Evolution in the shadows of genome reduction - resolving molecular mechanisms and early branches in polyextremophilic Red Algae

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

Inaugural-Dissertation

zur Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Alessandro Rossoni

aus Lodi

Düsseldorf, Juli 2019

aus dem Institut für Biochemie der Pflanzen

der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der

Mathematisch-Naturwissenschaftlichen Fakultät der

Heinrich-Heine-Universität Düsseldorf

Referent:	Prof. Dr. Andreas P.M. Weber
Korreferent:	Prof. Dr. Michael Feldbrügge

Tag der mündlichen Prüfung: 08.11.2019

EIDESSTATTLICHE ERKLÄRUNG

Ich versichere an Eides Statt, dass diese Dissertation von mir selbständig und ohne unzulässige fremde Hilfe unter Beachtung der „Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf" erstellt worden ist. Die Dissertation habe ich in der vorgelegten oder in ähnlicher Form noch bei keiner anderen Institution eingereicht. Ich habe bisher keine erfolglosen Promotionsversuche unternommen.

Düsseldorf, der 25.07.2019

____

Alessandro Rossoni

“People have done that”, *Prof. Dr. Andreas P.M. Weber*

Contents

Preface & Author’s Contribution

The present thesis starts with a summary in English and German language, followed by an introduction about the relevant topics and by 5 independent manuscripts. The author’s contribution (AWR) to each manuscript is listed below.

The first manuscript “**Unexpected conservation of the RNA splicing apparatus in the highly streamlined genome of Galdieria sulphuraria**” was published in BMC Evolutionary Biology in April 2018 (Qiu et al., 2018). This work reveals the unique nature of *Galdieria sulphuraria* 074W among red algae with respect to the conservation of the spliceosomal machinery and introns, possibly as a strategy for generating functional diversity to support their independent lifestyle in spite of its limited coding capacity. A complex spliceosomal machinery can compensate for gene loss in the shadows of nuclear genome reduction. AWR performed the wet-lab work (Cell culture, RNA-extraction, library preparation), handled the sequencing process and mapped the resulting reads onto the genome. AWR contributed parts of the manuscript.

The second manuscript, “**Streamlined genomes of polyextremophile red algae (Cyanidiales) maintain 1% horizontal gene transfers with diverse adaptive functions”** was published in eLife in June 2019 org (Rossoni et al., 2019). This work analyses the extent and role of horizontal gene transfer (HGT) in eukaryotes, one of the most critically disputed topics in evolutionary biology. The applicability of HGT versus Differential Loss is discussed. The presumption that eukaryotic HGT unfolds in the eukaryotic kingdom in the same manner that prokaryotic HGT unfolds in the prokaryotic kingdom is incorrect. The maintenance of HGT in a scenario of selection for genome reduction underlines the importance of HGT to eukaryote evolution. AWR co-designed the experiment, applied for additional funding (SMRT® Microbe), performed the wet-lab work (Cell culture, DNA-extraction, library preparation), handled the sequencing process, assembled the genomes and predicted gene models. Data analysis was performed by AWR, except the generation of phylogenetic inferences. The manuscript-draft was written by AWR. AWR also contributed to subsequent versions of the manuscript.

The third manuscript carries the title: "**Cold Acclimation of the Thermoacidophilic Red Alga Galdieria sulphuraria - Changes in Gene Expression and Involvement of Horizontally Acquired Genes**" and was published in Plant Cell Physiology in December 2018 (Rossoni and Schoenknecht et al., 2018). Here, we applied RNA-sequencing to obtain insights into the acclimation of the thermoacidophilie *Galdieria sulphuraria* 074W towards temperatures 20°C below its growth optimum and to study how horizontally acquired genes contribute to cold acclimation. Besides finding some interesting and unexpected acclimation strategies, such as the upregulation of C1 metabolism, we provide evidence that genes acquired by horizontal gene transfer play an important role in cold acclimation of this organism. The latter is an important contribution to the ongoing debate to which degree adaptive evolution in eukaryotes has been driven by horizontal gene transfer. AWR performed the wet-lab work (Cell culture, RNA-extraction, library preparation), handled the sequencing process and mapped the resulting reads onto the genome. AWR performed data analysis, wrote parts of the manuscript and significantly contributed to the outcome and discussion.

The forth manuscript "**Photorespiratory glycolate oxidase is essential for the red algae *Cyanidioschyzon merolae* under ambient CO_2 conditions**" was published in the Journal of Experimental Botany in February 2016. It illustrates the significance of photorespiration in red algae (Rademacher *et al*., 2016). This publication focuses on the relevance of the enzyme glycolate oxidase in the extremophile red alga *Cyanidioschyzon merolae* under photorespiratory conditions. Horizontally acquired genes are not significantly regulated under non-stress conditions, supporting the hypothesis of stress adaptational benefits. AWR performed RNA-Seq read mapping onto the genome and was partially involved in data analysis.

The fifth manuscript "**Systems Biology of Cold Adaptation in the Polyextremophilic Red Alga *Galdieria sulphuraria***" was published in the Journal Frontiers in Microbiology, section Extreme Microbiology, in May 2019. Cold stress was applied to *Galdieria sulphuraria* for more than 100 generations. Natural selection towards temperature tolerance is a systems biology problem which requires the gradual orchestration of an intricate gene network and deeply nested regulators. AWR co-designed the experiment, performed the wet-lab work, handled the sequencing process and performed data analysis. The manuscript-draft was written by AWR.

REFERENCES:

Manuscript 1: Qiu H, Rossoni AW, Weber APM, Yoon HS and Bhattacharya D. 2018. Unexpected conservation of the RNA splicing apparatus in the highly streamlined genome of Galdieria sulphuraria. BMC Evolutionary Biology 18.

Manuscript 2: Rossoni AW, Price DC, Seger M, Lyska D, Lammers P, Bhattacharya D and Weber APM. 2019. Streamlined genomes of polyextremophile red algae (Cyanidiales) maintain 1% horizontal gene transfers with diverse adaptive functions. eLife.

Manuscript 3: Rossoni AW & Schoenknecht G, Rupp RL, Lee HJ, Flachbart S, Weber APM, Eisenhut M. 2018. Cold Acclimation of the Thermoacidophilic Red Alga *Galdieria sulphuraria* - Changes in Gene Expression and Involvement of Horizontally Acquired Genes. Plant and Cell Physiology.

Manuscript 4: Rademacher N, Wrobel TJ, Rossoni AW, Kurz S, Bräutigam A, Weber APM, Eisenhut M. 2017. Transcriptional response of the extremophile red alga Cyanidioschyzon merolae to changes in CO_2 concentrations. Journal of Plant Physiology 217.

Manuscript 5: Rossoni AW, Weber APM. 2019. Systems Biology of Cold Adaptation in the Polyextremophilic Red Alga *Galdieria sulphuraria*”. 2019. Frontiers in Microbiology.

OTHER PUBLICATIONS:

Schmitz J, Dittmar JC, Brockmann JD, Schmidt M, Hüdig M, Rossoni AW, Maurino VG. 2017. Defense against Reactive Carbonyl Species Involves at Least Three Subcellular Compartments Where Individual Components of the System Respond to Cellular Sugar Status. The Plant Cell 17

Schmitz J, Rossoni AW, Maurino VG. 2018. Dissecting the Physiological Function of Plant Glyoxalase I and Glyoxalase I-Like Proteins. Frontiers in Plant Science 9

I.1 Summary

The new wealth of data made available by next generation sequencing technologies has spurred a novel generation of investigations that are challenging established concepts in the field of eukaryotic genome evolution. First, the classic notion of molecular neofunctionalization through genome/gene duplication and expansion, exemplified by plants and animals, is being contrasted by free-living organisms with highly reduced genomes exhibiting a surprising adaptational capacity. Second, the frequency and extent of eukaryotic horizontal gene transfer (HGT) as potential driver of adaptive evolution infringes the dogma of strict vertical inheritance in eukaryotes. Here, we chose the Cyanidiales, a group of unicellular and polyextremophile red algae, as model organisms for studying the above-mentioned controversies due to their broad ecological boundaries in spite of highly streamlined genomes (12 Mbp – 16 Mbp) and miniature gene inventories (4,400 – 6,500 genes).

