this dataset is not possible since there is a large amount of genome structural diversity between strains, and not all of the chromosomes are colinear between different strains. The structural diversity between strains could also explain the shallow gradient of decay, as sequencing reads from isolates that do not have similar whole genome structures to *G. sulphuraria* SAG 107.79 will not map to regions of the SAG 107.79 genome that are reflective of their loci in their own genome, further reflecting the need for many complete *G. sulphuraria* genomes. Finally, the shallow gradient may reflect that homologous recombination has occurred infrequently, and only under a specific set of conditions. The *G. sulphuraria* isolates taken for sequencing were obtained from stock centres, where strains may have undergone years of domestication outside of their natural habitat. These isolates have not been exposed to the rapidly changing extreme environment that *G. sulphuraria* cells are currently exposed to in nature today.

For the lineage specific LD calculations, specifically in the case of the USA lineage, the lack of significant linkage decay can be explained by the low number of variants present in the dataset, since the samples were extremely genetically similar if not identical to the reference genome. This would also explain the variation in $R^2$ for the inter-chromosomal variants for this lineage, which is not seen in any other dataset.

The Italian dataset is particularly unusual as there was a sufficient number of variants to calculate $R^2$, yet there was no negative correlation between $R^2$ and pairwise distance for the inter-chromosomal variants. There is slightly less variance between these samples than for the Taiwanese samples, as demonstrated in the principle component analysis, and this is reflected in the number of variants detected in this lineage. Moreover, the Taiwanese lineage shows a higher mean $R^2$ level across all variant pairs, nearly double that of the Italian lineage, meaning that a pair of variants on the same chromosome are much more likely to be linked, indicating that the rate of recombination in these samples is lower.

**The Molecular signatures for sexual capacity are in the *G. sulphuraria* genome, so what is the physical mechanism for meiosis in *G. sulphuraria*?**

*G. sulphuraria* demonstrates the genomic signs for sex, but how it conducts these sexual cycles on a molecular level is not fully understood, and it is not possible to conclude based on these genomic signs alone that meiosis is taking place. The presence of a HAP2 homolog in both *G. sulphuraria* genomes indicates that *G. sulphuraria* is capable of forming mating types [164], as this gene is directly implicated in gamete formation in other species. Recently, consistent with the genomic results presented here, *G. sulphuraria* gametes were observed in laboratory conditions. These gametes are without a cell wall and are motile. They additionally can proliferate asexually and undergo self-diploidisation, and mate with different haploid cells to form heterozygous diploids. Key genes involved in the haploid life cycle, BELL, KNOX, and MADS, were also identified [124], however, these genes are not at nearby loci in the *G. sulphuraria* SAG 107.79 genome, and sex determining regions in the *G. sulphuraria* genome have yet to be identified.

*Ostreococcus* species also have compact genomes (13 Mb) [204], the meiotic toolkit [205], and candidate mating type loci have been identified, based on reduced recombination rate and GC content in specific genomic regions [206]. However, the size of these regions is much larger than what would be possible for the *G. sulphuraria* genome - 650 Kb and 450 Kb, the former larger than the largest *G. sulphuraria* chromosome, and the latter less than 50 Kb shorter . This poses additional questions as to what a sex determining region would look like in a genome with many small chromosomes.

The motility of these cells is notable, since flagellae were lost in the ancestral genome reduction in the Rhodophytes [102]. It was reported that the motility of these gametes is actin dependent, reflecting an ancestral role for actin. It was also demonstrated that these gametes are isogamous [124], leading to additional questions about what variety of mating types *G. sulphuraria* may exhibit (since some species that exhibit isogamy can have hundreds of mating types).

Although haploid cells have been generated in a laboratory environment through lowering the pH and transferring cells to a $CO_2$ incubator, they have never been isolated from the environment, and it is still unknown how the haploid life cycle is

induced in the environment. These results are difficult to replicate, and the difficulty generating these haploid gametes may demonstrate that meiosis occurs infrequently in the environment, which would coincide with my results showing a low degree of linkage decay. It is interesting that transferring cells to a $CO_2$ incubator induced the formation of haploid cells, and this contradicts the oxidative damage hypothesis for the maintenance of meiosis. This provides a useful point for further investigation into the sexual cycles of *G. sulphuraria*.

*G. sulphuraria* conducts meiosis, and these findings further challenge the concept of assumed asexuality among micro-organisms and point to the diversity of sexual systems across the tree of life. Meiosis may represent a mechanism of adaptive evolution or genome regeneration for *G. sulphuraria*, further explaining how *G. sulphuraria* can maintain its adaptive, flexible lifestyle in extreme environments.

# Chapter 4: The Spontaneous Mutation Rate of *G. sulphuraria* SAG 107.79

## Introduction

Mutation affects nearly all aspects of biology, and is a major driver of adaptive evolution [207]. While there has been an increasing amount of studies on the genomic mechanisms of extremophily, unveiling a number of factors in extreme adaptation, including genome plasticity, codon bias, nucleotide skew, and horizontal gene transfers [19], research on the impact of spontaneous mutation on microbial communities populating extreme environments has been limited.

Prior to the widespread availability of high throughput whole genome sequencing, most strategies for determining the mutation rate and spectrum were indirect [208], [209]. These strategies included interspecies comparison of putatively neutral sites in specific genes, and analyses using reporter construct genes [210]. However, because selection can affect synonymous sites (mutations that occur within a coding sequence that do not affect the amino acid sequence) [211], mutation rates can vary significantly across different regions in the genome [212], these methods are likely to have significant biases.

The advent of next generation sequencing has allowed for accurate and unbiased estimation of the mutation rate, through long term mutation accumulation experiments. These are conducted by propagating replicate lines taken from a single colony through regular population bottlenecks, allowing cell lines to accumulate mutations in an unbiased fashion. Whole genome sequencing is then used to directly identify these mutations and estimate the genome wide rate and spectrum of spontaneous mutations [207]. These experiments have led to the unbiased estimation of mutation rate in a wide variety of species, revealing that although the base-substitution mutation rate across all organisms is low ( $< 10^{-7}$ mutations per nucleotide site per generation), there is a large variation in the mutation rate with rates in some species being over 1000-fold below this level [207].

For *G. sulphuraria*, I have shown in this thesis that it is a recombining organism with many small chromosomes, and hypothesised that these features play a role in extreme adaptation. Additionally, prior work has shown that horizontal gene transfer is a driver of adaptive evolution in *G. sulphuraria* [110][115]. I have shown that the *G.*

*sulphuraria* genome contains syntenic blocks of duplicated genes, amounting to 516 duplicated genes, 1032 genes in total for *G. sulphuraria* SAG 107.79. Rhodophyte genomes have undergone an ancestral genome reduction [102], and the genome size of *G. sulphuraria* is typical for a Rhodophyte. The maintenance of these duplications in a compact Rhodophyte genome is therefore curious. I therefore wanted to inquire as to whether maintaining these duplicated regions could provide any selective advantage to *G. sulphuraria*, by allowing for a higher spontaneous mutation rate in protein coding regions with redundant copies, providing a rapid, pre-emptive, mechanism of adaptive evolution. Moreover, the estimation of the genome wide mutation rate of *G. sulphuraria* represents the first mutation rate estimation of an extremophilic eukaryote.

## Methods

### Data Collection

A single (parent) colony of *G. sulphuraria* SAG 107.79 was propagated onto 28 plates (Allen's, pH 2, supplemented with 30 g $L^{-1}$ sucrose, solidified with Phyto-Agar) and cultivated at 37°C under continuous fluorescent illumination of 45 µmol photons·$m^{-2}$·$s^{-1}$. A single colony was propagated from each plate every 20 days. The parent colony was grown in liquid media (Allen's/Sucrose pH 2) for 30 days. After 10 generations, for each sample, a single colony was taken into liquid media (Allen's/Sucrose pH 2) and grown at 37 ˚C and constant light for 30 days. After 30 days, DNA was extracted from each sample by the methods described in "*Chapter 2: Assembly, Annotation, and Comparison of Complete G. sulphuraria Nuclear Genomes*", under "DNA Extraction". Each sample, including the parent sample, was sequenced by Novogene using the Illumina NovaSeq PE150 sequencing strategy. The total number of days from the initial propagation of the single colony to the extraction of DNA was 223 days. In order to estimate the number of cell divisions that had taken place (enabling the calculation of the mutation rate), the number of cells in single colonies after 20 days of growth was estimated using a hemocytometer and optical microscope.

### Variant Calling

Raw sequencing reads were assessed for quality using FastQC v0.11.7 [213], [214] and it was determined that trimming was not necessary since there was no significant reduction in quality across sequencing reads. Paired-end reads were

aligned to the *G. sulphuraria* SAG 107.79 2022 reference genome using the Burrow-Wheeler Aligner v.0.7.17 [136]. Aligned reads were filtered for properly paired reads with Mapping Quality > 40 using SAMtools v.1.10 [194]. The Picard Toolkit v 2.21.6 MarkDuplicates and AddOrReplaceReadGroups [195] were employed to further process read alignments for variant calling. Single sample variants were called with FreeBayes v1.3.6 [215], and filtered removing variants with a Phred-scaled quality score < 30. Variants were phased with WhatsHap v.1.4 [216] (retrieved from Bioconda [217]) using the *G. sulphuraria* SAG 107.79 ONT reads aligned to the *G. sulphuraria* SAG 107.79 reference genome with minimap2 v2.20 [138] in mode -a ava-ont. Repetitive, difficult to map regions were excluded from the analysis. These were identified as regions with high depth of coverage as determined by mosdepth v.0.2.8 [197] after all by all chromosome alignments using minimap2 v2.20 [138] in mode -ax ava-ont. Additionally, sites that were present in the parent sample or were present in 2 or more samples were remove using VCFtools v0.1.16 [218], as the likelihood of a novel mutation taking place at the same site in several samples is extremely low.

For the analysis of duplicated protein coding regions, and non-duplicated protein coding regions, only coding regions that did not overlap with unmappable genomics regions (as defined above) were analysed. Non-duplicated protein coding regions were randomly selected, to create an identical sample size.

**Measuring Transcript Abundance**

RNA sequencing reads for *G. sulphuraria* SAG 107.79 mapped to the *G. sulphuraria* SAG 107.79 genome assembly using STAR v. 2.7.3 [140], as detailed in "*Chapter 2: Assembly, Annotation, and Comparison of Complete G. sulphuraria Nuclear Genomes*", Genome Annotation". Transcripts were quantified using HTSeq v. 0.11.0 [219] with parameter --nonunique all as duplicated genes were being targeted and reads that aligned to both duplicated genes needed to be counted. The same set of duplicated and non-duplicated protein coding genes in mappable genomic regions were used for this analysis as were used for calculated the mutation rate.

All statistical analyses were carried out in Python3. Unless otherwise specified, ± indicates one standard error of the mean.

## Results

### *G. sulphuraria* estimated growth rate

There was a large variation in the size of *G. sulphuraria* colonies throughout the experiment. At each re-streaking, colonies of a similar size were selected. The average number of cells per colony throughout the experiment was estimated as 2694400 ± 172629.8, and colonies were assumed to have formed from a single cell. Exponential growth was assumed, therefore $s = a(1 + r)^t$ where ***s*** is the number of cells in a colony, ***a*** is the initial number of cells, *r* is the growth rate – that is the amount of time it takes for a cell to duplicate, and *t* is the number of time intervals. Based on these estimations and assumptions, the rate of growth was estimated as 1.0966 ± 0.83 cell divisions per day. There were 193 days between the streaking of the first colony, to taking the colonies into liquid media for DNA extraction. Since parent and daughter colonies were grown in liquid media for the same amount of time, it can be assumed that mutations occurred in an equally unbiased fashion during this time period in both the parent and daughter samples and these time periods can be excluded from the analysis. The estimated number of generations for this experiment is therefore 211.65 ± 159.9.

### The Genome Wide Mutation Rate of *G. sulphuraria*

Of the 28 daughter lines that were sequenced, 27 had sufficient coverage to accurately determine the accumulation of novel mutations. The parent line was sequenced at 142x depth of coverage. The mean depth of coverage across all lines was 105x.

In total, 97.03 % of the *G. sulphuraria* SAG 107.79 genome was calculated as mappable, comprising 13856770 bp. There are 1032 genes that are duplicated and colinear in the *G. sulphuraria* SAG 107.79 genome, however, over half of these genes are localised to difficult to map regions, leaving 503 coding regions for the analysis, making up 961116 bp. This is to be expected since the fact that these regions are present in several places in the genome makes them difficult to map to. The 503 sampled non duplicated coding regions made up 807915 bp.

Across all mappable regions in 27 *G. sulphuraria* SAG 107.79 lines, I identified 1741 single-nucleotide variants (SNVs), yielding an overall nucleotide substitution rate of $2.21 \times 10^{-8} (SE = 0.22 \times 10^{-8})$ per site per generation. Additionally, 183 multi-

nucleotide variations (MNVs), and 604 indels, yielding an MNV rate of $2.31 \times 10^{-9}$ $(SE = 0.17 \times 10^{-9})$ per site per generation, and an indel rate of $7.63 \times 10^{-9}$ $(SE = 0.68 \times 10^{-9})$ per site per generation. Together, this yields an estimation of the overall mutation rate for *G. sulphuraria* SAG 107.79 of $3.19 \times 10^{-8}$ $(SE = 0.28 \times 10^{-8})$ per site per generation.

| | **SNV** | **MNV** | **Indel** | **Total** |
|---|---|---|---|---|
| ***G. sulph*. Whole Genome** | 2.21 x $10^{-8}$±0.22 x $10^{-8}$ | 2.31 x $10^{-9}$±0.17 x $10^{-9}$ | 7.63 x $10^{-9}$±0.68 x $10^{-9}$ | 3.19 x $10^{-8}$±0.28 x $10^{-8}$ |
| ***G. sulph.* Duplicated Regions** | 7.43 x $10^{-8}$±0.29 x $10^{-8}$ | 7.46 x $10^{-9}$±0.11 x $10^{-9}$ | 4.19 x $10^{-9}$±0.95 x $10^{-9}$ | 8.59 x $10^{-8}$±0.33 x $10^{-8}$ |
| ***G. sulph.* Protein Coding** | 1.65 x $10^{-8}$±0.39 x $10^{-8}$ | 1.30 x $10^{-9}$±0.56 x $10^{-9}$ | 4.12 x $10^{-9}$±0.11 x $10^{-9}$ | 2.19 x $10^{-8}$±0.48 x $10^{-8}$ |
| ***Ostreococcus tauri*** | 4.19 x $10^{-10}$ | N/A | 0.60 x $10^{-10}$ | 4.79 x $10^{-10}$ |
| ***Chlamydomonas reinhardtii*** | 2.08 x $10^{-10}$ | N/A | N/A | 3.23 x $10^{-10}$ |
| ***Haloferax volcanii*** | 3.15 x $10^{-10}$±0.27 x $10^{-10}$ | N/A | 3.58 x $10^{-11}$±0.82 x $10^{-11}$ | 3.24 x $10^{-10}$±0.28 x $10^{-10}$* |

*Table 20: Whole genome and genomic region mutation rates for G. sulphuraria, and genome wide mutation rates for Ostreococcus tauri* [220]*, Chlamydomonas reinhardtii* [221]*, and the halophilic archaea, Haloferax volcanii* [222]*. All units are mutations per site per generation.*

The nucleotide substitution, multi-nucleotide variation, and indel rates for the duplicated and non-duplicated protein coding regions are shown in *Table 20*. The total mutation rate for non-duplicated protein coding regions is $2.19 \times 10^{-8}$ $(SE = 0.48 \times 10^{-8})$ per site per generation, whereas the total mutation rate for duplicated protein coding regions was nearly four times higher at $8.59 \times 10^{-8}$ $(SE = 0.33 \times 10^{-8})$ per site per generation, marking a significant difference (independent T-test: $T = 10.24,\ p = 2.89 \times 10^{-14}$) between the overall rate of mutation in duplicated and non-duplicated protein coding regions in the *G. sulphuraria* SAG 107.79 genome. The difference between the overall mutation rate for duplicated protein coding regions and the genome wide mutation rate was statistically significant (independent T-test: $T = 11.11,\ p = 1.43 \times 10^{-15}$), whereas there was no

significant difference between the non-duplicated protein coding regions and the overall mutation rate (independent T-test: $T = -1.79,\ p = 0.079$).