The reduced genetic capacity of the Cyanidiales results from of two phases of genome reduction along their evolutionary history which led to the loss of multiple core eukaryotic traits (e.g. flagella, basal bodies, GPI anchor biosynthesis, etc.) that are lacking in many red algae. In contrast to this trend, we observed the upkeep of a complex splicing machinery (SM) in the *Galdieria* lineage. *Galdieria sulphuraria* maintained 149 of the original >157 spliceosomal components that were present in the last common red algal ancestor [**Manuscript 1**]. In addition, it continued increasing its intron number over time – an effect that is positively correlated to increased transcript diversity. We subsequently hypothesized that the expansion of introns and the upkeep of an increased SM burden in a scenario of natural selection towards genome reduction may be explained by the benefits provided to *Galdieria sulphuraria* by this costly system, e.g. greater adaptive capacity through alternative splicing and increased transcript diversity. RNA expression analysis on cold stressed *Galdieria sulphuraria* revealed significant temperature-dependent alternative splicing, mostly through intron retention. We conclude that spliceosomal complexity and intron richness constitute a potential mechanism to counteract the consequences of genome reduction and for generating functional diversity to ameliorate gene loss in free living organisms.

Another possible strategy to boost the adaptational ability of free-living eukaryotes can be through the acquisition of foreign genes via HGT, one of the most hotly debated topics in modern evolutionary biology. In spite of increasing evidence for eukaryotic HGT from various genome sequencing projects, proponents of eukaryotic HGT have failed to deliver quantitative explanations for the absence of a eukaryotic pangenome as well as the lack of cumulative effects in horizontally acquired genes. In comparison to prokaryotes, the sequenced eukaryotic genomes are still few and broadly dispersed across the tree of life. We addressed these issues by sequencing 10 novel Cyanidiales genomes from 9 geographically isolated habitats (combined with the already-published *Cyanidioschyzon merolae* 10D, *Galdieria sulphuraria* 074W and *Galdieria phlegrea DBV009*) and determined that 1% (96 orthogroups) of their gene inventory is HGT-derived [**Manuscript 2**]. In addition to phylogenetic inference, HGT candidates significantly differ from native genes in GC-content (> 1%), number of splice sites (0.97 - 1.36 fewer exons per gene) and differential gene expression in response to stress (HGT candidates are enriched within the differentially expressed genes). The majority of HGT candidates also originated from extremophilic prokaryotes, e.g. *Sulfobacillus thermosulfidooxidans*, that share similar habitats as the Cyanidiales and encodes functions related to polyextremophile traits, further supporting the narrative regarding their beneficial role during adaptation. By analyzing gain and loss pattern we conclude that the absence of a eukaryotic pangenome and cumulative effects can be explained by the rarity of eukaryotic HGT and a stronger propensity of HGT candidates to gene erosion in comparison to native genes. Because eukaryotic HGT is the exception rather than the rule, its quantity in eukaryotic genomes does not need to increase as a function of time and likely reached equilibrium between acquisition and erosion in the distant past. The presumption that eukaryotic HGT will (and should) unfold in the eukaryotic kingdom in the same manner as does prokaryotic HGT among prokaryotes is highly questionable.

Subsequently, we tested the – thus far – hypothesized involvement of HGT genes in stress adaptation through multiple experiments. Gene expression was measured using RNA-Sequencing in cold-stressed *Galdieria sulphuraria* 074W [**Manuscript 3**] and “CO_2-stressed” *Cyanidioschyzon merolae* 10D [**Manuscript 4**]. In both cases, strong transcription of HGT genes was measured, providing evidence for the successful integration of HGT genes into the transcriptional machinery of the host (the average read count for HGT candidates in *Galdieria sulphuraria* 074W and *Cyanidioschyzon merolae* 10D were 130 CPM and 184 CPM, respectively). In *Galdieria sulphuraria* 074W, HGT genes reacted significantly stronger to temperature changes in comparison to native genes, supporting the presumed involvement of HGTs to stress adaptation. Furthermore, some interesting and unexpected acclimation strategies under cold stress were found, such as the upregulation of C1 metabolism, especially the S-adenosylmethionine cycle and folate cycle. In contrast, HGT genes in *Cyanidioschyzon merolae* 10D did not differ in response intensity from native genes. Since high CO_2 does not constitute a stressful condition in phototrophic organisms, no transcriptional differences between the two gene sets were expected. Although there is no direct way to test the past impact of HGT on the evolutionary trajectory of the Cyanidiales, taken together, we interpret these results as supporting indications for their positive contribution with regards to extremophile adaptation.

Finally, we analyzed the effects of prolonged exposure to cold temperature upon the genetics and growth phenotype of *Galdieria sulphuraria* RT22 for a period spanning >100 generations [**Manuscript 5**]. DNA-sequencing revealed 757 variants located on 429 genes (6.1% of the transcriptome) mostly encoding functions involved in cell cycle (33.8% of enriched GO-Terms), gene regulation (20.9%) and signaling (7.7%). At epigenetic level, variants targeting the intergenic region were enriched in CpG islands. As a consequence, cold stressed samples grew ~30% faster at the end of the experiment in comparison to the starting population. Rather than targeting specific pathways, natural selection towards temperature tolerance is a systems biology problem which requires the gradual orchestration of an intricate gene network and deeply nested regulators at genetic and epigenetic level equally affecting native and HGT genes.

I.2 Zusammenfassung

Die Vielfalt neuer genetischer Daten, welche im Verlauf des letzten Jahrzehnts mittels moderner Sequenziertechnologien generiert wurde, hat zu einer ganzen Reihe innovativer Untersuchungen geführt. Diese lang etablierten Grundsätze auf dem Gebiet der Evolutionstheorie hinterfragen. So steht die klassische Vorstellung molekularer Neofunktionalisierung durch Genom- und Genduplikationen in Eukaryoten, am besten exemplifiziert im Pflanzen- und Tierreich, im starken Kontrast zur überraschenden Anpassungsfähigkeit von freilebenden Mikroorganismen mit ihren sehr kleinen Genomen. Die Cyanidiales, eine Gruppe einzelliger und polyextremophiler Rotalgen (*Galdieria*, *Cyanidium* und *Cyanidioschyzon*), repräsentieren einen solchen Fall. Trotz ihrer Anpassungsfähigkeit an die verschiedensten Umweltbedingungen besitzen die Cyanidiales stark reduzierte Genome (12 Mbp - 16 Mbp) und verfügen nur über ein minimales Geninventar (4.400 - 6.500 Gene). In den Cyanidiales korreliert die Anzahl der Gene somit nicht positiv mit ihrer Adaptationsfähigkeit. Darüber hinaus besitzen die Cyanidiales die Fähigkeit, fremdes Genmaterial aus der Umgebung in die eigene DNA zu integrieren. Dieser Mechanismus des horizontalen Gentransfers (HGT) widerspricht jedoch dem Dogma der streng vertikalen Vererbung nach Darwin und Mendel in Eukaryoten. Ob, neben Bakterien und Archaeen, auch Eukaryoten fremde Gene in das eigene Erbgut integrieren können und daraus adaptive Vorteile erlangen, ist eine der größten Kontroversen der modernen Evolutionsbiologie.

Die reduzierte genetische Kapazität der Cyanidiales ist das Ergebnis zweier Phasen der Genomreduktion im Verlauf ihrer 1,3 Milliarden Jahre alten Evolutionsgeschichte. Im Verlauf ihrer Evolution sind mehrere eukaryotische Kernmerkmale verloren gegangen, wie z.B. Flagellen, Basalkörper, GPI-Ankerbiosynthese etc. Im Gegensatz zu diesem „Verlusttrend“ hat *Galdieria sulphuraria* eine komplexe Spleißmaschinerie (SM) behalten [**Manuskript 1**]. Darüber hinaus steigerte *Galdieria sulphuraria* die Anzahl der Introns im Verlauf der Zeit von 1.677 (letzter gemeinsamer Vorfahre aller Rotalgen) auf 13.245 - ein Effekt, der positiv mit einer erhöhten Transkriptvielfalt durch alternatives Spleißen korreliert. Diese wiederum hängt mit einem erhöhten Adaptationspotential zusammen. Daraus erfolgte die Hypothese, dass die Expansion von Introns und die Bewahrung einer komplexen SM in einem Szenario natürlicher Selektion zur Genomreduktion nur durch adaptive Vorteile für *Galdieria sulphuraria*

erklärt werden kann. Diese Hypothese wurde anhand einer RNA-Expressionsanalyse getestet, bei der *Galdieria sulphuraria* kälteren Temperaturen ausgesetzt wurde. Als Folge konnte signifikantes, temperaturabhängiges, alternatives Spleißen beobachtet werden, meist durch Intron-Retention. Wir schließen daraus, dass sowohl die Komplexität der Spleißmaschinerie als auch die hohe Anzahl an Introns potenzielle Kompensationsmechanismen für freilebende Organismen darstellen. Somit kann zusätzliche funktionelle Diversität erzeugt werden um den Folgen der Genomreduktion entgegenzuwirken.

Eine weitere mögliche Strategie zur Steigerung der Anpassungsfähigkeit könnte durch den Erwerb fremder Gene über HGT stattfinden. Trotz zunehmender Evidenz für eukaryotischen HGT aus verschiedenen Genomsequenzierungsprojekten konnten weder Erklärungen für das Fehlen eines eukaryotischen Pangenoms noch für das Fehlen kumulativer Effekte in HGT-Genen gefunden werden. Einer der wahrscheinlichsten Gründe dafür ist die geringe Anzahl an sequenzierten Genomen verwandter Spezies innerhalb des eukaryotischen Stammbaums. Wir haben uns dieser Problematik angenommen und 10 neue Cyanidiales-Genome assembliert. Anhand phylogenetischer Inferenz konnte festgestellt werden, dass 1% der Cyanidiales-Gene einen horizontalen Ursprung aufweist [**Manuskript 2**]. Diese 96 orthologen Gruppen (641 Gene insgesamt) weisen signifikante Unterschiede zur nativen Genpopulation auf, wie z.B. einen erhöhten GC-Gehalt und eine geringere Anzahl an Exons pro Gen. Zudem stammte die überwiegende Mehrheit der HGT-Gene von polyextremophilen Prokaryoten ab, welche ähnliche Lebensräume wie die Cyanidiales bewohnen. Diese kodieren mehrheitlich polyextremophile Eigenschaften, welche die Anpassung der Cyanidiales an extreme Habitate gefördert haben. Aus der Analyse der Genakquisitions- und Genverlustmuster innerhalb der Cyanidiales konnte gefolgert werden, dass sowohl das Fehlen eines eukaryotischen Pangenoms als auch die Abwesenheit kumulativer Effekte durch zwei Faktoren erklärt werden kann: erstens, die Seltenheit des eukaryotischen HGTs und zweitens die stärkere Neigung der HGT-Gene zur Generosion. Entsprechend muss die Menge an HGT-Genen in eukaryotischen Genomen nicht zwangsweise mit der Zeit zunehmen. Ein Gleichgewicht zwischen Erwerb und Erosion könnte somit bereits in der fernen Vergangenheit erreicht worden sein. Die Annahme, dass sich HGT in der eukaryotischen Lebensdomäne auf dieselbe Weise entfaltet wie HGT in der prokaryotischen Lebensdomäne, ist nicht zwingend korrekt.

Zusätzlich wurde die - bisher prognostizierte - Beteiligung von HGT-Genen in der Stressantwort verschiedener Cyanidiales als Indikator für ihre angenommene Rolle in der adaptiven Evolution gemessen. Die systemische Genexpressionsänderung von kältegestressten *Galdieria sulphuraria* 074W [**Manuskript 3**] und CO_2-gestressten *Cyanidioschyzon merolae* 10D [**Manuskript 4**] wurde mittels RNA-Sequenzierung gemessen. In beiden Fällen konnte eine starke Transkription von HGT-Genen bestätigt werden. Dies lieferte den ersten Nachweis für ihre funktionelle Integration in die Transkriptionsmaschinerie des Wirts. Die durchschnittliche Transkriptionsrate der HGT-Gene in *Galdieria sulphuraria* 074W betrug 130 CPM und in *Cyanidioschyzon merolae* 10D 184 CPM. Des Weiteren reagierten HGT-Gene in *Galdieria sulphuraria* 074W im Vergleich zu nativen Genen signifikant stärker auf Kältestress, eine Beobachtung, welche die vermutete Rolle von HGT in Bezug auf die Anpassung der Cyanidiales an abiotischen Stress unterstützt. Darüber hinaus wurden einige interessante und unerwartete Akklimatisierungsstrategien beobachtet, wie z.B. die Hochregulierung des C1-Metabolismus, insbesondere des S-Adenosylmethionins-Zyklus und des Folat-Zyklus. Im Gegensatz dazu unterscheidet sich die Änderung der Transkriptionsrate von HGT-Genen in *Cyanidioschyzon merolae* 10D nicht signifikant von der Transkriptionsrate der nativen Gene. Da ein hoher CO_2-Gehalt in phototrophen Organismen allgemein keinen Stresszustand hervorruft (der photosynthetische Oxidationsstress wird sogar reduziert), wurden *a priori* keine entsprechenden Transkriptionsunterschiede zwischen den beiden Genpopulationen erwartet. Zwar kann die Wirkung von HGT auf die vergangene Evolutionstrajektorie der Cyanidiales heute nicht mehr direkt getestet werden, jedoch interpretieren wir diese Ergebnisse als unterstützende Hinweise für einen positiven Beitrag von HGT auf die Evolution der Cyanidiales.

Zuletzt wurden die Auswirkungen einer längeren Kälteeinwirkung auf die Genetik und das Wachstum von *Galdieria sulphuraria* RT22 über einen Zeitraum von > 100 Generationen gemessen [**Manuskript 5**]. Die DNA-Sequenzierung ergab 757 Varianten, die sich auf 429 Gene befanden (6,1% des gesamten Geninventars) und überwiegend auf Genen lokalisiert sind, welche mit Funktionen wie Zellzyklus, Genregulation, oder Signalübertragung in Verbindung stehen. Auf epigenetischer Ebene waren Mutationen in CpG-Inseln signifikant angereichert. Kältegestresste Proben wuchsen am Ende des Experiments im Vergleich zur Ausgangsbevölkerung um ca. 30% schneller. Die natürliche Selektion zu höherer Temperaturtoleranz erfordert dementsprechend die schrittweise Orchestrierung eines komplizierten Gennetzwerks mit tief verschachtelten Regulationsprozessen auf genetischer und epigenetischer Ebene. Sie ist nicht auf individuelle Komponenten im Einzelkontext zu reduzieren.