Mutations were distributed evenly across duplicated coding regions (*Figure 19*). The majority of mappable duplicated coding regions contained at least 1 mutation across all samples, and there was no clear concentration of mutations over a particular region or chromosome within the duplicated coding regions. The random sample of non-duplicated coding regions is distributed evenly across the genome, and is representative of the distribution of coding regions.

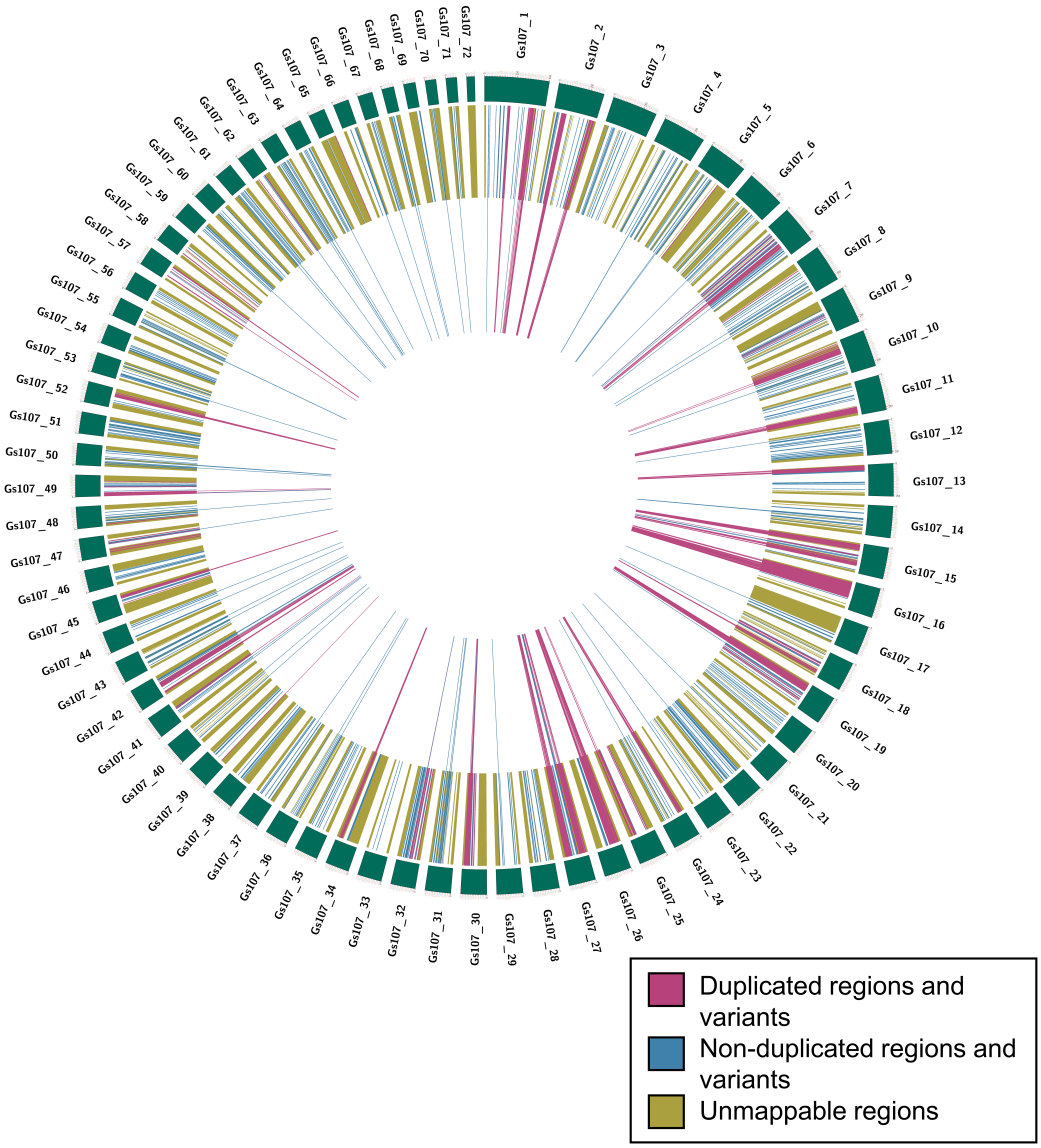



*Figure 19: Visualisation of the unmappable regions, and the duplicated and non-duplicated protein coding regions used in this analysis (outside track), and the variants found in the duplicated and non-duplicated protein coding regions (inside track).*

### Haplotype Loss was Observed in Mutation Accumulation Lines

The resequencing of *G. sulphuraria* SAG 107.79 mutation accumulation lines led to accidental observation of haplotype loss in the genome. This means that recombination events occurred within the genome, yet one of the recombination products was lost, and only one was retained (*Figure 20*), confirming the assertion made in "*Chapter 3: Understanding The Meiotic Capacity of G. sulphuraria: A Genomics Based Approach*", that *G. sulphuraria* is undergoing recombination.

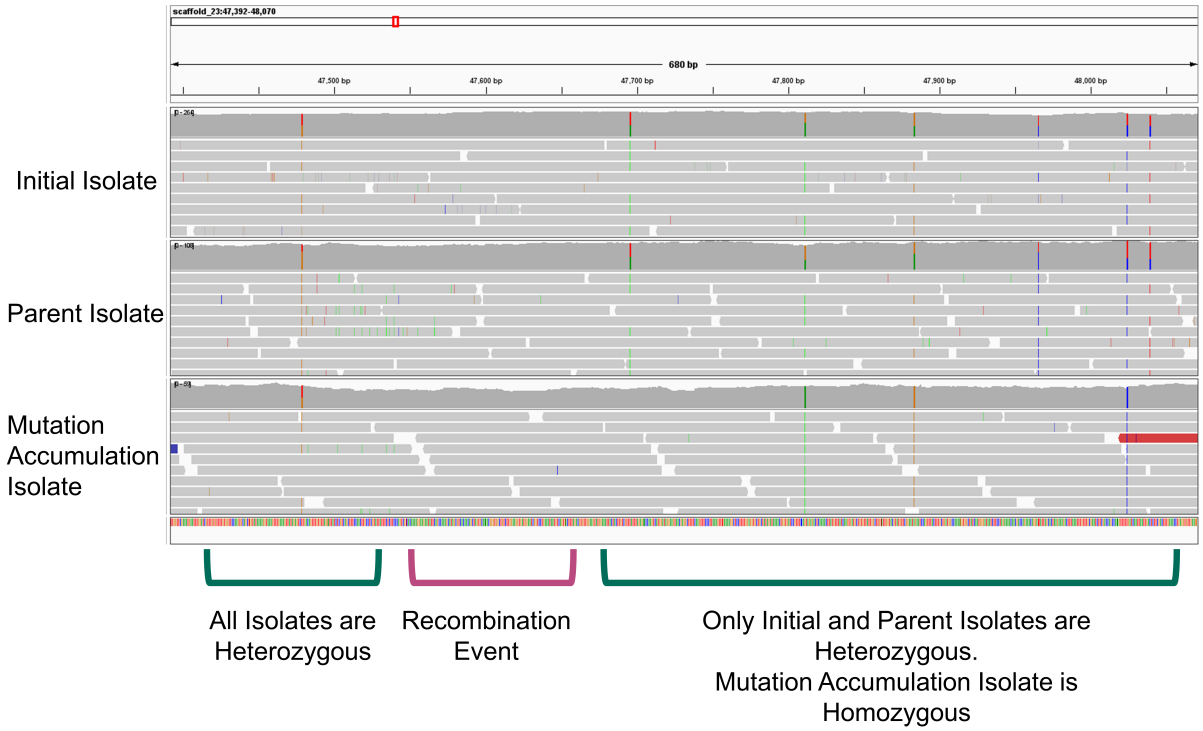



*Figure 20: Illumina sequencing reads showing in an example from chromosome Gs107_23 of haplotype loss having occurred between the sequencing of the parent isolate of the mutation accumulation experiment, and a final mutation accumulation isolate. The "initial isolate" is the SAG 107.79 Illumina sequencing reads described in "Chapter 2: Assembly, Annotation, and Comparison of Complete G. sulphuraria Nuclear Genomes".*

### Transcript Abundance of Duplicated Genes

Data from the yeast genome suggest that more highly expressed genes are retained in duplicate [223]. Conversely, the *G. sulphuraria* duplicated genes are significantly less expressed than the control non-duplicated genes (independent T-test: $T = -3.74,\ p = 1.95 \times 10^{-4}$). Mean expression levels were approximately double in the control set of genes (0.022 ± 0.0025) compared to the duplicated genes (0.012 ± 0.0010).

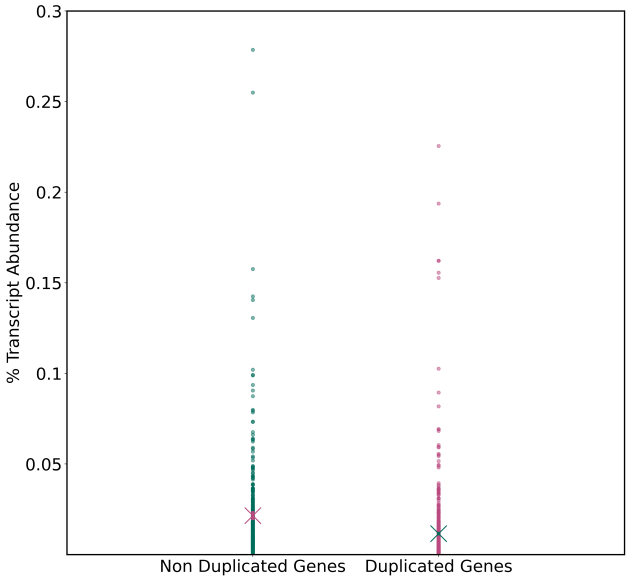



*Figure 21: % Transcript Abundance between duplicated and non-duplicated protein coding genes.*

## Discussion

Spontaneous mutation is a major source of novel genetic variation in nature, and is influenced by a variety of biological processes (prevention, production, and repair) [222]. These processes differ among varying genetic circumstances, and this variation leads to biases in the location, type, and number of accumulated mutations [212].

In terms of spontaneous mutation, extremophiles are an interesting case study, since the optimal conditions in which they grow in nature are often DNA-damaging, yet also rapidly changing. For extremophiles, the balance between maintaining genome stability and allowing for mutations, and therefore adaptations, to occur, is especially pertinent. Uncovering the mutation rate of these organisms further unveils how the not only tolerate, but thrive in extreme environments.

*G. sulphuraria* is unique since it is a eukaryotic polyextremophile that is also metabolically flexible, conducting both photosynthesis as well as having the ability to utilise a variety of complex carbon compounds for growth. While investigations have indicated that horizontal gene transfer has played a major role in the adaptation of *G. sulphuraria* its extreme habitat [35], [109], [110], [115], the impact of spontaneous mutation had not been investigated until now. Acknowledging that there was significant variation in the number of cells per colony and the resulting growth rate estimation, comparing the *G. sulphuraria* mutation rate to that of other species

should be treated with some caution. Nevertheless, the estimation of the *G. sulphuraria* mutation rate presented here is two orders of magnitude higher than that of the algae *Chlamydomonas reinhardtii* and *Ostreococcus tauri* [220], [221]. This 100-fold difference cannot be accounted for by variations in colony size and represents an extraordinarily fast rate of mutation for a micro-organism, indicating that *G. sulphuraria* is evolving very quickly.

Microbial communities have been reported to evolve faster in extreme environments [19]. This result was contrasting with previous reports that (hyper)thermophilies exhibited lower mutation rates than mesophiles [224], however, it has been suggested that this could be due to the unusual evolutionary pattern of (hyper)thermophiles such as distinct mutational spectra [225][226], and mutation repair strategies [20]. Moreover, these studies focused on a single environmental factor (temperature) and did not account for polyextremophily [19]. The disparity in spontaneous mutation rates between *G. sulphuraria* and mesophilic micro-organisms could reflect their distinct evolutionary history. Mesophiles are more likely to have reached a steady state of environmental adaptation, therefore most mutations may have deleterious or neutral effects on fitness and are less likely to be fixed in populations. On the other hand, the fitness of polyextremophiles such as *G. sulphuraria* may never be optimal for the rapidly changing extreme environment in which they inhabit, so adaptive evolution is expected to occur more frequently.

Higher spontaneous mutation rates are reported to correlate with smaller genome sizes [227]. The Rhodophytes have compact genomes [102], and *G. sulphuraria* is no exception to this, however the small genome size does not solely explain the high mutation rate of *G. sulphuraria*, especially considering that the mutation rate of the smallest free living eukaryote, with a genome size of 13 Mb [204], *Ostreococcus tauri*, is 100 fold lower than *G. sulphuraria* [220].

Mutational hotspots are regions of a genome which exhibit higher rates of spontaneous mutation than the rest of the genome. It was previously assumed that spontaneous mutations would occur across a genome in an unbiased fashion, however this has been shown to be untrue, with many species exhibiting mutational hotspots within their genomes. Many parasite genomes, such as *Neospora caninum*, *Plasmodium falciparum*, and *Toxoplasma gondii*, possess mutational hotspots [228]–

[230]. Though not extremophiles, in order to be successful, the Apicomplexa must evade an adaptive host immune response, and it is reported that these organisms maintain higher rates of spontaneous mutation on subtelomeric genes which encode antigens [231], representing an adaptive response to host adaptive immunity . Here I have shown that *G. sulphuraria* also has mutational hotspots, and that the regions with higher mutation rates are duplicated genomic regions. In duplicated protein coding regions, a deleterious mutation in a gene may not pose a significant disadvantage since there is a functional copy of the gene elsewhere in the genome, but maintaining a high mutation rate in that region also increases the likelihood of a beneficial mutation taking place, providing *G. sulphuraria* with a preemptive adaptive strategy to its rapidly changing extreme environment. Similar to parasites, *G. sulphuraria* also possesses a superfamily of genes that localise to the subtelomeres, the Archaeal ATPases. This superfamily is the largest superfamily of genes within the *G. sulphuraria* genome, with the *G. sulphuraria* 074W reported to possess over 100 Archaeal ATPases [110], and this is supported in the *G. sulphuraria* ACUF 138 genome. The exact function of these genes remains unknown, although higher Archaeal ATPase copy number is associated with higher optimum growth temperatures in archaea [110]. Due to the large number of long, repetitive subtelomeric regions in the *G. sulphuraria* genome, the subtelomeres are classified as difficult to map regions and were therefore excluded from this analysis. As sequencing technologies improve, I hope that the community will be able to circumvent this problem by employing newer, more accurate long read sequencing methods to measuring the accumulation of mutations, as these sequencing reads would be able to accurately map to long, repetitive, subtelomeric regions that, owing to the repetitive nature of these regions, short sequencing reads map to less accurately. As a hypothesis, given that the Archaeal ATPases exist in such high copy number, it is likely they also exhibit higher mutation rates than the rest of the genome, potentially acting as contingency genes.

The duplicated genes that make up the mutational hotspots within the *G. sulphuraria* genome are expressed, on average, half as much as non-duplicated genes. This contradicts data from other species [223] that suggest that maintained duplicated genes are often more highly expressed. Conversely, the data presented here suggest that either the non-duplicated genes are being actively overexpressed, or

that the duplicated genes are actively expressed less, in order to maintain an overall balance of transcript abundance. As duplicated genes are lost due to the eventual accumulation of deleterious mutations, *G. sulphuraria* must then rebalance transcript levels by increasing the transcription of the gene that was duplicated but whose duplicate was lost. As the duplicated genes are in colinear blocks, it is likely that the transcriptional factors associated with the duplicated genes are also duplicated and therefore aid in the balancing of transcription across these duplicated regions, however should a deleterious mutation occur in the gene encoding the transcription factor, and not the gene encoding the enzyme (or vice-versa), *G. sulphuraria* must have a way of managing the resulting transcriptional imbalance.

Although spontaneous mutation is a major evolutionary driver, the extraordinary rate at which *G. sulphuraria* is accumulating mutations should come at an enormous cost to the organism. It is noted that the vast majority of mutations are deleterious, therefore the chance of complete loss of function mutations occurring within this genome is high. For costs of this level of mutation to be met, the high mutation rate must offer a significant benefit to the organism.

In this chapter I have demonstrated an additional mechanism that could be implicated in extremophily in *G. sulphuraria*, and shown that *G. sulphuraria* is possibly one of the fastest evolving free living eukaryotes in nature. Owing to the evolutionary cost of a high mutation rate, maintaining this level of mutation must be of benefit to *G. sulphuraria*, and this could come in the form of enabling rapid adaptive evolution to rapid environmental changes.