II. Introduction

More than 150 years have passed since Charles Darwin published his “theory of evolution by the means of natural selection” in 1895 [1]. Since then, the theory of evolution has repeatedly and successfully stood the test of time. Today’s modern synthesis of evolution, first formulated by Theodosius Dobzhansky [2], has continuously been synchronized and expanded throughout multiple disciplines of the life sciences unifying, e.g., population biology, developmental biology, molecular genetics, epigenetics, and mathematical modeling. What is left of Darwin’s original notion is now part of a much greater extended evolutionary framework [3]. Since the advent of the genomic age, most notably perceived by the broader public through the sequencing of the human genome in 2001 [4], thousands of eukaryotic, bacterial and archaeal genomes have been wholly or partially decoded [5]. This new breadth of genomic data has led to a new generation of evolutionary questions that, while not changing the overarching evolutionary narrative, are challenging some of its established concepts. This thesis contributes to the elucidation of two novel concepts using the Cyanidiales as model organisms: First, the surprising adaptational capacity of organisms with highly streamlined genomes. Second, the frequency and extent of eukaryotic horizontal gene transfer as a driver of adaptive evolution.

The Cyanidiales

The red algae (Rhodophyta) are an ancient archaeplastidal phylum that is constituted of approximately ~7.000 species [6] which separated from the eukaryotic tree of life ca 1.6 billion years ago [7], shortly after eukaryotic photosynthesis was established through primary endosymbiosis [8-10]. With an age of more than 1.3 billion years, the Cyanidiales are not only the first divergent branch within the red algae, but also define one of the oldest extant eukaryotic organisms. Hence, they are most probably the closest living relatives to the red alga that gave rise to the plastid of Chromalveolata via secondary endosymbiosis [8, 11-13] (e.g., diatoms contribute up to 50% of the organic carbon fixed annually in the world's oceans [14]). The Cyanidiales comprise the rudimental lineages of *Galdieria*, *Cyanidioschyzon*, and *Cyanidum* which thrive in highly acidic and thermal habitats worldwide where few other organisms survive (e.g., fumaroles, hot springs, geysers, mining wastewaters, etc.). Here, they are the dominant photosynthetic organisms in these ecological niches and frequently constitute up to 90% of the total biomass and almost 100% of eukaryotic biomass [15].

Their ecological boundaries are defined by temperatures between 12°C – 56°C [16-19], highly acidic conditions (pH < 0 – 4.5) [20-22], increased concentrations of toxic heavy metals and other xenobiotics [23, 24] as well as up to 10% of salt [25]. However, specific Cyanidiales lineages such as the mesophilic *Cyanidium*, and in some cases *Galdieria phlegrea*, are also found in more temperate environments with neutral pH [21, 26-30] indicating that neither high acidity nor high temperature are obligate conditions for the prevalence of Cyanidiales at a given habitat [31].

Multiple independent descriptions of the Cyanidiales date back to the first half of the 19th century and misclassified them into different algal clades due to their uniform morphology and lack of distinguishing features. As a consequence, many studies were performed with mixed populations [32, 33]. 1933, Geitler named a species he described as *Cyanidium caldarium* and established the family Cyanidiaceae [34] which was correctly classified as a red alga (Rhodophyta) only in 1958 [35]. *Cyanidioschyzon merolae* was described as a Cyanidiales in 1978 [36]. This alga is smaller, divides by binary fission, is photoautotrophic and does not have a cell wall. Its genome was the first algal genome to be sequenced [37]. In 1981, *Galdieria sulphuraria* was separated from a *C. caldarium* culture based on the presence of linoleic acid [38] and its ability to grow heterotrophically on a broad array of carbon sources [39, 40]. Since *C. caldarium* and *G. sulphuraria* show no obvious phenotypic differences and live as mixed populations in the same environment, they are hard to differentiate. As a consequence, the majority of literature on *C. caldarium* published before 1981, such as publications referring to '*Cyanidium caldarium*' forma B [32] or '*Cyanidium caldarium*' M 8 [41], actually applies to *G. sulphuraria* [25, 42]. Today, six putative family-level taxa are acknowledged based on molecular haplotypes: *Galdieria sulphuraria*, *Galdieria phlegrea*, *Galdieria maxima*, the *Cyanidium* lineage, a mesophile *Cyanidium* lineage which has not been successfully isolated as laboratory culture so far, and the *Cyanidioschyzon* lineage [17, 20, 43]. New sampling sites in the APAC region indicate even greater biodiversity [16, 44]. The Cyanidiales have been used as model organisms for a shifting array of subjects, ranging from exploring the ecological boundaries of life and the species composition in extreme environments [45-48] to the study of photosynthetic components [49-51], photosynthetic regulation [52] and especially the role of the FtsZ ring during organelle division [53-55]. Today, their metabolic flexibility in combination with the polyextremophilic traits has lately risen the interest for industrial applications. Pilot studies are evaluating the bio-industrial potential of various Cyanidiales strains for urban wastewater treatment [56-58], bioremediation [59, 60] as food ingredient [61] and for industrial phycocyanin production [62, 63].

Concept 1: Eukaryotic evolution in the shadows of nuclear genome reduction

The sequenced genomes of *C. merolae 10D* [37, 64], *G. sulphuraria* 074W [65] and *G. phlegrea* DBV009 [66] are challenging the field of eukaryotic genome evolution with new concepts due to their complete or partial loss of and universal eukaryotic features. So far, eukaryotic genome evolution has been primarily attributed to gene/genome duplications, leading to the temporary amelioration of evolutionary constraints, which in turn promotes gene neofunctionalization, gene family expansion, but also gene loss, repeat decay, etc. [67, 68]. Whereas an increase in genome size over time is not positively correlated with organismal complexity [69], genome reduction has been tightly connected to gene loss, thus narrowing the ecological potential of free-living organisms by reducing their ability to innovate and adapt. The best examples for such scenarios would be symbiosis [70], parasitism [71], pathogenicity [72] and adaptation to novel habitats where a particular set of genes is no longer required for survival and lost due to neutral regressive evolution [73]. In this context, analysis of the Cyanidiales genomes revealed two phases of ancestral genome reduction along their evolutionary history. Their genomes are among the most streamlined and miniaturized genomes known today, with genome size ranging between 12 Mb – 16 Mb and a gene inventory limited to 4,400 – 6,500 genes. Consequently, multiple core eukaryotic traits are missing in *Galdieria*, *Cyanidioschyzon* and *Cyanidium*, such as flagella, basal bodies, the glycosyl-phosphatidylinositol anchor biosynthesis pathway, the autophagy regulation pathway and phytochrome based light sensing as well as a simplified cytoskeleton which is partially missing cytoskeletal motor proteins [74-76]. The ability of the Cyanidiales to live in such diverse environments contrasts the established canon of genome reduction and its consequences. This controversy is further boosted by similar findings in other free-living microorganisms such as *Picochlorum* [77, 78] and *Micromonas* [79]. The existence of free-living organisms with highly reduced genomes must point towards unknown evolutionary strategies for generating functional diversity to support independent lifestyles [80, 81]. Due to their species richness and diversity of habitats, the Cyanidiales offer the most remarkable model organism for studying the underlying mechanics of adaptation in free-living eukaryotes with reduced genome sizes.