# Chapter 5: Discussion

*G. sulphuraria* is a remarkable organism. Particularly, it is its genome that makes *G. sulphuraria* truly astounding. I have demonstrated that *G. sulphuraria* has a compact nuclear genome, with numerous tiny chromosomes, numerous tiny chromosomes. These chromosomes appear to have undergone recombination events, resulting in islands of duplicated genes, and chromosome structural differences between different *G. sulphuraria* isolates. Moreover, the genome exhibits extremely high spontaneous mutation rates. These interesting genomic features improve our understanding of *G. sulphuraria*, and demonstrate how *G. sulphuraria* can maintain an adaptive, flexible, lifestyle in its extreme environment.

The biotechnological potential of *G. sulphuraria* is perhaps what has been used to justify the extensive whole genome sequencing that has been conducted on this organism both as a part of this project, and sequencing projects in groups around the world. Its metabolic flexibility, and wide ranging extremophilic traits, making it robust to environmental change, but also conferring metal scavenging abilities, have led to *G. sulphuraria* being a superb candidate organism for a broad range of industrial purposes, including remediation, pigment production, and lignocellulose degradation. The improved understanding of the *G. sulphuraria* genome detailed in this thesis may aid in the optimisation of *G. sulphuraria* for use in biotechnology, or provide a solid baseline for gene discovery within the alga with the benefit of complete genome annotations. However, I do not feel that the industrial applications of *G. sulphuraria* is necessary as a justification to defend the effort and costs of thorough whole genome sequencing. The unusual characteristics of the *G. sulphuraria* genome, and their implications of the understanding of the mechanisms of extremophily, are justification enough.

*G. sulphuraria* is extremely successful in its environment, making up 90% of the total biomass and almost all of the eukaryotic biomass. It is also anciently diverged, with the Cyanidiophyceae branching from the Rhodophytes ~1.5 billion years ago. *Galdieria* species and their most “recent” ancestors have thrived in these extreme environments for over 1 billion years, without becoming extinct. From a perspective of scientific curiosity, the sequencing of a collection of *G. sulphuraria* isolates has uncovered a multitude of interesting genomic traits that perhaps begin to explain how

this organism has remained so successful throughout history. These genomic features are the topic of discussion in this thesis.

**Adaptive Evolution as a Mechanism for Extreme Adaptation**

*G. sulphuraria* is unique. Being an extreme eukaryote that resides in a variety of rapidly changing hot, acidic, and heavy metal rich habitats, with the ability to grow both phototrophically and heterotrophically on a plethora of complex carbon sources, *G. sulphuraria* is particularly devoid of suitable organisms to compare it to, on both a genomic and ecological level. The closest neighbour to *G. sulphuraria*, *C. merolae*, shares no collinear regions in the genome, and is an obligate phototroph that does not occupy the most extreme habitats that *G. sulphuraria* is so successful in. Moreover, despite the completion of the *C. merolae* genome, very little is known about its lifecycle, nor much about the life cycle of the Cyanidiophyceae in general, likely due to the sexual portion (should there be any) of the life cycles only occurring under specific conditions, as is the case with *G. sulphuraria*, as demonstrated by Hirooka et al. [124].

Comparisons of *G. sulphuraria* to other extremophiles are tenuous, as the Cyanidiophyceae are the only eukaryotic extremophiles, the rest being bacteria and archaea. Bacteria and archaea are completely different biologically, with circular genomes, no organelles, and no alternative splicing, to name a few differences. Prokaryotes also have different mechanisms of rapid adaptation – namely horizontal gene transfer. Horizontal gene transfer is thought to play such a major role in prokaryote evolution that there is doubt as to whether prokaryotic phylogenies can be resolved with any accuracy [16]. Although horizontal gene transfer has been reported to play a role in *G. sulphuraria* adaptation, the level of horizontally acquired genes in the genome is relatively low compared to a prokaryote. Horizontally gene transfer candidates making up 1 – 5 % of protein coding genes does not explain alone how *G. sulphuraria* can adapt to extreme environments. There is not a confirmed mechanism of how *G. sulphuraria* rapidly acquires genes horizontally. I hypothesise that *G. sulphuraria* acquires novel genetic material during the haploid phase, as the cell wall is absent during this phase and DNA uptake is more probable. Noting that although horizontal gene transfer has played some role in extreme adaptation in *G. sulphuraria*, in contrast with prokaryotes the vast majority of genes have been vertically inherited in *G. sulphuraria*. As a result, the evolutionary

mechanisms driving adaptation of *G. sulphuraria* remain distinctly different to those of the prokaryotes that share its habitat.

Considering the rapidly changing extreme environment *G. sulphuraria* thrives in, a useful comparison that does continue to appear is between *G. sulphuraria* and protozoan parasites, such as *Plasmodium* and *Leishmania*. Parasites face a different kind of “extreme environment” – the host immune system. Like *G. sulphuraria*, these organisms have a fairly concise gene complement (between ~5000 and ~8000 genes), although they have slightly larger genomes than *G. sulphuraria* (22 Mb and 32 Mb) [232][233]. The parasitic lifestyle is an interesting point of comparison to that of *G. sulphuraria*. To enable survival, parasites must rapidly and constantly adapt in order to evade the hosts constantly adapting immune response, and consequently have evolved various strategies of adaptive immune evasion, including subjecting subtelomeric DNA, harbouring the genes which encode antigens, to higher spontaneous mutation rates [234][231]. *G. sulphuraria* faces similar challenges, though not from an adaptive immune response but from a rapidly changing hot and acidic environment. The main similarity between *G. sulphuraria* and the protozoa I described, is that they are almost never perfectly adapted to their environment and therefore can never reach a steady state in their evolution, because by the time they have adapted to one change in their environment, the next environmental change begins to occur. I therefore suggest that the consequences of these evolutionary pressures on the genomes of *G. sulphuraria* and protozoan parasites may be similar.

*G. sulphuraria* often inhabits the most extreme environments within the broad range of habitats that the Cyanidiophyceae inhabit. It is the rapid rate and variety of extreme environmental change that *G. sulphuraria* is exposed to that makes it special – other Cyanidiophyceae reside in much more stable environments. Examples of these rapid changes may include; rapidly increasing water temperatures prior to a volcanic event, rapid variations in water acidity, variations in heavy metal content and concentration, and desiccation. These are all stresses that constantly challenge the fitness of *G. sulphuraria*. *G. sulphuraria* is the only eukaryotic species to operate in these environments, and if it were not able to rapidly adapt, it simply would not be able to survive these environmental changes, resulting in extinction.

Similar to parasites and their hosts, while the environment remains so volatile *G. sulphuraria* cannot outpace the rapid environmental changes it faces. Parasites co-evolve with their hosts, trapped in a constant cycle of adaptive evolution that enables evasion of a constantly adapting host immune response. The environment that *G. sulphuraria* resides in does not adapt in response to *G. sulphuraria*, however it is constantly changing. *G. sulphuraria* therefore represents a unique model in which to study adaptive evolution in response to a rapid, and potentially lethal, abiotic change. In this thesis I have elucidated some of the mechanisms of adaptive evolution in *G. sulphuraria*.

**Karyotype and Ploidy**

Karyotype is the form and number of chromosomes an individual species exhibits. A diverse range of karyotypes exist in the tree of life. Prokaryotes have single, circular, chromosomes. Eukaryotic chromosomes are linear, and vary greatly in size and number, with some species, such as the male jack jumper ant, *Myrmecia pilosula*, having a single chromosome [235], and others having up to 226 chromosome pairs, which is the highest number of chromosomes recorded in a diploid, in the atlas blue butterfly, *Polyommatus atlanticus* [236]. Karyotype can be determined through microscopy, or pulse field gel electrophoresis, however for organisms with many small chromosomes that are difficult to visualise, such as *G. sulphuraria*, long read sequencing is being increasingly applied to this area in order to estimate the size and number of chromosomes in a given species.

The completed nuclear genome assemblies have shown clearly that *G. sulphuraria* has a multitude of small chromosomes. These vary in size and number between different *G. sulphuraria* isolates. Based on the currently completed *G. sulphuraria* genome assemblies presented here, and in recently published data from Hirooka et al. [124], the *G. sulphuraria* genome size and chromosome number ranges from 13 Mb to 16.5 Mb, and 72-76 chromosomes, isolate dependent. Although Hirooka et al. reported 80 telomere-to-telomere chromosomes [124], four pairs of these are haplotigs, and for this estimation they were removed as they do not represent the number of chromosomes in the *G. sulphuraria* haploid genome, hence I describe the number of haploid chromosomes as 76 in this genome.

Using DNA sequencing to determine karyotype still has limitations. Firstly, without significant coverage of the longest sequencing reads (15 Kb < in the case of *G. sulphuraria*), genome assemblers struggle to correctly assemble contigs telomere to telomere. This is because subtelomeres consist of repetitive, shared DNA, that often is not unique to a specific chromosome and instead is shared among all subtelomeres in all chromosomes. Therefore, to correctly assemble each subtelomere, sequencing reads must span from the telomere, through the subtelomere, and into the unique regions of the chromosome, in order for correct assembly. Secondly, genome assemblies do not tell us what the chromosomes actually look like in terms of structure and organisation within the *G. sulphuraria* cell. The assemblies only tell us the length and DNA content of the chromosomes.

To understand the chromosomal architecture on a cellular level, microscopy must be employed. Muravenko et al. attempted to visualise *G. sulphuraria* chromosomes using light microscopy and concluded that *G. sulphuraria* had 2 chromosomes, however these images were scaled to 3 μm [237]. At this scale, chromosomes from model organisms have been visualised, demonstrating how nuclear DNA is packaged in these species [238], yet chromosomes in these model genomes, such as the human genome, are an order of magnitude larger than *G. sulphuraria's* longest chromosome. The human chromosome 1 is 249 Mb long [239], almost 20 times larger than the entire *G. sulphuraria* genome, let alone the longest *G. sulphuraria* chromosome. At 3 μm scale, *G. sulphuraria* chromosomes of 100-500 Kb in length are not going to be distinguishable. To investigate the organisation and packaging of *G. sulphuraria* chromosomes, more advanced microscopy techniques that operate at the nanoscale, such as stimulated emission depletion nanoscopy, should be used [238]. Recruiting advanced microscopy techniques to karyotyping *G. sulphuraria* could further uncover how this multitude of small chromosomes supports *G. sulphuraria*'s adaptive, flexible lifestyle, as well as understanding how these small chromosomes are packaged.

The packaging of *G. sulphuraria* chromosomes is particularly interesting given the presence of predicted reverse gyrases in the genome. These enzymes facilitate genome stability of hyperthermophiles, and enable archaeal hyperthermophiles to maintain AT rich genomes [18]. It has even been suggested that reverse gyrase is essential for hyperthermophilic life, with deletion of the reverse gyrase gene resulting

in loss of hyperthermophily in *Pyrococcus furiosus,* with growth inhibited above 95 ˚C [240]. Interestingly, reverse gyrase has been considered a hyperthermophile specific protein and therefore the reverse gyrase gene should not be present in the genome of a moderate thermophile, *G. sulphuraria* [241]. In order to confirm reverse gyrase activity, the reverse gyrase candidates I have described in "*Chapter 2: Assembly, Annotation, and Comparison of Complete G. sulphuraria Nuclear Genomes*" should be biochemically assessed for DNA supercoiling activity and ATP-dependency, since bioinformatics alone cannot confirm their activity, only predict it. Moreover, the recent establishment of transformation methods that can produce knockout *G. sulphuraria* cells [124] will enable the impact of this putative reverse gyrase on thermophily in *G. sulphuraria* to be examined. Confirmation of reverse gyrase activity would challenge the consensus that these are the only hyperthermophile specific proteins, and would demonstrate that reverse gyrase mediated DNA supercoiling is a key mechanism for thermophily in *G. sulphuraria*, and explain the maintenance of an AT rich genome in this thermophile.

Ploidy refers to the number of copies of chromosomes in an individual cell. Polyploids have more than two copies of every chromosome (diploids have two copies), and ploidy levels can vary within the cell cycle. The available evidence points to *G. sulphuraria* being at least diploid, however the possibility of a more diverse ploidy level has not been discussed. Oxford Nanopore Technologies (ONT) sequencing coverage varied across different *G. sulphuraria* chromosomes in all assemblies. Assuming that the number of ONT sequencing reads is proportional to the amount of DNA provided to the sequencer, this would indicate that there are more copies of certain chromosomes within the *G. sulphuraria* genome. The would also, in part, explain why genome assemblers failed to resolve all haplotypes, since they were assuming diploid ploidy levels, where there may be aneuploidy. Aneuploidy is a phenomenon where the number of chromosomes in the "diploid" is not an exact multiple of the number of chromosomes in the haploid. Aneuploidy usually results from errors during meiosis leading to the formation of gametes with abnormal numbers of chromosomes, and then when these gametes fuse with gametes with a normal number of chromosomes, the resulting cell will have an abnormal number of chromosomes for what should otherwise be a diploid. Usually this is a case of a pair of homologous chromosomes having an extra chromosome,

and in terms of genome assembly, could lead the genome assembler to resolve two contigs, instead of one. Biologically, aneuploidy is not well tolerated in higher organisms. In humans, aneuploidy is detrimental and incompatible with survival, aside from a few exceptions that are always accompanied with various pathologies [242]. On the other hand, aneuploidy can be tolerated and occurs naturally in some unicellular eukaryotes, and other several multicellular species, and is possible that *G. sulphuraria* tolerates aneuploidy also.

In fungi, has been a lot of work on ploidy variation, particularly as it relates to the sexual life cycle and timing of meiosis and syngamy. The budding yeast *Saccharomyces* are prone to mating whenever compatible cells encounter one another are primarily diploid, whereas fission yeast, *Schizosaccharomyces*, that sporulate soon after mating, are mostly haploid. These assumptions have been challenged after the sequencing of additional natural isolates of Saccharomyces cerevisiae, revealing extensive variation in ploidy level, with diploid (31%), triploid (10%), and tetraploid lineages (59%) [243]. Widespread genomic analyses suggest that species have undergone ancient polyploidisation events, resulting in duplicated genes [244]. This is a possible explanation to the initial gene duplications seen within the *G. sulphuraria* genomes, in that a polyploidisation event took place, creating additional copies of certain chromosomes, followed by non-homologous recombination of these duplicated chromosomes to other chromosomes in the genome, resulting in a novel chromosome. The persistence of these duplicated genes within the genome is what is especially curious, however.

There are three main theories that explain the persistence of duplicated genes. 1) duplicated genes persist when the gene copies are immediately and actively preserved, 2) persisting duplicates are genes that take on novel, advantageous, functions before decay (neofunctionalisation), and 3) surviving duplicate genes are those that lose nonoverlapping functions, meaning that both copies must be retained for full gene function [244]. The widespread gene duplication is in the form of duplicated collinear blocks of genes, meaning that the duplicated genes are linked. Given that the duplicated genes of *G. sulphuraria* are subject to significantly higher mutation rates than non-duplicated genes, these genes are unlikely to be immediately and actively preserved. So what of neofunctionalisation and the loss of non-overlapping functions? Both of these explanations are possible, and likely have

occurred within some duplicated genes, however there is another potential explanation. A duplicated gene, on average, had half as much transcript abundance when compared with non-duplicated genes. If genes had taken on new functions, or non-overlapping functions had been lost, one would expect the transcript abundance of the functioning copy of a gene to increase, in order to produce the correct amount of functioning protein. The evidence indicates that this is not occurring as if it were, there would be an increased variation in transcript abundance among duplicated genes, while the mean transcript abundance still remained lower as non-functioning gene copies became less expressed. This perhaps suggests that the duplicated gene blocks are maintained within the genome as a contingency mechanism, so that if loss of function mutations occur within these genes, or if there is a loss of the chromosome altogether, there is at least one other functioning copy of the gene elsewhere in the genome.

On the diversity of karyotype between *G. sulphuraria* strains, it is interesting that *G. sulphuraria* has recently been demonstrated to form gametes that can undergo self-diploidisation. Consider a scenario where a *G. sulphuraria* diploid cell with 72 chromosome pairs has undergone meiosis, but that the homologous chromosomes have segregated in error and the resulting haploid cells have 70, and 74 chromosomes. Should these haploid cells undergo self-diploidisation, the resulting diploids would have 70 and 74 chromosome pairs. Given the extraordinarily high substitution rate of this organism, genes on these extra chromosomes in the 74 chromosome diploid could quickly evolve beneficial new functions (neofunctionisation), and begin to be maintained within the genome. This could explain the diversity in the structure and number of chromosomes between *G. sulphuraria* strains.