One evolutionary mechanism to ameliorate the consequences of gene loss is through spliceosomal complexity. Eukaryotic genes are never fully translated to the final protein product. They contain pieces of the non-coding DNA located at the beginning (5' untranslated region), the end (3' untranslated region), or between (introns) the information-carrying regions (exons) which have to be processed first. The processing ("splicing") is performed by the spliceosomal machinery (SM). Introns are removed ("spliced") from the pre-mRNA and remaining exons are re-joined together to form the mature mRNA. However, not every exon is always fully included in the mature mRNA. Differential splicing describes the regulated process during which certain exons can be treated equally to introns (spliced from the mRNA) under one condition but kept under a different condition. As a consequence, one gene can lead to multiple proteins products (isoforms) which can differ in biological function. Genetic information can therefore be compacted and stored more efficiently in less space. At the same time, mutations can reprogram the inclusion/exclusion patterns of exons and introns during splicing. As a result, new isoforms are generated, which potentially provide novel functions and adaptive advantages [82]. The complexity of the SM is positively correlated to transcript diversity via alternative splicing [83]. However, the upkeep of a complex splicing machinery, e.g., ~200 SM proteins in the human genome [84], stands in clear contrast to the evolutionary trajectory of the Cyanidiales pushing for genome reduction and gene loss (especially since other core eukaryotic traits were lost). The sequenced Cyanidiales genomes, together with genomes of other red algae [75, 85, 86], suggest a varying degree of different evolutionary trajectories within this phylum regarding spliceosomal complexity [80, 87]. *Galdieria sulphuraria* 074W is the most intron-rich red alga (13,245 introns in 7174 transcripts) and maintained 149 of the >157 SM proteins that were present in the red algal ancestor. Also, it has continued increasing its intron number over time. In contrast, *Cyanidioschyzon merolae* 10D had the least number of introns (27 introns in 4803 transcripts) and lost the majority of its SM machinery (54 components left). RNA expression analysis in *Galdieria sulphuraria* 074W reported temperature dependent alternative splicing of more 1766 introns. Although not all red algae follow the same strategy, *Galdieria sulphuraria* 074W provides evidence for the adaptive advantages of an increased intron and SM burden in a scenario of natural selection towards streamlined genomes [87].

Concept 2: Horizontal Gene Transfer - an additional venue of adaptive evolution in eukaryotes?

One of the possible coping strategies is the acquisition of genes through horizontal gene transfer (HGT). HGT, equivalent to lateral gene transfer (LGT), is the inter- and intraspecific transmission of genes between a donor species and the acceptors of this gene. Organisms can thereby improve their genetics through the uptake and integration of foreign DNA. HGT in Bacteria [88-90] and Archaea [91] is widely accepted and recognized as an essential driver of evolution leading to the formation of pan-genomes [92, 93]. A pan-genome comprises all genes shared by any defined phylogenetic group of organisms as well as the genes unique to any single species in this group, where core genes often comprise central metabolic processes shared by all species, whereas genes present only in a subset of species are often associated with the origin of adaptive traits. This phenomenon is so pervasive in Bactria that it has been questioned whether prokaryotic genealogies can be reconstructed with any confidence using standard phylogenetic methods [94, 95]. Eukaryotes, on the other hand, were believed to transmit their nuclear and organellar genomes from one generation to the next in a vertical manner following classic Darwinian/Mendelian inheritance mechanics. In contrast to this notion, as the number of sequenced eukaryotic genomes sequencing has exponentially increased in the last decade, an increasing body of data has pointed towards the existence of HGT in these taxa as well, although at much lower rates than in prokaryotes [96]. However, the frequency and impact of eukaryotic HGT outside the context of endosymbiosis and pathogenicity remain one of the most hotly debated topics in evolutionary biology. Because the correct identification of HGT is rarely trivial and unambiguous, much space is left for interpretation and erroneous assignments. Since the first reports of eukaryotic HGT, skeptics have challenged its existence declaring eukaryotic HGT is Lamarckian, thus false, and merely a result from analysis artifacts [97, 98]. Proponents of eukaryotic HGT have failed to deliver explanations for the apparent absence of eukaryotic pan-genomes as well as the lack of cumulative effects which would be observed when genes derived from HGT increasingly diverge from their non-eukaryotic orthologs as a function of time.

In the context of HGT, the Cyanidiales became more broadly known after publication of the genome sequences of *Galdieria sulphuraria* 074W [65, 99] and *Galdieria phlegrea* DBV009 [66]. Comparative genomics postulated a “hot start” at the root of the red algae phylum, which triggered a billion-year lasting adaptation towards polyextremophily at the cost of genome reduction [100]. Hence, the ancestor of today’s Cyanidiales was likely to be a unicellular, proto-rhodophyte living in aquatic thermal environments. Whereas the Cyanidiales have maintained this evolutionary trajectory of extremophile adaptation for a period longer than one billion years, it was successively lost in the other red algal groups whose descendants are now largely mesophilic. The gene sets of *Galdieria sulphuraria* 074W [65] and *Galdieria phlegrea* DBV009 [66] revealed 73/63 instances of phylogenetic inferences other than the cyanobacterial plastid endosymbiont which were interpreted as gene gains from non-eukaryotic species via eukaryotic HGT. The majority of HGT genes was hypothesized to have provided selective advantages during the evolution towards polyextremophily strong enough to counteract the background selection pushing towards genome reduction and reduced gene inventory. In *Galdieria* genomes, the functional annotations of HGT genes were depicted as indicators of the selective forces that acted on the Cyanidiales. In this context, a dominant proportion of the inherited gene functions were connected to ecologically important traits for extremophile survival, such as heavy metal detoxification, xenobiotic detoxification, ROS scavenging, and metabolic functions related to carbon, fatty acid, and amino acid turnover. The authors postulated a relevant role in the systems biology of *Galdieria sulphuraria* (systems biology comprises the cellular phenotypes and response networks that emerge from interactions between individual system components). Hence, the success of the Cyanidiales is not only derived from constant adaptation towards polyextremophily with genome reduction as a possible evolutionary motor for major radiation but also "boosted" through HGT which expanded and optimized favorable traits [80, 99]. Expressional changes in HGT genes as a consequence of temperature stress were more pronounced in comparison to native genes in *Galdieria sulphuraria*, indicating their physiological relevance in the role of environmental stress adaptation [101, 102]. At the same time, significant differential expression of HGT genes was absent in *Cyanidioschyzon merolae* exposed to high CO_2, a condition which does not implicate stress [51, 102]. HGT genes are thus not only present as non-coding DNA sequence but also well integrated into the transcriptional machinery. Further, HGT genes identified differ significantly from the native genes in various genomic features (e.g., GC-content, number of exons per gene, etc.) and originated from extremophilic prokaryotes that thrive in similar habitats as the Cyanidiales [102] yet another indication for their adaptive potential. However, microevolution at decreased growth temperatures over a period spanning > 100 generations of *Galdieria sulphuraria* did not translate into a major selective force acting specifically on HGT genes. Rather, natural selection towards temperature tolerance is a systems biology problem which requires the gradual orchestration of an intricate gene network and deeply nested regulators at genetic and epigenetic level [103].

Analysis of multiple Cyanidiales genomes has significantly contributed to the ongoing debate regarding the existence, frequency, and impact of eukaryotic HGT [102]. The absence of a eukaryotic pangenome and cumulative effects can be explained by the rarity of eukaryotic HGT and the propensity of HGT candidates to gene erosion. Because eukaryotic HGT is the exception rather than the rule, its quantity in eukaryotic genomes does not need to increase as a function of time and likely reached equilibrium between acquisition and erosion in the distant past. The presumption that eukaryotic HGT will (and should) unfold in the eukaryotic kingdom in the same manner as does prokaryotic HGT among prokaryotes is highly questionable.

1. Darwin, C., *On the origin of species by means of natural selection, or preservation of favoured races in the struggle for life*. 1859: London : John Murray, 1859.
2. Dobzhansky, T., *Genetics and the origin of species*. 1937, New York: Columbia Univ. Press.
3. Laland, K., et al., *Does evolutionary theory need a rethink?* 2014. **514**(7521): p. 161.
4. Venter, J.C., et al., *The sequence of the human genome.* 2001. **291**(5507): p. 1304-1351.
5. O'Leary, N.A., et al., *Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation.* 2015. **44**(D1): p. D733-D745.
6. Guiry, M.D.J.J.o.p., *How many species of algae are there?* 2012. **48**(5): p. 1057-1063.
7. Brawley, S.H., et al., *Insights into the red algae and eukaryotic evolution from the genome of Porphyra umbilicalis (Bangiophyceae, Rhodophyta).* 2017. **114**(31): p. E6361-E6370.
8. Yoon, H.S., et al., *A molecular timeline for the origin of photosynthetic eukaryotes.* Mol Biol Evol, 2004. **21**(5): p. 809-18.

9. Parfrey, L.W., et al., *Estimating the timing of early eukaryotic diversification with multigene molecular clocks.* 2011. **108**(33): p. 13624-13629.
10. Archibald, John M., *Endosymbiosis and Eukaryotic Cell Evolution.* Current Biology, 2015. **25**(19): p. R911-R921.
11. Cavalier-Smith, T., *Principles of protein and lipid targeting in secondary symbiogenesis: euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree.* J Eukaryot Microbiol, 1999. **46**(4): p. 347-66.
12. Reeb, V.C., et al., *Interrelationships of chromalveolates within a broadly sampled tree of photosynthetic protists.* 2009. **53**(1): p. 202-211.
13. Reyes-Prieto, A., A.P. Weber, and D. Bhattacharya, *The origin and establishment of the plastid in algae and plants.* Annu Rev Genet, 2007. **41**: p. 147-68.
14. Falkowski, P.G. and J.A. Raven, *Aquatic photosynthesis*. 2013: Princeton University Press.
15. Doemel, W.N. and T. Brock, *The physiological ecology of Cyanidium caldarium.* Microbiology, 1971. **67**(1): p. 17-32.
16. Hsieh, C.J., et al., *Analysis of rbcL sequences reveals the global biodiversity, community structure, and biogeographical pattern of thermoacidophilic red algae (Cyanidiales).* 2015. **51**(4): p. 682-694.
17. Ciniglia, C., et al., *Hidden biodiversity of the extremophilic Cyanidiales red algae.* Mol Ecol, 2004. **13**(7): p. 1827-38.
18. Toplin, J., et al., *Biogeographic and phylogenetic diversity of thermoacidophilic cyanidiales in Yellowstone National Park, Japan, and New Zealand.* 2008. **74**(9): p. 2822-2833.
19. Barcytė, D., et al., *Burning coal spoil heaps as a new habitat for the extremophilic red alga Galdieria sulphuraria*. Vol. 18. 2017. 19-29.
20. Ciniglia, C., et al., *Cyanidiophyceae in Iceland: plastid rbc L gene elucidates origin and dispersal of extremophilic Galdieria sulphuraria and G. maxima (Galdieriaceae, Rhodophyta).* 2014. **53**(6): p. 542-551.
21. Iovinella, M., et al., *Cryptic dispersal of Cyanidiophytina (Rhodophyta) in non-acidic environments from Turkey.* Extremophiles, 2018: p. 1-11.
22. W. Gross, S.G., *Physiological characterization of the red alga Galdieria sulphuraria isolated from a mining area.* Nova Hedwigia Beihefte, 2001. **123**: p. 523 -530.
23. Albertano, P., G. Pinto, and R. Taddei, *Evaluation of toxic effects of heavy metals on unicellular algae. II. Growth curves with different concentrations of heavy metals.* Delpinoa, 1980. **21**: p. 23-34.
24. Castenholz, R.W. and T.R. McDermott, *The Cyanidiales: ecology, biodiversity, and biogeography*, in *Red Algae in the Genomic Age*. 2010, Springer. p. 357-371.
25. Albertano, P., et al., *The taxonomic position of Cyanidium, Cyanidioschyzon and Galdieria: an update.* 2000. **433**(1-3): p. 137-143.
26. Azua-Bustos, A., et al., *Ancient photosynthetic eukaryote biofilms in an Atacama Desert coastal cave.* Microb Ecol, 2009. **58**(3): p. 485-96.
27. Del Rosal, Y., et al., *Cyanidium sp. colonizadora de cuevas turísticas.* 2015.
28. Friedmann, I.J.I.J.o.S., *Progress in the biological exploration of caves and subterranean waters in Israel.* 1964. **1**(1): p. 5.
29. Skuja, H., *Alghe cavernicole nelle zone illuminate delle Grotte di Castellana.* Le Grotte d'Italia, 1970. **4**: p. 193 - 202.
30. Leclerc, J., A. Couté, and P.J.C.a. Dupuy, *climat annuel de deux grottes et d'une englise du Poitou, ou vivent des colonies pures d'algues sciaphiles.* 1983.

31. Gross, W., *Revision of comparative traits for the acido-and thermophilic red algae Cyanidium and Galdieria*, in *Enigmatic Microorganisms and life in Extreme Environments*. 1999, Springer. p. 437-446.
32. De Luca, P. and R. Taddei, *On the necessity of a systematic revision of the thermal acidophilic alga Cyanidium caldarium Tilden Geitler.* Webbia, 1976. **30**: p. 197-218.
33. Reeb, V. and D. Bhattacharya, *The thermo-acidophilic cyanidiophyceae (Cyanidiales)*, in *Red algae in the genomic age*. 2010, Springer. p. 409-426.
34. Geitler, L., *Diagnosen neuer Blaualgen von den Sunda-Inseln.* Archiv für Hydrobiologie, 1933. **Suppl. 12**: p. 622- 634.
35. Hirose, H., *Rearrangement of the systematic position of a thermal alga, Cyanidium caldarium.* Botanical Magazine, 1958. **71**: p. 347-352.
36. De Luca, P., R. Taddei, and L. Varano, *Cyanidioschyzon merolae, a new alga of thermal acidic environments.* Webbia, 1978. **33**(1): p. 37-44.
37. Matsuzaki, M., et al., *Genome sequence of the ultrasmall unicellular red alga Cyanidioschyzon merolae 10D.* Nature, 2004. **428**(6983): p. 653-7.
38. Boenzi, D., P. De Luca, and R. Taddei, *Fatty acids in "Cyanidium".* Giornale Botanico Italiano, 1977. **111**: p. 129-134.
39. Merola, A., et al., *Revision of Cyanidium caldarium. Three species of acidophilic algae.* Giornale Botanico Italiano, 1981. **115**: p. 189-195.
40. Gross, W. and C. Schnarrenberger, *Heterotrophic growth of 2 strains of the acido-thermophilic red alga Galdieria sulphuraria.* Plant and Cell Physiology, 1995. **36**(4): p. 633-638.
41. Nagashima, H. and I. Fukuda, *Morphological properties of Cyanidiurn caldarium and related algae in Japan.* Japanese Journal of Phycology, 1981. **29**: p. 237-242.
42. Seckbach, J., *Systematic problems with Cyanidium caldarium and Galdieria sulphuraria and their implications for molecular biology studies.* Journal of Phycology, 1991. **27**(6): p. 794-796.
43. Yoon, H.S., et al., *Establishment of endolithic populations of extremophilic Cyanidiales (Rhodophyta).* 2006. **6**(1): p. 78.
44. Hsieh, C.J., et al., *The effects of contemporary selection and dispersal limitation on the community assembly of acidophilic microalgae.* Journal of phycology, 2018. **54**(5): p. 720-733.
45. Doemel, W.N. and T.D. Brock, *Upper temperature limit of Cyanidium caldarium.* Archiv für Mikrobiologie, 1970. **72**(4): p. 326-&.
46. Tansey, M.R. and T.D. Brock, *Upper temperature limit for eukaryotic organisms.* Proceedings of the National Academy of Sciences of the United States of America, 1972. **69**(9): p. 2426-&.
47. Rothschild, L.J. and R.L.J.N. Mancinelli, *Life in extreme environments.* 2001. **409**(6823): p. 1092.
48. Misumi, O., et al., *Cytological Studies of Metal Ion Tolerance in the Red Algae Cyanidioschyzon merolae*. Vol. 73. 2008. 437-443.
49. Sørensen, L., et al., *Purification of the photosynthetic pigment C-phycocyanin from heterotrophic Galdieria sulphuraria.* 2013. **93**(12): p. 2933-2938.
50. Okumura, A., et al., *Aromatic structure of Tyrosine-92 in the extrinsic PsbU protein of red algal Photosystem II is important for its functioning.* 2007. **581**(27): p. 5255-5258.
51. Rademacher, N., et al., *Photorespiratory glycolate oxidase is essential for the survival of the red alga Cyanidioschyzon merolae under ambient CO2 conditions.* J Exp Bot, 2016. **67**(10): p. 3165-75.