Since *G. sulphuraria* haploids can undergo self-diploidisation, it may be possible that diploid *G. sulphuraria* cells can undergo endoreplication, during which the entire genome is duplicated within the cell, without the cell undergoing cell division, leading to polyploidy. It has been reported in other systems, particularly asexual systems, that polyploidy followed by homologous recombination is a method by which these systems avoid mutational meltdown in the absence of sex [245]. Since there are multiple copies of each chromosome, a mutation can be identified and repaired as there are many copies of the wild type chromosome. These mechanisms serve to

lower the mutation rate, and considering that *G. sulphuraria* has an extremely high mutation rate, it raises doubt whether polyploidisation as a regeneration mechanism is occurring. Alternatively, *G. sulphuraria* could use polyploidisation followed by recombination to repair mutations under a specific set of conditions, similar to how the formation of haploids is only stimulated by certain conditions [124].

Haplotype loss is also occurring within the *G. sulphuraria* genome, further supporting the notion that ploidy variation could be taking place, but further investigations are needed to fully understand how variations in ploidy aid in *G. sulphuraria*'s adaptive lifestyle.

**Mating and Speciation**

Of the *G. sulphuraria* isolates for which there is long read sequencing data and assemblies, no two isolates have the same chromosome structure. These structural differences are not insignificant, as some assemblies have chromosomes that are twice as long as the longest reported chromosome in another assembly. This then begs the question, are these isolates capable of mating with each other? And if not, is it then right to assign these to the same species? Species are often defined as groups of organisms capable of interbreeding and producing fertile offspring.

The genus of *Leishmania*, the protozoan parasites responsible for the neglected tropical disease Leishmaniasis, contains over 20 species with karyotypes varying from 34 to 36 chromosomes. Some of these species have been separated for 20-100 million years, yet retain a high degree of synteny [246]. The life cycle of protozoa are very different to red algae, but it is interesting that there is such a wide range of *Leishmania* species, even though the karyotypes are much similar and the chromosomes are more syntenic, while all of the *G. sulphuraria* lineages are designated as a single species.

Here, I seek to examine in terms of the karyotypes of these isolates, whether it is possible that two *G. sulphuraria* isolates from different lineages could mate. Since these lineages are geographically isolated, there will never have been an opportunity for these inter-lineage mating events to have taken place recently in nature, yet the ability to form *G. sulphuraria* crosses could prove useful for creating new *G. sulphuraria* lines with specific desired traits for industry (such as improved efficiency

at degrading a certain carbon source, improved remediation abilities, increased temperature tolerance etc.).

In the case of the *G. sulphuraria* isolates assembled and discussed in this thesis, ACUF 138, ACUF 017, and SAG 107.79, the longest chromosome of SAG 107.79 is 114.5 Kb longer than the longest chromosome of ACUF 138, and there are significant structural rearrangements between some chromosomes. Therefore, could these two karyotypes mate to produce fertile offspring? I present two scenarios upon fusion of SAG 107.79 and ACUF 138 gametes, 1) the resulting karyotype leads to cell death as there is too much copy number imbalance or essential genes are absent, 2) endoreplication occurs, creating a tetraploid genome, and then chromosomes are selectively or randomly segregated at cell division, forming two diploid cells, possibly with an entirely novel karyotype. Thinking about *Galdieria* in terms of strict mendelian inheritance would lead you to believe that only the first scenario could be true. However, this would then mean that at least every *G. sulphuraria* lineage is its own species, if not more. In the second scenario, different lineages may be able to successfully mate, producing a novel *G. sulphuraria* isolate. Endoreplication does not create an imbalance as the entire genome is duplicated within the cell, and is usually well tolerated, however in this hypothetical case the resulting cell would have two different *G. sulphuraria* genomes, and there would be a degree of imbalance in terms of gene copy number because the same syntenic gene blocks may not be duplicated in both genomes.

The plastid, mitochondrial, and nuclear phylogenies have been shown to be incongruent [120], meaning that each genome has taken a different evolutionary path. However, they always segregate into the same six lineages, and isolates do not move into different lineages for different phylogenies. This is what you would expect for distinct species, but this could be explained solely by geographic location and isolation of the lineages. Within the lineages, the three phylogenies are also incongruent. The single lineage linkage disequilibrium data did not support recombination, but this could be explained by a low number of samples and resulting variants. Long read sequencing of more recently separated strains could further demonstrate how genome structural diversity has evolved, and may aid in the determination of *Galdieria* species through karyotyping.

Now that protocols for the formation of *G. sulphuraria* gametes are forming, the community will soon have the ability to test for these cryptic species by co-culturing different strains and attempting to form crosses. The genome sequencing of any potential crosses could also elucidate the mechanisms for *G. sulphuraria* genome structural diversity, and determine if there are any biases in where structural variation takes place in the genome – similar to the biases in mutation rate I observed in "*Chapter 4: The Spontaneous Mutation Rate of G. sulphuraria SAG 107.79".*

**Recombination, Gene Duplication, and the Mutational Load**

I have reported here that *G. sulphuraria* has an extraordinarily fast rate of mutation. It is surprising that given this rate of mutation, and acknowledging that the vast majority of mutations are deleterious, that mutational meltdown has not occurred. I believe it would be useful to repeat the mutation accumulation experiment over a significantly longer time period (2 years +), and take intermediary samples every 3-6 months for sequencing. Anecdotally, there was no degradation in colony size, or fitness over the mutation accumulation period, but over longer time periods, when more mutations have been allowed to occur, this may not be the case. By repeating the experiment in the aforementioned manner, there can be a robust assessment of whether mutational meltdown does occur, how long it takes for it to occur, and how many mutations a *G. sulphuraria* cell can tolerate before it occurs.

Meiosis has been shown to occur under a specific set of conditions, however *G. sulphuraria* has been cultured continuously for years in stock centres and laboratories usually at 37 ˚C, in pH 0-2 Allen's media, sometimes supplemented with sucrose. These isolates have not been subject to any specific set of conditions that could signal the induction of meiosis, endoreplication, or any other genome regeneration pathway. I therefore wonder if the organism has a way of detecting when too many deleterious mutations have occurred (perhaps as fitness degrades), and then inducing genome regeneration.

Mutation rate is negatively correlated with genome size, and based on a genome size of 13 -16 Mb, the *G. sulphuraria* mutation rate should be in the range of $1 \times 10^{-10}$ to $5 \times 10^{-10}$ [227]. The estimation calculated in "

 is 100 times higher than this value at 3.19 x $10^{-8}$. The protozoan parasites also do not strictly fit the correlation between genome size and mutation rate, exhibiting mutation rates 10 times higher than expected for genomes of that size. The *G. sulphuraria* spontaneous mutation rate is closer to that of the eubacteria *Mesoplasma florum*, and some double stranded viruses [227]. *M. florum* demonstrates slightly higher mutation rates than expected for its genome size. This is particularly interesting as this species are part of the class Mollicutes, that developed from Gram-positive bacteria through reductive evolution and adopting a parasitic lifestyle, however *M. florum* is not parasitic nor pathogenic, but evolved from parasitic pathogens and adopted a free living lifestyle [247]. The common ancestor of *M. florum* and the *Mycobacterium* was likely undergoing Red Queen-like evolution. When scaled for genome size, the mutation rate of this eubacteria is not significantly higher or lower than parasitic pathogenic eubacteria of the same class [227]. This could indicate that once an organism evolves mechanisms to enable a high mutation rate, it retains that high mutation rate even if the environmental conditions no longer require such a rapid rate of evolution, which could also explain why domesticated *G. sulphuraria* isolates still exhibit high rates of mutation.

The maintenance of duplicated genes within the genome, and the fact that these genes appear to be subject to higher mutation rates than non-duplicated protein coding genes, does support the notion that perhaps a portion of these genes are contingency genes, or that *G. sulphuraria* is capable of maintaining selectively higher mutation rates in certain genomic regions and is therefore capable of having contingency genes in genomic regions that were not mappable in these experiments. There is certainly more scope for further probing these biases in the accumulation of mutations, particularly in the identification of mappable regions where very few mutations accumulate. I predict that the regions with the lowest rate of mutation will be the genes identified in the *G. sulphuraria* pangenome in Iovinella and Lock [120], as these are the genes that are well conserved between all sequenced *G. sulphuraria* isolates included in the study, and are therefore likely the slowest evolving genes. The extraordinarily high mutation rate the *G. sulphuraria* nuclear genome appears to exhibit, without undergoing mutational meltdown, remains an enigma.

In addition to the cost of maintaining this high mutation rate, *G. sulphuraria* also undergoes the costly processes of recombination and sex. The RQH suggests that sexual reproduction is necessary to outpace parasitism, and this could perhaps apply to organisms with extremely flexible and adaptive lifestyles such as *G. sulphuraria*. This ties in with the fitness associated recombination model, in which organisms that are not in an optimal state of fitness for their environment invest more into sexual reproduction. This is the model that I believe applies to the *G. sulphuraria* genome and best explains the maintenance of sex in the genome.

**Is Genomic Plasticity Solely Due to Extremophily?**

The genomic data I have detailed in this thesis demonstrate the remarkable plasticity of the *G. sulphuraria* nuclear genome. The mitochondrial genome has also been reported to have many extraordinary features including extreme GC skew and increased mutation [44]. It demonstrated that two types of mitochondrial genome exist within the Cyanidiophyceae, those of *Galdieria* species, G-type, and those of the other Cyanidiophyceae species, C-type. G-type genomes exhibited strand skew, higher GC content, higher repeat content, and were shorter than C-type genomes.

There are two main differences between G-type and C-type species. G-type species inhabit more extreme environments, and are metabolically flexible, while C-type species inhabit less extreme environments, and are obligate phototrophs. The *G. sulphuraria* plastid genomes do not exhibit any unusual features, and *G. sulphuraria* conducts photosynthesis normally like the other Cyanidiophyceae. So, is it extremophily that solely explains the extreme features of both the mitochondrial and the nuclear genome, or is it metabolic flexibility? I suspect that it is a mixture of both factors on both genomes, while metabolic flexibility and mixotrophy has an increased effect on the mitochondrial genome, whereas extremophily mostly explains the plasticity of the nuclear genome.

**Implications of Genome Plasticity on Biotechnology**

The diverse properties of *G. sulphuraria* have demonstrated applications in biotechnology, such as phycocyanin production, metal bioremediation, and processing of agricultural waste. The ability of *G. sulphuraria* to produce phycocyanin, to acquire heavy metals, and to process agricultural waste is contingent on *G. sulphuraria* retaining functioning genes that are involved in these

processes. *G. sulphuraria*'s high mutation rate and genome plasticity could have a negative impact on its potential uses in biotechnology, as it the genome may not be stable enough to produce consistent outcomes. However, the high mutation rate and genome plasticity could also prove advantageous.

We know that *G. sulphuraria* has survived for millions of years in its environment, and it can tolerate its high mutation rate, even though we do not know exactly how it does this. We can therefore reasonably assume that *G. sulphuraria* will not undergo mutational meltdown, especially when in liquid cultures that are used in industry (as opposed to single colonies, that have smaller isolated populations and may die out). Therefore, there is potential to exploit this high genome plasticity for rapid directed evolution for strain development for industry.

Currently, microalgal strain development for biotechnology involves mutagenic treatment, which can involve being subject to UV or ionising radiation, interchelating agents, and a variety of other chemical mutagens including depurinating agents and base analogues, followed by determination of survival rate and screening for the desired mutation [248]. Since *G. sulphuraria* is extremely mutagenic, further subjecting it to chemical and physical mutagens could rapidly lead to the acquisition of novel desired traits. Moreover, at this high rate of mutation simply culturing *G. sulphuraria* under selective growth conditions could result in *G. sulphuraria* optimising itself to industrial conditions without the need for any external mutagenic strategies. Selective growth conditions could involve growing *G. sulphuraria* in the dark on a set of specific carbon sources, and determining if the cultures can optimise themselves and increase carbon uptake on their own.

In addition to these mutagenic strategies, it has now been demonstrated that *G. sulphuraria* gametes can undergo transformation [124], so not only is there potential to form *G. sulphuraria* strain crosses, but there is a possibility of introducing entirely new traits into the genome. These techniques are new, there is currently only one publication reporting them, and they have not been successfully repeated. One major question remains as to how long transformed DNA can persist within such a mutagenic and plastic genome, and if in the absence of continued selection, transformants could simply revert back to their original state over time. As protocols for the induction of meiosis and transformation in *G. sulphuraria* become more

robust, the breeding and development of new *G. sulphuraria* strains with improved traits for specific industrial processes can occur. This represents an exciting and important step in the optimisation of *G. sulphuraria* for use in biotechnology, and it is a very exciting time in the study of eukaryotic extremophiles.

**Third Generation Sequencing**

The advent of 3rd generation sequencing technologies, specifically from Oxford Nanopore Technologies, has enabled the completion of three *G. sulphuraria* genomes. These genome assemblies have revealed a large degree of structural diversity between *G. sulphuraria* genomes, that simply would not have been detected without long read sequencing. At the time the sequencing for this thesis was performed, ONT error rates using the Guppy v. 4.2.2 base calling software (the neural network that translates the raw signals generated by the nanopore into nucleic acid sequence) ranged from 2 – 11% [249], which necessitated the use of Illumina reads to correct the widespread long read sequencing errors.

Oxford Nanopore Technologies are now reporting significantly lower error rates, with accuracy as high as 99.92% for some samples [250], removing the need for assembly polishing with Illumina reads, and possibly Illumina sequencing for evolutionary genomics altogether. This increased sequencing accuracy is comparable with short read technologies and will enable accurate variant calling using long reads. This will be of specific benefit to the *G. sulphuraria* genome which has at least 140 subtelomeric regions, that are difficult to map short reads to, resulting in variants not being accurately detectable in those regions. Additionally, resequencing *G. sulphuraria* with accurate long reads would enable the detection of new structural variants that have possibly been introduced to the genome over the course of mutation accumulation experiments, possibly elucidating the mechanism of structural variation within the genome.

**The *G. sulphuraria* Nuclear Genome is Remarkable, Yet Bewildering**

Thanks to developments in sequencing technologies, coupled with the increased ability to analyse these data, I have uncovered how from a single base, to chromosome wide level, that *G. sulphuraria* has a nuclear genome that exhibits a multitude of extraordinary features. Firstly, the *G. sulphuraria* has a compact genome with numerous tiny chromosomes. This enabled the identification of duplicated

genomic regions, and vast genome wide structural diversity between *G. sulphuraria* isolates. I established that this was likely due to recombination, and that considering that *G. sulphuraria* is a sexual organism, there are now further questions on the implication of genome structural differences between isolates on speciation within the genera. The multitude of small chromosomes, extraordinarily high mutation rates, and recombination represent exceptional genome plasticity that must serve as a mechanism of the extreme eukaryote maintaining an adaptive and flexible lifestyle within a rapidly changing extreme environment. Moreover, the three completed *G. sulphuraria* genome assemblies I have constructed for this thesis, and the diverse range of genomic analysis protocols, provide a lasting legacy for continued analysis of the genome of this truly remarkable organism.

The study of the *G. sulphuraria* genome has created conundrum after conundrum. In *G. sulphuraria*'s quest to conquer a wide range of extreme environments, from thermal springs, to acid mine drainage sites, and more, it has developed a genome whose features challenge the expectations of a free living eukaryote. In summary, it shows us that "Life finds a way".