52. Kawase, Y., S. Imamura, and K.J.F.I. Tanaka, *A MYB-type transcription factor, MYB2, represses light-harvesting protein genes in Cyanidioschyzon merolae.* 2017. **591**(16): p. 2439-2448.
53. Takahara, M., et al., *Isolation, characterization, and chromosomal mapping of an ftsZ gene from the unicellular primitive red alga Cyanidium caldarium RK-1.* 2000. **37**(2): p. 143-151.
54. Miyagishima, S.-y., M. Takahara, and T.J.T.P.C. Kuroiwa, *Novel filaments 5 nm in diameter constitute the cytosolic ring of the plastid division apparatus.* 2001. **13**(3): p. 707-721.
55. Kuroiwa, T., et al., *The division apparatus of plastids and mitochondria*, in *International review of cytology*. 1998, Elsevier. p. 1-41.
56. Henkanatte-Gedera, S., et al., *Removal of dissolved organic carbon and nutrients from urban wastewaters by Galdieria sulphuraria: Laboratory to field scale demonstration.* 2017. **24**: p. 450-456.
57. Selvaratnem, T., et al., *Feasibility of Algal Systems for Sustainable Wastewater Treatment*. 2014.
58. Selvaratnam, T., et al., *Algal biofuels from urban wastewaters: Maximizing biomass yield using nutrients recycled from hydrothermal processing of biomass.* 2015. **182**: p. 232-238.
59. Fukuda, S.-y., et al., *Cellular accumulation of cesium in the unicellular red alga Galdieria sulphuraria under mixotrophic conditions.* 2018: p. 1-5.
60. Barragan, M.H. and S.W. Harcum, *The feasibility of using Cyanidium Caldarium to bioremediate copper from acid mine drainage.*
61. Graziani, G., et al., *Microalgae as human food: chemical and nutritional characteristics of the thermo-acidophilic microalga Galdieria sulphuraria.* 2013. **4**(1): p. 144-152.
62. Sloth, J.K., et al., *Accumulation of phycocyanin in heterotrophic and mixotrophic cultures of the acidophilic red alga Galdieria sulphuraria.* 2006. **38**(1-2): p. 168-175.
63. Sloth, J.K., et al., *Growth and phycocyanin synthesis in the heterotrophic microalga Galdieria sulphuraria on substrates made of food waste from restaurants and bakeries.* 2017. **238**: p. 296-305.
64. Nozaki, H., et al., *A 100%-complete sequence reveals unusually simple genomic features in the hot-spring red alga Cyanidioschyzon merolae.* BMC Biol, 2007. **5**: p. 28.
65. Schonknecht, G., et al., *Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote.* Science, 2013. **339**(6124): p. 1207-10.
66. Qiu, H., et al., *Adaptation through horizontal gene transfer in the cryptoendolithic red alga Galdieria phlegrea.* Curr Biol, 2013. **23**(19): p. R865-6.
67. Innan, H. and F. Kondrashov, *The evolution of gene duplications: classifying and distinguishing between models.* Nature Reviews Genetics, 2010. **11**(2): p. 97-108.
68. Lefébure, T., et al., *Less effective selection leads to larger genomes.* 2017: p. gr. 212589.116.
69. Eddy, S.R.J.C.b., *The C-value paradox, junk DNA and ENCODE.* 2012. **22**(21): p. R898-R899.
70. McCutcheon, J.P. and N.A.J.N.R.M. Moran, *Extreme genome reduction in symbiotic bacteria.* 2012. **10**(1): p. 13.
71. Spanu, P.D., et al., *Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism.* 2010. **330**(6010): p. 1543-1546.

72. Peyretaillade, E., et al., *Extreme reduction and compaction of microsporidian genomes.* 2011. **162**(6): p. 598-606.
73. Albalat, R. and C.J.N.R.G. Cañestro, *Evolution by gene loss.* 2016. **17**(7): p. 379.
74. Qiu, H., H.S. Yoon, and D. Bhattacharya, *Red Algal Phylogenomics Provides a Robust Framework for Inferring Evolution of Key Metabolic Pathways.* PLoS Curr, 2016. **8**.
75. Collén, J., et al., *Genome structure and metabolic features in the red seaweed Chondrus crispus shed light on evolution of the Archaeplastida.* 2013. **110**(13): p. 5247-5252.
76. Brawley, S.H., et al., *Insights into the red algae and eukaryotic evolution from the genome of Porphyra umbilicalis (Bangiophyceae, Rhodophyta).* Proc Natl Acad Sci U S A, 2017. **114**(31): p. E6361-E6370.
77. Foflonker, F., et al., *Genomic Analysis of Picochlorum Species Reveals How Microalgae May Adapt to Variable Environments.* Molecular Biology and Evolution, 2018.
78. Foflonker, F., et al., *Genome of the halotolerant green alga P icochlorum sp. reveals strategies for thriving under fluctuating environmental conditions.* 2015. **17**(2): p. 412-426.
79. Worden, A.Z., et al., *Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes Micromonas.* Science, 2009. **324**(5924): p. 268-72.
80. Bhattacharya, D., et al., *When Less is More: Red Algae as Models for Studying Gene Loss and Genome Evolution in Eukaryotes.* Critical Reviews in Plant Sciences, 2018. **37**(1): p. 81-99.
81. Olson, M.V.J.T.A.J.o.H.G., *When less is more: gene loss as an engine of evolutionary change.* 1999. **64**(1): p. 18-23.
82. Black, D.L.J.A.r.o.b., *Mechanisms of alternative pre-messenger RNA splicing.* 2003. **72**(1): p. 291-336.
83. Matlin, A.J., F. Clark, and C.W.J.N.r.M.c.b. Smith, *Understanding alternative splicing: towards a cellular code.* 2005. **6**(5): p. 386.
84. Hegele, A., et al., *Dynamic protein-protein interaction wiring of the human spliceosome.* 2012. **45**(4): p. 567-580.
85. Nakamura, Y., et al., *The first symbiont-free genome sequence of marine red alga, Susabi-nori (Pyropia yezoensis).* 2013. **8**(3): p. e57122.
86. Bhattacharya, D., et al., *Genome of the red alga Porphyridium purpureum.* Nat Commun, 2013. **4**: p. 1941.
87. Qiu, H., et al., *Unexpected conservation of the RNA splicing apparatus in the highly streamlined genome of Galdieria sulphuraria.* BMC Evol Biol, 2018. **18**(1): p. 41.
88. Doolittle, W.F., *Lateral genomics.* Trends Cell Biol, 1999. **9**(12): p. M5-8.
89. Ochman, H., J.G. Lawrence, and E.A. Groisman, *Lateral gene transfer and the nature of bacterial innovation.* Nature, 2000. **405**(6784): p. 299-304.
90. Boucher, Y., et al., *Lateral gene transfer and the origins of prokaryotic groups.* Annu Rev Genet, 2003. **37**: p. 283-328.
91. Nelson-Sathi, S., et al., *Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea.* Proc Natl Acad Sci U S A, 2012. **109**(50): p. 20537-42.
92. Tettelin, H., et al., *Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome".* Proc Natl Acad Sci U S A, 2005. **102**(39): p. 13950-5.