# Appendix

| Isolate | Isolation Site | Country | Habitat | Culture Collection |
|---|---|---|---|---|
| 138 | San Salvador | El Salvador | N/A | ACUF |
| 002.2 | Piscarelli | Italy | Endolithic site | ACUF |
| 111 | Caserta | Italy | Acidic rock | ACUF |
| 017 | Solfatara | Italy | Fumarols | ACUF |
| 21 | Vulcano Island | Italy | N/A | ACUF |
| 638 | Güglükonak | Turkey | Thermal bath | ACUF |
| 4512T | Güglükonak | Turkey | Thermal bath | ACUF |
| PISC6 | Piscarelli | Italy | Endolithic site | Novel |
| RI1 | Rio Tinto | Spain | Acidic pool | Aguilera |
| SOL1 | Solfatara | Italy | Fumarols | Novel |
| SOL2 | Solfatara | Italy | Fumarols | Novel |
| SOL3 | Solfatara | Italy | Fumarols | Novel |
| 136 | Mexicali | Mexico | N/A | ACUF |
| 141G | Yellowstone National Park | USA | Acidic hot spring | ACUF |
| 142 | N/A | Iceland | N/A | ACUF |
| 1067 | Azores | Portugal | Endolithic site | CCALA |
| 965 | Soos | Czechia | Diatom field | CCALA |
| 5573 | Yellowstone National Park | USA | Acidic soil | CCMEE |
| 5610 | Yellowstone National Park | USA | Acidic crust | CCMEE |
| 5657 | Owakudani | Japan | Acidic hot water | CCMEE |
| 5658 | Owakudani | Japan | Acidic hot water | CCMEE |
| 5665 | Kusatsu | Japan | Acidic hot water | CCMEE |
| 5672 | Owakudani | Japan | Acidic hot water | CCMEE |
| 5712 | Craters of the Moon | New Zealand | Acidic steam hole | CCMEE |
| p501 | N/A | N/A | N/A | IPPAS |
| p503 | Kamchatka | Russia | N/A | IPPAS |
| 107.79 | California | USA | Acidic hot water | SAG |
| 074 | Java | Indonesia | Fumarols | ACUF |

| | | | | |
|---|---|---|---|---|
| THAL033 | GengZiPeng | Taiwan | Acidic hot water | THAL |
| THAL054 | DaYouKeng | Taiwan | Acidic pool | THAL |
| 388 | Landmannalaugar | Iceland | Acidic soil | ACUF |
| 402 | Niasjvellir | Iceland | Acidic soil | ACUF |
| 427 | Gunnhuver | Iceland | Acidic soil | ACUF |
| 455 | Viti | Iceland | Acidic soil | ACUF |
| p501 | Kamchatka | Russia | N/A | IPPAS |
| 009 | Nepi | Italy | Acidic pool | ACUF |
| 647 | Çermik | Turkey | Thermal bath | ACUF |
| 663 | Güglükonak | Turkey | Acidic hot water | ACUF |
| 345TY | Biloris | Turkey | Thermal bath | ACUF |
| 5716 | N/A | New Zealand | N/A | N/A |
| AG1 | Rio Tinto | Spain | Acidic stream | Aguilera |
| CEMI | Rio Tinto | Spain | Acidic stream | Aguilera |
| 141DG | Yellowstone National Park | USA | Acidic hot spring | ACUF |
| 141Y | Yellowstone National Park | USA | Acidic hot spring | ACUF |
| 108.79 | California | USA | Acidic hot water | SAG |
| 107.79 | California | USA | Acidic hot water | SAG |
| 5639 | N/A | N/A | N/A | N/A |
| 5680 | N/A | N/A | N/A | N/A |

*Table 21: List of Galdieria strains, their isolation site, and culture collection. Abbreviations and references of culture collections is as follows: ACUF = Algal Collection University Federico II* [125]*, CCALA = Culture Collection of Autotrophic Organisms* [191]*, CCMEE = Culture Collection of Microorganisms from Extreme Environments* [251]*, IPPAS = Culture Collection of Microalgae, THAL = Tung-Hai Algal Lab Culture Collection* [127]*, SAG = Culture Collection University of Göttingen* [126]*, Aguilera = Aguilera et. al, 2007* [252]*.*

# Bibliography


[1] M. Crichton, *Jurassic Park*. Alfred A. Knopf, 1990.

[2] L. J. Rothschild and R. L. Mancinelli, “Life in extreme environments,” vol. 409, 2001.

[3] C. Brininger, S. Spradlin, L. Cobani, and C. Evilia, “The more adaptive to change, the more likely you are to survive: Protein adaptation in extremophiles,” *Seminars in Cell and Developmental Biology*, vol. 84. Elsevier Ltd, pp. 158–169, 2018, doi: 10.1016/j.semcdb.2017.12.016.

[4] A. Krüger, C. Schäfers, C. Schröder, and G. Antranikian, “Towards a sustainable biobased industry – Highlighting the impact of extremophiles,” *New Biotechnology*, vol. 40. Elsevier B.V., pp. 144–153, 2018, doi: 10.1016/j.nbt.2017.05.002.

[5] A. Chien, D. B. Edgar, and J. M. Trela, “Deoxyribonucleic acid polymerase from the extreme thermophile Thermus aquaticus.,” *J. Bacteriol.*, vol. 127, no. 3, pp. 1550–1557, Sep. 1976, doi: 10.1128/jb.127.3.1550-1557.1976.

[6] S. Ishino and Y. Ishino, “DNA polymerases as useful reagents for biotechnology - the history of developmental research in the field.,” *Front. Microbiol.*, vol. 5, p. 465, 2014, doi: 10.3389/fmicb.2014.00465.

[7] A. Bhalla, N. Bansal, S. Kumar, K. M. Bischoff, and R. K. Sani, “Improved lignocellulose conversion to biofuels with thermophilic bacteria and thermostable enzymes.,” *Bioresour. Technol.*, vol. 128, pp. 751–759, Jan. 2013, doi: 10.1016/j.biortech.2012.10.145.

[8] R. Cavicchioli, K. S. Siddiqui, D. Andrews, and K. R. Sowers, “Low-temperature extremophiles and their applications.,” *Curr. Opin. Biotechnol.*, vol. 13, no. 3, pp. 253–261, Jun. 2002, doi: 10.1016/s0958-1669(02)00317-8.

[9] B. Van den Burg, “Extremophiles as a source for novel enzymes,” *Current Opinion in Microbiology*, vol. 6, no. 3. Elsevier Ltd, pp. 213–218, 2003, doi: 10.1016/S1369-5274(03)00060-2.

[10] E. A. Gaucher, J. T. Kratzer, and R. N. Randall, “Deep Phylogeny — How a Tree Can Help Characterize Early Life on Earth,” *Cold Spring Harb. Perspect.*

*Biol.*, pp. 1–17, 2010.

[11] J. Van Etten, C. H. Cho, H. S. Yoon, and D. Bhattacharya, “Extremophilic red algae as models for understanding adaptation to hostile environments and the evolution of eukaryotic life on the early earth,” *Semin. Cell Dev. Biol.*, 2022, doi: 10.1016/j.semcdb.2022.03.007.

[12] E. V Pikuta *et al.*, “Microbial Extremophiles at the Limits of Life,” *Crit. Rev. Microbiol.*, vol. 33, no. 3, pp. 183–209, 2007, doi: 10.1080/10408410701451948.

[13] J. Seckbach and H. Stan-Lotter, *Extremophiles as Astrobiological Models*. Scrivener Publishing LLC, 2020.

[14] V. Mastascusa *et al.*, “Extremophiles survival to simulated space conditions: an astrobiology model study.,” *Orig. life Evol. Biosph. J. Int. Soc. Study Orig. Life*, vol. 44, no. 3, pp. 231–237, Sep. 2014, doi: 10.1007/s11084-014-9397-y.

[15] D. Burr *et al.*, “Exocube: an Astrobiology Exposure Platform Onboard the International Space Station,” in *44th COSPAR Scientific Assembly. Held 16-24 July*, Jul. 2022, vol. 44, p. 2759.

[16] H. Philippe and C. J. Douady, “Horizontal gene transfer and phylogenetics.,” *Curr. Opin. Microbiol.*, vol. 6, no. 5, pp. 498–505, Oct. 2003, doi: 10.1016/j.mib.2003.09.008.

[17] Q. Wang, Z. Cen, and J. Zhao, “The Survival Mechanisms of Thermophiles at High Temperatures: An Angle of Omics,” *Physiology*, vol. 30, no. 2, pp. 97–106, Mar. 2015, doi: 10.1152/physiol.00066.2013.

[18] M. Heine and S. B. C. Chandra, “The linkage between reverse gyrase and hyperthermophiles: a review of their invariable association.,” *J. Microbiol.*, vol. 47, no. 3, pp. 229–234, Jun. 2009, doi: 10.1007/s12275-009-0019-8.

[19] S. J. Li *et al.*, “Microbial communities evolve faster in extreme environments,” *Sci. Rep.*, vol. 4, 2014, doi: 10.1038/srep06205.

[20] M. van Wolferen, M. Ajon, A. J. M. Driessen, and S.-V. Albers, “How hyperthermophiles adapt to change their lives: DNA exchange in extreme conditions.,” *Extremophiles*, vol. 17, no. 4, pp. 545–563, Jul. 2013, doi:

10.1007/s00792-013-0552-6.

[21] J. Seckbach, “Algae and Cyanobacteria in Extreme Environments,” 2007.

[22] H. S. Yoon, K. M. Müller, R. G. Sheath, F. D. Ott, and D. Bhattacharya, “Defining the Major Lineages of the Red Algae (Rhodophyta),” *J. Phycol.*, vol. 42, no. 2, pp. 482–492, Apr. 2006, doi: https://doi.org/10.1111/j.1529-8817.2006.00210.x.

[23] S. M. Adl *et al.*, “The New Higher Level Classification of Eukaryotes with Emphasis on the Taxonomy of Protists,” *J. Eukaryot. Microbiol.*, vol. 52, no. 5, pp. 399–451, Oct. 2005, doi: https://doi.org/10.1111/j.1550-7408.2005.00053.x.

[24] J. E. Tilden, “Observations on Some West American Thermal Algæ,” *Bot. Gaz.*, vol. 25, no. 2, pp. 89–105, 1898, [Online]. Available: http://www.jstor.org/stable/2464465.

[25] W. A. Setchell, “Algae of North America,” *Phycotheca Boreali-Americana*, vol. Fasc. 18, no. No. 851, 1901.

[26] L. Geitler, “Die Cyanophyceen der Deutschen Limnologischen Sunda-Expedition,” *Hydrobiol. fur Arch.*, 1936.

[27] H. Hirose, “Studies on a Thermal Alga, Cyanidium caldarium,” *Bot. Mag. Tokyo*, vol. 63, no. 745–746, pp. 107–111, 1950.

[28] M. B. Allen, “Studies with Cyanidium caldarium, an anomalously pigmented chlorophyte,” *Arch. Mikrobiol.*, vol. 32, pp. 270–277, 1959.

[29] A. Merola, R. Castaldo, P. De Luca, R. Gambardella, A. Musacchio, and R. Taddei, “Revision of cyanidium caldarium. Three species of acidophilic algae,” *G. Bot. Ital.*, vol. 115, no. 4–5, pp. 189–195, 1981, doi: 10.1080/11263508109428026.

[30] J. Seckbach, “Systematic problems with Cyanidium caldarium and Galdieria sulphuraria and their implications for molecular biology studies.,” *J. Phycol.*, vol. 27, pp. 794–796, 1991.

[31] J. Seckbach, R. Ikan, H. Nagashima, and I. Fukuda, “New Phylogenetic Status

for acid hot spring algae," *Endocytobiology*, pp. 241–254, 1993.

[32] O. Y. Sentsova, "The study of Cyanidiophyceae in Russia," in *Evolutionary Pathways and Enigmatic Algae: Cyanidium caldarium (Rhodophyta) and Related Cells*, J. Seckbach, Ed. Dordrecht: Springer Netherlands, 1994, pp. 167–174.

[33] S. Cozzolino, P. Caputo, O. De Castro, A. Moretti, and G. Pinto, "Molecular variation in Galdieria sulphuraria (Galdieri) merola and its bearing on taxonomy," *Hydrobiologia*, vol. 433, no. 1994, pp. 145–151, 2000, doi: 10.1023/A:1004035224715.

[34] S.-L. Liu, Y.-R. Chiang, H. S. Yoon, and H.-Y. Fu, "Comparative Genome Analysis Reveals Cyanidiococcus Gen. Nov., A New Extremophilic Red Algal Genus Sister To Cyanidioschyzon (Cyanidioschyzonaceae, Rhodophyta)," *J. Phycol.*, vol. 56, pp. 1428–1442, 2020.

[35] H. Qiu *et al.*, "Adaptation through horizontal gene transfer in the cryptoendolithic red alga Galdieria phlegrea," *Current Biology*. 2013, doi: 10.1016/j.cub.2013.08.046.

[36] C. Ciniglia *et al.*, "Cyanidium chilense (Cyanidiophyceae, Rhodophyta) from tuff rocks of the archeological site of Cuma, Italy," *Phycol. Res.*, vol. 67, no. 4, pp. 311–319, 2019, doi: 10.1111/pre.12383.

[37] P. A. Cohen and F. A. Macdonald, "The Proterozoic Record of Eukaryotes," *Paleobiology*, vol. 41, no. 4, pp. 610–632, 2015, doi: 10.1017/pab.2015.25.

[38] W. W. Fischer, J. Hemp, and J. S. Valentine, "How did life survive Earth's great oxygenation?," *Curr. Opin. Chem. Biol.*, vol. 31, pp. 166–178, 2016, doi: https://doi.org/10.1016/j.cbpa.2016.03.013.

[39] C. Ciniglia, H. S. Yoon, A. Pollio, G. Pinto, and D. Bhattacharya, "Hidden biodiversity of the extremophilic Cyanidiales red algae," *Mol. Ecol.*, vol. 13, no. 7, pp. 1827–1838, 2004, doi: 10.1111/j.1365-294X.2004.02180.x.

[40] H. S. Yoon *et al.*, "Establishment of endolithic populations of extremophilic Cyanidiales (Rhodophyta)," *BMC Evol. Biol.*, vol. 6, no. 1, p. 78, 2006, doi: 10.1186/1471-2148-6-78.

[41] C. H. Cho *et al.*, “Potential causes and consequences of rapid mitochondrial genome evolution in thermoacidophilic Galdieria (Rhodophyta),” *BMC Evol. Biol.*, vol. 20, no. 1, 2020, doi: 10.1186/s12862-020-01677-6.

[42] B. J. Finlay, “Global dispersal of free-living microbial eukaryote species.,” *Science*, vol. 296, no. 5570, pp. 1061–1063, May 2002, doi: 10.1126/science.1070710.

[43] A. Eren *et al.*, “Genetic structure of Galdieria populations from Iceland,” *Polar Biol.*, vol. 41, no. 9, pp. 1681–1691, 2018, doi: 10.1007/s00300-018-2308-3.

[44] K. Jain *et al.*, “Extreme features of the Galdieria sulphuraria organellar genomes: a consequence of polyextremophily?,” *Genome Biol. Evol.*, vol. 7, no. 1, pp. 367–380, Dec. 2014, doi: 10.1093/gbe/evu290.

[45] M. Vítová, Ed., “Microalgae.” IntechOpen, Rijeka, 2020, doi: 10.5772/intechopen.83737.

[46] D. Barcytė, J. Elster, and L. Nedbalová, “Plastid-encoded rbcL phylogeny suggests widespread distribution of Galdieria phlegrea (Cyanidiophyceae, Rhodophyta),” *Nord. J. Bot.*, vol. 36, no. 7, p. e01794, Jul. 2018, doi: https://doi.org/10.1111/njb.01794.

[47] D. Moreira, A. I. López-Archilla, R. Amils, and I. Marín, “Characterization of two new thermoacidophilic microalgae: Genome organization and comparison with Galdieria sulphuraria,” *FEMS Microbiol. Lett.*, vol. 122, no. 1–2, pp. 109–114, 1994.

[48] D. Barcyte, L. Nedbalova, A. Culka, F. Kosek, and J. Jehlicka, “Burning coal spoil heaps as a new habitat for the extremophilic red alga Galdieria sulphuraria,” *Fottea*, vol. 18, no. 1, pp. 19–29, 2018, doi: 10.5507/fot.2017.015.

[49] W. N. Doemel and T. D. Brock, “The Physiological Ecology of Cyanidium caldarium,” *Microbiology*, vol. 67, no. 1, pp. 17–32, 1971, doi: https://doi.org/10.1099/00221287-67-1-17.

[50] D. Pleissner, A. V. Lindner, and N. Händel, “Heterotrophic cultivation of Galdieria sulphuraria under non-sterile conditions in digestate and hydrolyzed straw,” *Bioresour. Technol.*, vol. 337, p. 125477, 2021, doi:

https://doi.org/10.1016/j.biortech.2021.125477.

[51] S. Sekar and M. Chandramohan, “Phycobiliproteins as a commodity: trends in applied research, patents and commercialization,” *J. Appl. Phycol.*, vol. 20, no. 2, pp. 113–136, 2008, doi: 10.1007/s10811-007-9188-1.

[52] N. T. Eriksen, “Production of phycocyanin--a pigment with applications in biology, biotechnology, foods and medicine.,” *Appl. Microbiol. Biotechnol.*, vol. 80, no. 1, pp. 1–14, Aug. 2008, doi: 10.1007/s00253-008-1542-y.

[53] Y.-K. Lee, “Commercial production of microalgae in the Asia-Pacific rim.” 1997.

[54] L. Sørensen, A. Hantke, and N. T. Eriksen, “Purification of the photosynthetic pigment C-phycocyanin from heterotrophic Galdieria sulphuraria,” *J. Sci. Food Agric.*, vol. 93, no. 12, pp. 2933–2938, Sep. 2013, doi: https://doi.org/10.1002/jsfa.6116.

[55] R. A. Schmidt, M. G. Wiebe, and N. T. Eriksen, “Heterotrophic high cell-density fed-batch cultures of the phycocyanin-producing red alga Galdieria sulphuraria,” *Biotechnol. Bioeng.*, vol. 90, no. 1, pp. 77–84, Apr. 2005, doi: https://doi.org/10.1002/bit.20417.

[56] M. Wan *et al.*, “Comparison of C-phycocyanin from extremophilic Galdieria sulphuraria and Spirulina platensis on stability and antioxidant capacity,” *Algal Res.*, vol. 58, p. 102391, 2021, doi: https://doi.org/10.1016/j.algal.2021.102391.

[57] M. Palmieri *et al.*, “Galdieria sulphuraria ACUF 427 Freeze-Dried Biomass as Novel Biosorbent for Rare Earth Elements,” *Microorganisms*, vol. 10, no. 11, 2022, doi: 10.3390/microorganisms10112138.

[58] B. Volesky and Z. R. Holan, “Biosorption of Heavy Metals,” *Biotechnol. Prog.*, vol. 11, no. 3, pp. 235–250, May 1995, doi: https://doi.org/10.1021/bp00033a001.

[59] M. Iovinella *et al.*, “Bioremoval of Yttrium (III), Cerium (III), Europium (III), and Terbium (III) from Single and Quaternary Aqueous Solutions Using the Extremophile Galdieria sulphuraria (Galdieriaceae, Rhodophyta),” *Plants*, vol. 11, no. 10. 2022, doi: 10.3390/plants11101376.

[60] X. Ju *et al.*, “Effective and selective recovery of gold and palladium ions from metal wastewater using a sulfothermophilic red alga, Galdieria sulphuraria,” *Bioresour. Technol.*, vol. 211, pp. 759–764, 2016, doi: https://doi.org/10.1016/j.biortech.2016.01.061.

[61] S. A. Ostroumov, I. V Tropin, and A. V Kiryushin, “Removal of Cadmium and Other Toxic Metals from Water: Thermophiles and New Biotechnologies,” *Russ. J. Gen. Chem.*, vol. 88, no. 13, pp. 2962–2966, 2018, doi: 10.1134/S1070363218130224.

[62] S. Pan, K. L. Dixon, T. Nawaz, A. Rahman, and T. Selvaratnam, “Evaluation of Galdieria sulphuraria for nitrogen removal and biomass production from raw landfill leachate,” *Algal Res.*, vol. 54, p. 102183, 2021, doi: https://doi.org/10.1016/j.algal.2021.102183.

[63] C. Ciniglia *et al.*, “Cyanidiophyceae in Iceland: Plastid rbcL gene elucidates origin and dispersal of extremophilic Galdieria sulphuraria and G. maxima (Galdieriaceae, Rhodophyta),” *Phycologia*, vol. 53, no. 6, pp. 542–551, 2014, doi: 10.2216/14-032.1.

[64] J. D. Watson and F. H. C. Crick, “The structure of DNA,” in *Cold Spring Harbor symposia on quantitative biology*, 1953, vol. 18, pp. 123–131.

[65] B. Maddox, *Rosalind Franklin: the dark lady of DNA*. First edition. New York : HarperCollins, 2002., 2002.

[66] R. W. HOLLEY *et al.*, “Structure of a Ribonucleic Acid.,” *Science*, vol. 147, no. 3664, pp. 1462–1465, Mar. 1965, doi: 10.1126/science.147.3664.1462.

[67] F. Sanger, G. G. Brownlee, and B. G. Barrell, “A two-dimensional fractionation procedure for radioactive nucleotides.,” *J. Mol. Biol.*, vol. 13, no. 2, pp. 373–398, Sep. 1965, doi: 10.1016/s0022-2836(65)80104-8.

[68] W. Min Jou, G. Haegeman, M. Ysebaert, and W. Fiers, “Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein.,” *Nature*, vol. 237, no. 5350, pp. 82–88, May 1972, doi: 10.1038/237082a0.

[69] W. Fiers *et al.*, “Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene.,” *Nature*, vol. 260, no.

5551, pp. 500–507, Apr. 1976, doi: 10.1038/260500a0.

[70] R. Wu, “Nucleotide sequence analysis of DNA: I. Partial sequence of the cohesive ends of bacteriophage λ and 186 DNA,” *J. Mol. Biol.*, vol. 51, no. 3, pp. 501–521, 1970, doi: https://doi.org/10.1016/0022-2836(70)90004-5.

[71] R. Wu and A. D. Kaiser, “Structure and base sequence in the cohesive ends of bacteriophage lambda DNA.,” *J. Mol. Biol.*, vol. 35, no. 3, pp. 523–537, Aug. 1968, doi: 10.1016/s0022-2836(68)80012-9.

[72] R. Padmanabhan, R. Padmanabhan, and R. Wu, “Nucleotide sequence analysis of DNA: IX. Use of oligonucleotides of defined sequence as primers in DNA sequence analysis,” *Biochem. Biophys. Res. Commun.*, vol. 48, no. 5, pp. 1295–1302, 1972, doi: https://doi.org/10.1016/0006-291X(72)90852-2.

[73] J. M. Heather and B. Chain, “The sequence of sequencers: The history of sequencing DNA,” *Genomics*. 2016, doi: 10.1016/j.ygeno.2015.11.003.

[74] F. Sanger, S. Nicklen, and A. R. Coulson, “DNA sequencing with chain-terminating inhibitors.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 74, no. 12, pp. 5463–5467, Dec. 1977, doi: 10.1073/pnas.74.12.5463.

[75] S. Levy *et al.*, “The Diploid Genome Sequence of an Individual Human,” *PLOS Biol.*, vol. 5, no. 10, p. e254, Sep. 2007, [Online]. Available: https://doi.org/10.1371/journal.pbio.0050254.

[76] NHGRI, “The Cost of Sequencing a Human Genome.” https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost, NHGRI.

[77] E. C. Hayden, “Technology: The $1,000 genome.,” *Nature*, vol. 507, no. 7492. England, pp. 294–295, Mar. 2014, doi: 10.1038/507294a.

[78] M. Ronaghi, S. Karamohamed, B. Pettersson, M. Uhlén, and P. Nyrén, “Real-Time DNA Sequencing Using Detection of Pyrophosphate Release,” *Anal. Biochem.*, vol. 242, no. 1, pp. 84–89, 1996, doi: https://doi.org/10.1006/abio.1996.0432.

[79] M. Ronaghi, M. Uhlén, and P. Nyrén, “A sequencing method based on real-time pyrophosphate.,” *Science*, vol. 281, no. 5375, p. 363,365, Jul. 1998, doi:

10.1126/science.281.5375.363.

[80] T. Hu, N. Chitnis, D. Monos, and A. Dinh, “Next-generation sequencing technologies: An overview,” *Hum. Immunol.*, vol. 82, no. 11, pp. 801–811, 2021, doi: 10.1016/j.humimm.2021.02.012.

[81] A. Rhoads and K. F. Au, “PacBio Sequencing and Its Applications.,” *Genomics. Proteomics Bioinformatics*, vol. 13, no. 5, pp. 278–289, Oct. 2015, doi: 10.1016/j.gpb.2015.08.002.

[82] J. Clarke, H.-C. Wu, L. Jayasinghe, A. Patel, S. Reid, and H. Bayley, “Continuous base identification for single-molecule nanopore DNA sequencing.,” *Nat. Nanotechnol.*, vol. 4, no. 4, pp. 265–270, Apr. 2009, doi: 10.1038/nnano.2009.12.

[83] Illumina, “High Performance Long Read Assay Enables Contiguous Data up to 10Kb on Existing Illumina Platforms.” https://emea.illumina.com/science/genomics-research/articles/infinity-high-performance-long-read-assay.html (accessed Aug. 02, 2020).

[84] M. O. Dayhoff and R. S. Ledley, “Comprotein: a computer program to aid primary protein structure determination,” 1962.

[85] L. Pauling, E. Zuckerkandl, T. Henriksen, and R. Lövstad, “Chemical Paleogenetics. Molecular ‘Restoration Studies’ of Extinct Forms of Life.,” *Acta Chem. Scand.*, vol. 17 supl., pp. 9–16, 1963, doi: 10.3891/acta.chem.scand.17s-0009.

[86] W. M. Fitch, “Distinguishing Homologous from Analogous Proteins,” *Syst. Biol.*, vol. 19, no. 2, pp. 99–113, 1970, doi: 10.2307/2412448.

[87] J. E. Haber and D. E. J. Koshland, “An evaluation of the relatedness of proteins based on comparison of amino acid sequences.,” *J. Mol. Biol.*, vol. 50, no. 3, pp. 617–639, Jun. 1970, doi: 10.1016/0022-2836(70)90089-6.

[88] D. G. Higgins and P. M. Sharp, “CLUSTAL: a package for performing multiple sequence alignment on a microcomputer.,” *Gene*, vol. 73, no. 1, pp. 237–244, Dec. 1988, doi: 10.1016/0378-1119(88)90330-7.

[89] F. Sievers and D. G. Higgins, “Clustal Omega, accurate alignment of very

large numbers of sequences.," *Methods Mol. Biol.*, vol. 1079, pp. 105–116, 2014, doi: 10.1007/978-1-62703-646-7_6.

[90] R. Stallman, "The GNU Manifesto," 1985.

[91] Cern, "The Birth of the Web." https://www.home.cern/science/computing/birth-web.

[92] J. Gauthier, A. T. Vincent, S. J. Charette, and N. Derome, "A brief history of bioinformatics.," *Brief. Bioinform.*, vol. 20, no. 6, pp. 1981–1996, Nov. 2019, doi: 10.1093/bib/bby063.

[93] S. Mangul *et al.*, "Systematic benchmarking of omics computational tools," *Nat. Commun.*, vol. 10, no. 1, p. 1393, 2019, doi: 10.1038/s41467-019-09406-4.

[94] C. Turnbull *et al.*, "The 100,000 Genomes Project: bringing whole genome sequencing to the NHS," *BMJ*, vol. 361, 2018, doi: 10.1136/bmj.k1687.

[95] J. Wise, "Genome sequencing of children promises a new era in oncology," *BMJ Br. Med. J.*, vol. 364, Jan. 2019, doi: https://doi.org/10.1136/bmj.l105.

[96] C. N. Ibe, "Democratizing plant genomics to accelerate global food production," *Nat. Genet.*, vol. 54, no. 7, pp. 911–913, 2022, doi: 10.1038/s41588-022-01122-y.

[97] T. E. B. Project, "Earth Biogenome Project." https://www.earthbiogenome.org/.

[98] S. W. Lo and D. Jamrozy, "Genomics and epidemiological surveillance," *Nat. Rev. Microbiol.*, vol. 18, no. 9, p. 478, 2020, doi: 10.1038/s41579-020-0421-0.

[99] K. A. Saravanan *et al.*, "Role of genomics in combating COVID-19 pandemic," *Gene*, vol. 823, p. 146387, 2022, doi: https://doi.org/10.1016/j.gene.2022.146387.

[100] M. Matsuzaki *et al.*, "Genome sequence of the ultrasmall unicellular red alga Cyanidioschyzon merolae 10D." 2004, [Online]. Available: http://www.ncbi.nlm.nih.gov/COG/new/kognitor.html.

[101] A. P. M. Weber, G. G. Barbier, R. P. Shrestha, R. J. Horst, A. Minoda, and C. Oesterhelt, "A Genomics Approach to Understanding the Biology of Thermo-Acidophilic Red Algae," in *Algae and Cyanobacteria in Extreme Environments*,

J. Seckbach, Ed. Dordrecht: Springer Netherlands, 2007, pp. 503–518.

[102] H. Qiu, D. C. Price, E. C. Yang, H. S. Yoon, and D. Bhattacharya, “Evidence of Ancient Genome Reduction in Red Algae (Rhodophyta),” *J. Phycol.*, vol. 51, no. 4, pp. 624–636, 2015, doi: 10.1111/jpy.12294.

[103] D. Bhattacharya, H. Qiu, J. Lee, H. Su Yoon, A. P. M. Weber, and D. C. Price, “When Less is More: Red Algae as Models for Studying Gene Loss and Genome Evolution in Eukaryotes,” *CRC. Crit. Rev. Plant Sci.*, vol. 37, no. 1, pp. 81–99, Jan. 2018, doi: 10.1080/07352689.2018.1482364.

[104] J. Moran, P. G. McKean, and M. L. Ginger, “Eukaryote Flagella: Variations in Form, Function, and Composition during Evolution,” *Bioscience*, vol. 64, no. 12, pp. 1103–1114, Nov. 2014, [Online]. Available: http://www.jstor.org/stable/90006997.

[105] A. R. Burmeister, “Horizontal Gene Transfer.,” *Evolution, medicine, and public health*, vol. 2015, no. 1. England, pp. 193–194, Jul. 2015, doi: 10.1093/emph/eov018.

[106] H. Tettelin *et al.*, “Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial ‘pan-genome’.,” *Proc. Natl. Acad. Sci.*, vol. 102, no. 39, pp. 13950–13955, Sep. 2005, doi: 10.1073/pnas.0506758102.

[107] G. Vernikos, D. Medini, D. R. Riley, and H. Tettelin, “Ten years of pan-genome analyses.,” *Curr. Opin. Microbiol.*, vol. 23, pp. 148–154, Feb. 2015, doi: 10.1016/j.mib.2014.11.016.

[108] W. F. Doolittle and T. D. P. Brunet, “What Is the Tree of Life?,” *PLOS Genet.*, vol. 12, no. 4, p. e1005912, Apr. 2016, [Online]. Available: https://doi.org/10.1371/journal.pgen.1005912.

[109] G. Schönknecht, A. P. M. Weber, and M. J. Lercher, “Horizontal gene acquisitions by eukaryotes as drivers of adaptive evolution,” *BioEssays*, vol. 36, no. 1, pp. 9–20, 2014, doi: https://doi.org/10.1002/bies.201300095.

[110] G. Schönknecht *et al.*, “Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote,” *Science (80-. ).*, vol. 339, no. 6124,

pp. 1207–1210, 2013, doi: 10.1126/science.1231707.

[111] W. F. Martin, “Too Much Eukaryote LGT,” *BioEssays*, vol. 39, no. 12. John Wiley and Sons Inc., 2017, doi: 10.1002/bies.201700115.

[112] S. L. Salzberg, O. White, J. Peterson, and J. A. Eisen, “Microbial genes in the human genome: lateral transfer or gene loss?,” *Science*, vol. 292, no. 5523, pp. 1903–1906, Jun. 2001, doi: 10.1126/science.1061036.

[113] M. J. Stanhope, A. Lupas, M. J. Italia, K. K. Koretke, C. Volker, and J. R. Brown, “Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates.,” *Nature*, vol. 411, no. 6840, pp. 940–944, Jun. 2001, doi: 10.1038/35082058.

[114] C. Ku and W. F. Martin, “A natural barrier to lateral gene transfer from prokaryotes to eukaryotes revealed from genomes: the 70 % rule,” *BMC Biol.*, pp. 1–12, 2016, doi: 10.1186/s12915-016-0315-9.

[115] A. W. Rossoni *et al.*, “The genomes of polyextremophilic cyanidiales contain 1% horizontally transferred genes with diverse adaptive functions,” *Elife*, vol. 8, 2019, doi: 10.7554/eLife.45017.

[116] J. Van Etten and D. Bhattacharya, “Horizontal Gene Transfer in Eukaryotes: Not if, but How Much?,” *Trends in Genetics*, vol. 36, no. 12. Elsevier Ltd, pp. 915–925, 2020, doi: 10.1016/j.tig.2020.08.006.

[117] G. Glöckner, A. Rosenthal, and K. Valentin, “The Structure and Gene Repertoire of an Ancient Red Algal Plastid Genome,” *J. Mol. Evol.*, vol. 51, no. 4, pp. 382–390, 2000, doi: 10.1007/s002390010101.

[118] C. J. Hsieh, S. H. Zhan, Y. Lin, S. L. Tang, and S. L. Liu, “Analysis of rbcL sequences reveals the global biodiversity, community structure, and biogeographical pattern of thermoacidophilic red algae (Cyanidiales).,” *J. Phycol.*, vol. 51, no. 4, pp. 682–694, Aug. 2015, doi: 10.1111/jpy.12310.

[119] M. Iovinella *et al.*, “Cryptic dispersal of Cyanidiophytina (Rhodophyta) in non-acidic environments from Turkey.,” *Extremophiles*, vol. 22, no. 5, pp. 713–723, Sep. 2018, doi: 10.1007/s00792-018-1031-x.

[120] M. Iovinella *et al.*, “Resolving Complexities in Taxonomic Lineages of the

Organellar and Nuclear Genomes of Galdieria through Comparative Phylogenomic Analysis," *bioRxiv*, 2022, doi: 10.1101/2022.10.04.510841.

[121] D. A. Benson *et al.*, "GenBank.," *Nucleic Acids Res.*, vol. 41, no. Database issue, pp. D36-42, Jan. 2013, doi: 10.1093/nar/gks1195.

[122] National Center for Biotechnology Information, "GenBank." https://www.ncbi.nlm.nih.gov/genbank/.

[123] A. Davis *et al.*, "Using MinION nanopore sequencing to generate a de novo eukaryotic draft genome: preliminary physiological and genomic description of the extremophilic red alga Galdieria sulphuraria strain SAG 107.79," *bioRxiv*, p. 76208, 2016, doi: 10.1101/076208.

[124] S. Hirooka, T. Itabashi, and T. M. Ichinose, "Life cycle and functional genomics of the unicellular red alga Galdieria for elucidating algal and plant evolution and industrial use," pp. 1–12, 2022, doi: 10.1073/pnas.2210665119/-/DCSupplemental.Published.

[125] G. Pinto, A. Pollio, C. Ciniglia, A. De Natale, and A. Del Mondo, "Algal Collection of University Federico II of Naples (ACUF)," 2014. www.acuf.net.

[126] M. Lorenz, "The Culture Collection of Algae (SAG) at Goettingen University." https://www.uni-goettingen.de/en/culture+collection+of+algae+%28sag%29/184982.html.

[127] T. Algal Molecular Ecology Lab, Department of Life Science & Center for Ecology and Environment, Tunghai University, "Tung-Hai Algal Lab (THAL) Culture Collection." http://algae.thu.edu.tw/lab/?page_id=42.

[128] A. M. Bolger, M. Lohse, and B. Usadel, "Genome analysis Trimmomatic: a flexible trimmer for Illumina sequence data," vol. 30, no. 15, pp. 2114–2120, 2014, doi: 10.1093/bioinformatics/btu170.

[129] A. Bankevich *et al.*, "SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing," vol. 19, no. 5, pp. 455–477, 2012, doi: 10.1089/cmb.2012.0021.

[130] R. R. Wick, L. M. Judd, and K. E. Holt, "Performance of neural network basecalling tools for Oxford Nanopore sequencing," pp. 1–10, 2019.

[131] S. Koren, B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman, and A. M. Phillippy, “Canu: scalable and accurate long-read assembly via adaptive k -mer weighting and repeat separation,” pp. 722–736, 2017, doi: 10.1101/gr.215087.116.Freely.

[132] R. Vaser and M. Šikić, “Time- and memory-efficient genome assembly with Raven,” *Nat. Comput. Sci.*, vol. 1, no. 5, pp. 332–336, 2021, doi: 10.1038/s43588-021-00073-4.

[133] H. Liu, S. Wu, A. Li, and J. Ruan, “SMARTdenovo: a de novo assembler using long noisy reads,” *Gigabyte*, 2021, doi: 10.46471/gigabyte.15.

[134] “medaka: Sequence correction provided by ONT Research.” https://github.com/nanoporetech/medaka.

[135] B. J. Walker *et al.*, “Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement,” vol. 9, no. 11, 2014, doi: 10.1371/journal.pone.0112963.

[136] H. Li and R. Durbin, “Fast and accurate short read alignment with Burrows–Wheeler transform,” *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009, doi: 10.1093/bioinformatics/btp324.

[137] J. W. Davey, S. J. Davis, J. C. Mottram, and P. D. Ashton, “Tapestry: validate and edit small eukaryotic genome assemblies with long reads,” *Biorxiv*, 2020.

[138] H. Li, “Minimap2: pairwise alignment for nucleotide sequences,” *Bioinformatics*, vol. 34, no. 18, pp. 3094–3100, 2018, doi: 10.1093/bioinformatics/bty191.

[139] M. Grabherr, B. Haas, and M. Yassour, “Full-length transcriptome assembly from RNA-Seq data without a reference genome,” *Nat. Biotechnol.*, vol. 29, no. 7, pp. 644–652, 2011, doi: 10.1038/nbt.1883.

[140] A. Dobin *et al.*, “STAR: ultrafast universal RNA-seq aligner,” *Bioinformatics*, vol. 29, no. 1, pp. 15–21, 2013, doi: 10.1093/bioinformatics/bts635.

[141] “Funannotate.” https://github.com/nextgenusfs/funannotate.

[142] F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V Kriventseva, and E. M.

Zdobnov, “BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs,” *Bioinformatics*, vol. 31, no. 19, pp. 3210–3212, Oct. 2015, doi: 10.1093/bioinformatics/btv351.

[143] P. Jones *et al.*, “Sequence analysis InterProScan 5 : genome-scale protein function classification,” vol. 30, no. 9, pp. 1236–1240, 2014, doi: 10.1093/bioinformatics/btu031.

[144] C. Camacho *et al.*, “BLAST+: architecture and applications.,” *BMC Bioinformatics*, vol. 10, p. 421, Dec. 2009, doi: 10.1186/1471-2105-10-421.

[145] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, “Basic local alignment search tool.,” *J. Mol. Biol.*, vol. 215, no. 3, pp. 403–410, Oct. 1990, doi: 10.1016/S0022-2836(05)80360-2.

[146] H. Feinberg, A. Changela, and A. Mondragón, “Protein-nucleotide interactions in E. coli DNA topoisomerase I.,” *Nat. Struct. Biol.*, vol. 6, no. 10, pp. 961–968, Oct. 1999, doi: 10.1038/13333.

[147] R. C. Edgar, “MUSCLE: multiple sequence alignment with high accuracy and high throughput.,” *Nucleic Acids Res.*, vol. 32, no. 5, pp. 1792–1797, 2004, doi: 10.1093/nar/gkh340.

[148] Y. Wang *et al.*, “MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity.,” *Nucleic Acids Res.*, vol. 40, no. 7, p. e49, Apr. 2012, doi: 10.1093/nar/gkr1293.

[149] University of York, “Viking High Performance Computing Cluster.” https://wiki.york.ac.uk/display/RCS/Viking+-+University+of+York+Research+Computing+Cluster.

[150] RStudio Team, “RStudio: Integrated Development Environment for R.” Boston, MA, 2020, [Online]. Available: http://www.rstudio.com/.

[151] B. Gel and E. Serra, “karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data,” *Bioinformatics*, vol. 33, no. 19, pp. 3088–3090, Oct. 2017, doi: 10.1093/bioinformatics/btx346.

[152] A. Smit, R. Hubley, and P. Green, “RepeatMasker Open-4.0.” www.repeatmasker.org.

[153] A. Maréchal and N. Brisson, “Recombination and the maintenance of plant organelle genome stability,” *New Phytol.*, vol. 186, no. 2, pp. 299–317, Apr. 2010, doi: https://doi.org/10.1111/j.1469-8137.2010.03195.x.

[154] A. J. Bendich, “Circular Chloroplast Chromosomes: The Grand Illusion,” *Plant Cell*, vol. 16, no. 7, pp. 1661–1666, Jul. 2004, doi: 10.1105/tpc.160771.

[155] J. W. Lilly, M. J. Havey, S. A. Jackson, and J. Jiang, “Cytogenomic Analyses Reveal the Structural Plasticity of the Chloroplast Genome in Higher Plants,” *Plant Cell*, vol. 13, no. 2, pp. 245–254, Feb. 2001, doi: 10.1105/tpc.13.2.245.

[156] G. Bourque *et al.*, “Ten things you should know about transposable elements,” *Genome Biol.*, vol. 19, no. 199, pp. 1–12, 2018, doi: https://doi.org/10.1186/s13059-018-1577-z.

[157] L. Chen *et al.*, “The genome of Sulfolobus acidocaldarius, a model organism of the Crenarchaeota.,” *J. Bacteriol.*, vol. 187, no. 14, pp. 4992–4999, Jul. 2005, doi: 10.1128/JB.187.14.4992-4999.2005.

[158] P. López-García and P. Forterre, “Control of DNA topology during thermal stress in hyperthermophilic archaea: DNA topoisomerase levels, activities and induced thermotolerance during heat and cold shock in Sulfolobus,” *Mol. Microbiol.*, vol. 33, no. 4, pp. 766–777, Aug. 1999, doi: https://doi.org/10.1046/j.1365-2958.1999.01524.x.

[159] European Bioinformatics Institute, “DNA Topoisomerase I, bacterial-type.” https://www.ebi.ac.uk/interpro/entry/InterPro/IPR005733/.

[160] D. Speijer, J. Lukeš, and M. Eliáš, “Sex is a ubiquitous, ancient, and inherent attribute of eukaryotic life,” *Proc. Natl. Acad. Sci.*, vol. 112, no. 29, pp. 8827–8834, 2015, doi: 10.1073/PNAS.1501725112.

[161] A. M. Schurko and J. M. Logsdon, “Using a meiosis detection toolkit to investigate ancient asexual ‘scandals’ and the evolution of sex,” *BioEssays*, vol. 30, no. 6. pp. 579–589, 2008, doi: 10.1002/bies.20764.

[162] J. Dacks and A. J. Roger, “The First Sexual Lineage and the Relevance of Facultative Sex,” *J. Mol. Evol.*, vol. 48, no. 6, pp. 779–783, 1999, doi: 10.1007/pl00013156.

[163] P. G. Hofstatter and D. J. G. Lahr, “All Eukaryotes Are Sexual, unless Proven Otherwise: Many So-Called Asexuals Present Meiotic Machinery and Might Be Able to Have Sex,” *BioEssays*, vol. 41, no. 6, 2019, doi: 10.1002/bies.201800246.

[164] P. G. Hofstatter, G. M. Ribeiro, A. L. Porfírio-Sousa, and D. J. G. Lahr, “The Sexual Ancestor of all Eukaryotes: A Defense of the ‘Meiosis Toolkit’: A Rigorous Survey Supports the Obligate Link between Meiosis Machinery and Sexual Recombination.,” *BioEssays*, vol. 42, no. 9, p. e2000037, Sep. 2020, doi: 10.1002/bies.202000037.

[165] L. Peacock *et al.*, “Identification of the meiotic life cycle stage of *Trypanosoma brucei* in the tsetse fly,” *Proc. Natl. Acad. Sci.*, vol. 108, no. 9, pp. 3671–3676, 2011, doi: 10.1073/pnas.1019423108.

[166] L. Peacock, M. Bailey, M. Carrington, and W. Gibson, “Meiosis and haploid gametes in the pathogen Trypanosoma brucei.,” *Curr. Biol.*, vol. 24, no. 2, pp. 181–186, Jan. 2014, doi: 10.1016/j.cub.2013.11.044.

[167] C. M. O’Gorman, H. T. Fuller, and P. S. Dyer, “Discovery of a sexual cycle in the opportunistic fungal pathogen Aspergillus  fumigatus.,” *Nature*, vol. 457, no. 7228, pp. 471–474, Jan. 2009, doi: 10.1038/nature07528.

[168] S. K. Maciver, “Ancestral Eukaryotes Reproduced Asexually, Facilitated by Polyploidy: A Hypothesis,” *BioEssays*, vol. 41, no. 1900152, pp. 1–8, 2019, doi: 10.1002/bies.201900152.

[169] S. C. Stearns, “The evolutionary maintenance of sexual reproduction: The solutions proposed for a longstanding problem,” *J. Genet.*, vol. 69, no. 1, pp. 1–10, 1990, doi: 10.1007/BF02931662.

[170] E. Hörandl and F. Hadacek, “The oxidative damage initiation hypothesis for meiosis.,” *Plant Reprod.*, vol. 26, no. 4, pp. 351–367, Dec. 2013, doi: 10.1007/s00497-013-0234-7.

[171] S. P. Otto, “The Evolutionary Enigma of Sex.,” *Am. Nat.*, vol. 174, no. S1, pp. S1–S14, Jul. 2009, doi: 10.1086/599084.

[172] J. A. Birdsell and C. Wills, “The Evolutionary Origin and Maintenance of Sexual

Recombination: A Review of Contemporary Models," in *Evolutionary Biology*, R. J. Macintyre and M. T. Clegg, Eds. Boston, MA: Springer US, 2003, pp. 27–138.

[173] S. Schoustra, H. D. Rundle, R. Dali, and R. Kassen, "Fitness-Associated Sexual Reproduction in a Filamentous Fungus," *Curr. Biol.*, vol. 20, no. 15, pp. 1350–1355, 2010, doi: https://doi.org/10.1016/j.cub.2010.05.060.

[174] A. Del Mondo *et al.*, "A spotlight on Rad52 in cyanidiophytina (Rhodophyta): A relic in algal heritage," *Plants*, vol. 8, no. 2, 2019, doi: 10.3390/plants8020046.

[175] E. Hörandl, "A combinational theory for maintenance of sex.," *Heredity (Edinb).*, vol. 103, no. 6, pp. 445–457, Dec. 2009, doi: 10.1038/hdy.2009.85.

[176] Z. Wei *et al.*, "The cotton endocycle-involved protein SPO11-3 functions in salt stress via integrating leaf stomatal response, ROS scavenging and root growth," *Physiol. Plant.*, vol. 167, no. 1, pp. 127–141, Sep. 2019, doi: https://doi.org/10.1111/ppl.12875.

[177] A. M. Nedelcu, "Sex as a response to oxidative stress: stress genes co-opted for sex.," *Proceedings. Biol. Sci.*, vol. 272, no. 1575, pp. 1935–1940, Sep. 2005, doi: 10.1098/rspb.2005.3151.

[178] A. M. Nedelcu and R. E. Michod, "Sex as a response to oxidative stress: the effect of antioxidants on sexual induction in a facultatively sexual lineage.," *Proceedings. Biol. Sci.*, vol. 270 Suppl, no. Suppl 2, pp. S136-9, Nov. 2003, doi: 10.1098/rsbl.2003.0062.

[179] K. S. Bawa and J. H. Beach, "Evolution of Sexual Systems in Flowering Plants," *Ann. Missouri Bot. Gard.*, vol. 68, no. 2, pp. 254–274, Nov. 1981, doi: 10.2307/2398798.

[180] S. M. Coelho, L. Mignerot, and J. M. Cock, "Origin and evolution of sex-determination systems in the brown algae," *New Phytologist*, vol. 222, no. 4. Blackwell Publishing Ltd, pp. 1751–1756, 2019, doi: 10.1111/nph.15694.

[181] J. A. Fraser and J. Heitman, "Evolution of fungal sex chromosomes," *Mol. Microbiol.*, vol. 51, no. 2, pp. 299–306, 2004, doi: 10.1046/j.1365-2958.2003.03874.x.

[182] R. F. Hoekstra, “On the asymmetry of sex: Evolution of mating types in isogamous populations,” *J. Theor. Biol.*, vol. 98, no. 3, pp. 427–451, 1982, doi: https://doi.org/10.1016/0022-5193(82)90129-1.

[183] J. Lehtonen, H. Kokko, and G. A. Parker, “What do isogamous organisms teach us about sex and the two sexes?,” *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 371, no. 1706, p. 20150532, Oct. 2016, doi: 10.1098/rstb.2015.0532.

[184] S. M. Coelho, J. Gueno, A. P. Lipinska, J. M. Cock, and J. G. Umen, “UV Chromosomes and Haploid Sexual Systems,” *Trends in Plant Science*, vol. 23, no. 9. Elsevier Ltd, pp. 794–807, 2018, doi: 10.1016/j.tplants.2018.06.005.

[185] C. Destombe, M. Valero, P. Vernet, and D. Couvet, “What controls haploid—diploid ratio in the red alga, Gracilaria verrucosa?,” *J. Evol. Biol.*, vol. 2, no. 5, pp. 317–338, Sep. 1989, doi: https://doi.org/10.1046/j.1420-9101.1989.2050317.x.

[186] K. Avia *et al.*, “Genetic diversity in the UV sex chromosomes of the brown alga Ectocarpus,” *Genes (Basel).*, vol. 9, no. 6, 2018, doi: 10.3390/genes9060286.

[187] A. M. Schurko, M. Neiman, and J. M. Logsdon, “Signs of sex: what we know and how we know it,” *Trends in Ecology and Evolution*. 2009, doi: 10.1016/j.tree.2008.11.010.

[188] E. J. Pritham, “Transposable elements and factors influencing their success in eukaryotes.,” *J. Hered.*, vol. 100, no. 5, pp. 648–655, 2009, doi: 10.1093/jhered/esp065.

[189] R. C. Lewontin, “The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models,” *Genetics*, vol. 49, no. 1, pp. 49–67, 1964, doi: 10.1093/genetics/49.1.49.

[190] S. Kim *et al.*, “Recombination and linkage disequilibrium in Arabidopsis thaliana,” *Nat. Genet.*, vol. 39, no. 9, pp. 1151–1155, 2007, doi: 10.1038/ng2115.

[191] “Culture Collection of Autotrophic Organisms.” https://ccala.butbn.cas.cz.

[192] A. Marchler-Bauer and S. H. Bryant, “CD-Search: protein domain annotations on the fly.,” *Nucleic Acids Res.*, vol. 32, no. Web Server issue, pp. W327-31,

Jul. 2004, doi: 10.1093/nar/gkh454.

[193] S. Lu *et al.*, “CDD/SPARCLE: the conserved domain database in 2020.,” *Nucleic Acids Res.*, vol. 48, no. D1, pp. D265–D268, Jan. 2020, doi: 10.1093/nar/gkz991.

[194] H. Li *et al.*, “The Sequence Alignment/Map format and SAMtools,” vol. 25, no. 16, pp. 2078–2079, 2009, doi: 10.1093/bioinformatics/btp352.

[195] Broad Institute, “Picard Toolkit v.2.21.6.” http://broadinstitute.github.io/picard/.

[196] A. McKenna *et al.*, “The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.,” *Genome Res.*, vol. 20, no. 9, pp. 1297–1303, Sep. 2010, doi: 10.1101/gr.107524.110.

[197] B. S. Pedersen and A. R. Quinlan, “Mosdepth: quick coverage calculation for genomes and exomes.,” *Bioinformatics*, vol. 34, no. 5, pp. 867–868, Mar. 2018, doi: 10.1093/bioinformatics/btx699.

[198] P. Danecek *et al.*, “Twelve years of SAMtools and BCFtools.,” *Gigascience*, vol. 10, no. 2, Feb. 2021, doi: 10.1093/gigascience/giab008.

[199] S. Purcell, “PLINK.” http://pngu.mgh.harvard.edu/purcell/plink/.

[200] S. Purcell *et al.*, “PLINK: a tool set for whole-genome association and population-based linkage analyses.,” *Am. J. Hum. Genet.*, vol. 81, no. 3, pp. 559–575, Sep. 2007, doi: 10.1086/519795.

[201] A. Miles, “scikit-allel,” 2015. https://scikit-allel.readthedocs.io/en/stable/index.html.

[202] U. Abdu, A. González-Reyes, A. Ghabrial, and T. Schüpbach, “The Drosophila spn-D gene encodes a RAD51C-like protein that is required exclusively during meiosis.,” *Genetics*, vol. 165, no. 1, pp. 197–204, Sep. 2003, doi: 10.1093/genetics/165.1.197.

[203] J. J. Sekelsky, K. S. McKim, G. M. Chin, and R. S. Hawley, “The Drosophila meiotic recombination gene mei-9 encodes a homologue of the yeast excision repair protein Rad1.,” *Genetics*, vol. 141, no. 2, pp. 619–627, Oct. 1995, doi: 10.1093/genetics/141.2.619.

[204] E. Derelle *et al.*, “Genome analysis of the smallest free-living eukaryote Ostreococcus tauri unveils many unique features.” 2006, [Online]. Available: www.pnas.orgcgidoi10.1073pnas.0604795103.

[205] N. Grimsley, B. Péquin, C. Bachy, H. Moreau, and G. Piganeau, “Cryptic sex in the smallest eukaryotic marine green alga,” *Mol. Biol. Evol.*, vol. 27, no. 1, pp. 47–54, 2010, doi: 10.1093/molbev/msp203.

[206] R. Blanc-mathieu *et al.*, “Population genomics of picophytoplankton unveils novel chromosome hypervariability,” no. July, 2017.

[207] M. Lynch *et al.*, “Genetic drift, selection and the evolution of the mutation rate,” *Nature Reviews Genetics*, vol. 17, no. 11. Nature Publishing Group, pp. 704–714, 2016, doi: 10.1038/nrg.2016.104.

[208] D. Graur and W.-H. L. Li, *Fundamentals of Molecular Evolution*. 2000.

[209] S. Wielgoss *et al.*, “Mutation Rate Inferred From Synonymous Substitutions in a Long-Term Evolution Experiment With Escherichia coli.,” *G3 (Bethesda).*, vol. 1, no. 3, pp. 183–186, Aug. 2011, doi: 10.1534/g3.111.000406.

[210] J. W. Drake, “A constant rate of spontaneous mutation in DNA-based microbes.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 88, no. 16, pp. 7160–7164, Aug. 1991, doi: 10.1073/pnas.88.16.7160.

[211] D. S. Lawrie, P. W. Messer, R. Hershberg, and D. A. Petrov, “Strong Purifying Selection at Synonymous Sites in D. melanogaster,” *PLOS Genet.*, vol. 9, no. 5, p. e1003527, May 2013, [Online]. Available: https://doi.org/10.1371/journal.pgen.1003527.

[212] P. L. Foster, A. J. Hanson, H. Lee, E. M. Popodi, and H. Tang, “On the mutational topology of the bacterial genome.,” *G3 (Bethesda).*, vol. 3, no. 3, pp. 399–407, Mar. 2013, doi: 10.1534/g3.112.005355.

[213] S. Andrews, “FastQC: A Quality Control Tool for High Throughput Sequence Data,” 2010. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

[214] “FastQC.” Jun. 2015, [Online]. Available: https://qubeshub.org/resources/fastqc.

[215] E. Garrison and G. Marth, “Haplotype-based variant detection from short-read sequencing.” arXiv, 2012, doi: 10.48550/ARXIV.1207.3907.

[216] M. Martin *et al.*, “WhatsHap: fast and accurate read-based phasing,” *bioRxiv*, 2016, doi: 10.1101/085050.

[217] B. Grüning *et al.*, “Bioconda: sustainable and comprehensive software distribution for the life sciences,” *Nat. Methods*, vol. 15, no. 7, pp. 475–476, 2018, doi: 10.1038/s41592-018-0046-7.

[218] P. Danecek *et al.*, “The variant call format and VCFtools,” *Bioinformatics*, vol. 27, no. 15, pp. 2156–2158, 2011, doi: 10.1093/bioinformatics/btr330.

[219] S. Anders, P. T. Pyl, and W. Huber, “HTSeq - a Python framework to work with high-throughput sequencing data.,” *Bioinformatics*, vol. 31, no. 2, pp. 166–169, Jan. 2015, doi: 10.1093/bioinformatics/btu638.

[220] M. Krasovec, A. Eyre-Walker, S. Sanchez-Ferandin, and G. Piganeau, “Spontaneous Mutation Rate in the Smallest Photosynthetic Eukaryotes.,” *Mol. Biol. Evol.*, vol. 34, no. 7, pp. 1770–1779, Jul. 2017, doi: 10.1093/molbev/msx119.

[221] R. W. Ness, A. D. Morgan, N. Colegrave, and P. D. Keightley, “Estimate of the spontaneous mutation rate in Chlamydomonas reinhardtii,” *Genetics*, vol. 192, no. 4, pp. 1447–1454, 2012, doi: 10.1534/genetics.112.145078.

[222] S. Kucukyildirim, M. Behringer, E. M. Williams, T. G. Doak, and M. Lynch, “Estimation of the Genome-Wide Mutation Rate and Spectrum in the Archaeal Species Haloferax volcanii,” *Genetics*, vol. 215, no. 4, pp. 1107–1116, Aug. 2020, doi: 10.1534/genetics.120.303299.

[223] C. Seoighe and K. H. Wolfe, “Yeast genome evolution in the post-genome era,” *Curr. Opin. Microbiol.*, vol. 2, no. 5, pp. 548–554, 1999, doi: 10.1016/S1369-5274(99)00015-6.

[224] J. W. Drake, “Avoiding Dangerous Missense: Thermophiles Display Especially Low Mutation Rates,” *PLOS Genet.*, vol. 5, no. 6, p. e1000520, Jun. 2009, [Online]. Available: https://doi.org/10.1371/journal.pgen.1000520.

[225] D. W. Grogan, G. T. Carver, and J. W. Drake, “Genetic fidelity under harsh

conditions: Analysis of spontaneous mutation in the thermoacidophilic archaeon Sulfolobus acidocaldarius," *Proc. Natl. Acad. Sci.*, vol. 98, no. 14, pp. 7928–7933, Jul. 2001, doi: 10.1073/pnas.141113098.

[226] R. R. Mackwan, G. T. Carver, J. W. Drake, and D. W. Grogan, "An unusual pattern of spontaneous mutations recovered in the halophilic archaeon Haloferax volcanii.," *Genetics*, vol. 176, no. 1, pp. 697–702, May 2007, doi: 10.1534/genetics.106.069666.

[227] W. Sung, M. S. Ackerman, S. F. Miller, T. G. Doak, and M. Lynch, "Drift-barrier hypothesis and mutation-rate evolution," *Proc. Natl. Acad. Sci.*, vol. 109, no. 45, pp. 18488–18492, Nov. 2012, doi: 10.1073/pnas.1216223109.

[228] L. Calarco and J. Ellis, "Species diversity and genome evolution of the pathogenic protozoan parasite, Neospora caninum," *Infect. Genet. Evol.*, vol. 84, p. 104444, 2020, doi: https://doi.org/10.1016/j.meegid.2020.104444.

[229] W. L. Hamilton *et al.*, "Extreme mutation bias and high AT content in Plasmodium falciparum.," *Nucleic Acids Res.*, vol. 45, no. 4, pp. 1889–1901, Feb. 2017, doi: 10.1093/nar/gkw1259.

[230] L. Calarco, J. Barratt, and J. Ellis, "Genome Wide Identification of Mutational Hotspots in the Apicomplexan Parasite Neospora caninum and the Implications for Virulence.," *Genome Biol. Evol.*, vol. 10, no. 9, pp. 2417–2431, Sep. 2018, doi: 10.1093/gbe/evy188.

[231] J. D. Barry, M. L. Ginger, P. Burton, and R. Mcculloch, "Why are parasite contingency genes often associated with telomeres?," doi: 10.1016/S0.

[232] M. J. Gardner *et al.*, "Genome sequence of the human malaria parasite Plasmodium falciparum.," *Nature*, vol. 419, no. 6906, pp. 498–511, Oct. 2002, doi: 10.1038/nature01097.

[233] B. Kazemi, "Genomic organization of Leishmania species.," *Iran. J. Parasitol.*, vol. 6, no. 3, pp. 1–18, Aug. 2011.

[234] L. C. Strotz, M. Simões, M. G. Girard, L. Breitkreuz, J. Kimmig, and B. S. Lieberman, "Getting somewhere with the Red Queen: chasing a biologically modern definition of the hypothesis," *Biol. Lett.*, vol. 14, no. 5, p. 20170734,

May 2018, doi: 10.1098/rsbl.2017.0734.

[235] M. W. Crosland and R. H. Crozier, “Myrmecia pilosula, an Ant with Only One Pair of Chromosomes.,” *Science*, vol. 231, no. 4743, p. 1278, Mar. 1986, doi: 10.1126/science.231.4743.1278.

[236] V. A. Lukhtanov, “The blue butterfly Polyommatus (Plebicula) atlanticus (Lepidoptera, Lycaenidae) holds the record of the highest number of chromosomes in the non-polyploid eukaryotic organisms.,” *Comp. Cytogenet.*, vol. 9, no. 4, pp. 683–690, 2015, doi: 10.3897/CompCytogen.v9i4.5760.

[237] O. Muravenko, I. Selyakh, N. Kononenko, and I. Stadnichuk, “Chromosome numbers and nuclear dna contents in the red microalgae cyanidium caldarium and three galdieria species,” *Eur. J. Phycol.*, vol. 36, no. 3, pp. 227–232, 2001, doi: 10.1080/09670260110001735378.

[238] J. Sims, P. Schlögelhofer, and M.-T. Kurzbauer, “From Microscopy to Nanoscopy: Defining an Arabidopsis thaliana Meiotic Atlas at the Nanometer Scale,” *Front. Plant Sci.*, vol. 12, 2021, doi: 10.3389/fpls.2021.672914.

[239] S. G. Gregory *et al.*, “The DNA sequence and biological annotation of human chromosome 1,” *Nature*, vol. 441, no. 7091, pp. 315–321, 2006, doi: 10.1038/nature04727.

[240] G. L. Lipscomb, E. M. Hahn, A. T. Crowley, and M. W. W. Adams, “Reverse gyrase is essential for microbial growth at 95 °C.,” *Extremophiles*, vol. 21, no. 3, pp. 603–608, May 2017, doi: 10.1007/s00792-017-0929-z.

[241] H. Atomi, R. Matsumi, and T. Imanaka, “Reverse gyrase is not a prerequisite for hyperthermophilic life.,” *J. Bacteriol.*, vol. 186, no. 14, pp. 4829–4833, Jul. 2004, doi: 10.1128/JB.186.14.4829-4833.2004.

[242] N. K. Chunduri and Z. Storchová, “The diverse consequences of aneuploidy,” *Nat. Cell Biol.*, vol. 21, no. 1, pp. 54–62, 2019, doi: 10.1038/s41556-018-0243-8.

[243] T. K. Ezov *et al.*, “Molecular-genetic biodiversity in a natural population of the yeast Saccharomyces cerevisiae from ‘Evolution Canyon’: microsatellite polymorphism, ploidy and controversial sexual status.,” *Genetics*, vol. 174, no.

3, pp. 1455–1468, Nov. 2006, doi: 10.1534/genetics.106.062745.

[244] A. C. Gerstein and S. P. Otto, “Ploidy and the causes of genomic evolution.,” *J. Hered.*, vol. 100, no. 5, pp. 571–581, 2009, doi: 10.1093/jhered/esp057.

[245] S. K. Maciver, “Asexual Amoebae Escape Muller’s Ratchet through Polyploidy.,” *Trends Parasitol.*, vol. 32, no. 11, pp. 855–862, Nov. 2016, doi: 10.1016/j.pt.2016.08.006.

[246] Y. Sterkers, L. Lachaud, N. Bourgeois, L. Crobu, P. Bastien, and M. Pagès, “Novel insights into genome plasticity in Eukaryotes: mosaic aneuploidy in Leishmania,” *Mol. Microbiol.*, vol. 86, no. 1, pp. 15–23, Oct. 2012, doi: https://doi.org/10.1111/j.1365-2958.2012.08185.x.

[247] B. Vincent *et al.*, “Inferring the Minimal Genome of Mesoplasma florum by Comparative Genomics and Transposon Mutagenesis,” *mSystems*, vol. 3, no. 3, pp. e00198-17, Apr. 2018, doi: 10.1128/mSystems.00198-17.

[248] R. Bleisch *et al.*, “Strain Development in Microalgal Biotechnology - Random Mutagenesis Techniques,” *Life*, vol. 12, no. 7. 2022, doi: 10.3390/life12070961.

[249] C. Delahaye and J. Nicolas, “Sequencing DNA with nanopores: Troubles and biases,” *PLoS One*, vol. 16, no. 10, p. e0257521, Oct. 2021, [Online]. Available: https://doi.org/10.1371/journal.pone.0257521.

[250] Oxford Nanopore Technologies, “Nanopore Sequencing Accuracy,” 2022. https://nanoporetech.com/accuracy.

[251] Environmental Molecular Sciences Laboratory, “The CCMEE - Culture Collection of Microorganisms from Extreme Environments as a platform for biodiversity research at EMSL.” https://www.emsl.pnnl.gov/project/49673.

[252] A. Aguilera, L. Amaral-Zettler, V. Souza-Egipsy, E. Zettler, and R. Amils, “Eukaryotic Community Structure from Río Tinto (SW, Spain), a Highly Acidic River,” in *Algae and Cyanobacteria in Extreme Environments*, J. Seckbach, Ed. Dordrecht: Springer Netherlands, 2007, pp. 465–485.