93. Vernikos, G., et al., *Ten years of pan-genome analyses.* Curr Opin Microbiol, 2015. **23**: p. 148-54.

94. Philippe, H. and C.J. Douady, *Horizontal gene transfer and phylogenetics.* Curr Opin Microbiol, 2003. **6**(5): p. 498-505.

95. Doolittle, W.F. and T.D. Brunet, *What Is the Tree of Life?* PLoS Genet, 2016. **12**(4): p. e1005912.

96. Danchin, E.G., *Lateral gene transfer in eukaryotes: tip of the iceberg or of the ice cube?* BMC Biol, 2016. **14**(1): p. 101.

97. Martin, W.F., *Eukaryote lateral gene transfer is Lamarckian.* Nat Ecol Evol, 2018. **2**(5): p. 754.

98. Martin, W.F., *Too Much Eukaryote LGT.* Bioessays, 2017. **39**(12).

99. Schönknecht, G., A.P. Weber, and M.J. Lercher, *Horizontal gene acquisitions by eukaryotes as drivers of adaptive evolution.* Bioessays, 2014. **36**(1): p. 9-20.

100. Qiu, H., et al., *Evidence of ancient genome reduction in red algae (Rhodophyta).* J Phycol, 2015. **51**(4): p. 624-36.

101. Rossoni, A.W., et al., *Cold Acclimation of the Thermoacidophilic Red Alga Galdieria sulphuraria - Changes in Gene Expression and Involvement of Horizontally Acquired Genes.* Plant and Cell Physiology, 2018: p. pcy240-pcy240.

102. Rossoni, A.W., et al., *The genomes of polyextremophilic Cyanidiales contain 1% horizontally transferred genes with diverse adaptive functions.* eLife, 2019. **8**: p. e45017.

103. Rossoni, A.W. and A.P.M. Weber, *Systems Biology of Cold Adaptation in the Polyextremophilic Red Alga Galdieria sulphuraria.* 2019. **10**(927).

III. Manuscripts

Manuscript 1

Unexpected conservation of the RNA splicing apparatus in the highly streamlined genome of Galdieria sulphuraria

Qiu *et al. BMC Evolutionary Biology* (2018) 18:41
https://doi.org/10.1186/s12862-018-1161-x

BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

Unexpected conservation of the RNA splicing apparatus in the highly streamlined genome of *Galdieria sulphuraria*

Huan Qiu[1], Alessandro W. Rossoni[2], Andreas P. M. Weber[2], Hwan Su Yoon[3] and Debashish Bhattacharya[1,4*]

Abstract

Background: Genome reduction in intracellular pathogens and endosymbionts is usually compensated by reliance on the host for energy and nutrients. Free-living taxa with reduced genomes must however evolve strategies for generating functional diversity to support their independent lifestyles. An emerging model for the latter case is the Rhodophyta (red algae) that comprises an ecologically widely distributed, species-rich phylum. Red algae have undergone multiple phases of significant genome reduction, including extremophilic unicellular taxa with limited nuclear gene inventories that must cope with hot, highly acidic environments.

Results: Using genomic data from eight red algal lineages, we identified 155 spliceosomal machinery (SM)-associated genes that were putatively present in the red algal common ancestor. This core SM gene set is most highly conserved in *Galdieria* species (150 SM genes) and underwent differing levels of gene loss in other examined red algae (53-145 SM genes). Surprisingly, the high SM conservation in *Galdieria sulphuraria* coincides with the enrichment of spliceosomal introns in this species (2 introns/gene) in comparison to other red algae (< 0.34 introns/gene). Spliceosomal introns in *G. sulphuraria* undergo alternatively splicing, including many that are differentially spliced upon changes in culture temperature.

Conclusions: Our work reveals the unique nature of *G. sulphuraria* among red algae with respect to the conservation of the spliceosomal machinery and introns. We discuss the possible implications of these findings in the highly streamlined genome of this free-living eukaryote.

Keywords: Genome reduction, RNA splicing, Intron, Rhodophyta

Background

The study of eukaryote genome evolution has focused primarily on how genomes grow in size and complexity over time (e.g., via genome duplication [1] and transposable element accumulation [2] often due to neutral, population level processes) in model organisms such as vertebrates and land plants. In contrast, there is limited information arising from the opposite perspective (i.e., genome reduction), despite its prevalence in many lineages [3]. In addition, knowledge about genome reduction, which has been studied primarily in highly specialized endosymbionts and pathogens [4] has limited implications for free-living species and the maintenance of their biodiversity. Therefore, understanding the impact of genome reduction in free-living organisms, particularly in eukaryotes that have complex genomes, provides a novel avenue to understand and test the underlying principles of genome evolution.

An emerging model for elucidating the impacts of genome reduction in free-living eukaryotes is the Rhodophyta (red algae). This monophyletic algal lineage comprises an ecologically widely distributed and species-rich phylum (ca. 7000 species) [5]. Analysis of genomic

* Correspondence: d.bhattacharya@rutgers.edu
[1] Department of Ecology, Evolution and Natural Resources, Rutgers University, New Brunswick, NJ 08901, USA
[4] Department of Biochemistry and Microbiology, Rutgers University, New Brunswick, NJ 08901, USA
Full list of author information is available at the end of the article

and transcriptomic data have shown that red algae underwent at least two phases of massive genome reduction [6]. The first is in the stem lineage, where about one-quarter of the gene inventory was shed [6] and the second is in the ancestor of the anciently diverged extremophiles, Cyanidiophytina, such as *Cyanidioschyzon merolae* [7] and *Galdieria sulphuraria* [8], that thrive in volcanic hot-spring areas [6, 9]. As a consequence of adaptation to their unusual environment, *G. sulphuraria* (6.5 K nuclear genes) and *C. merolae* (4.7 K nuclear genes) contain smaller gene inventories than their mesophilic red algal sisters which encode ~ 10 K nuclear genes [10–12].

Alternative splicing provides a major avenue of post-transcriptional regulation in eukaryotes [13]. Here, using analysis of genomic and RNA-seq data from *G. sulphuraria*, we show: 1) selective retention of the spliceosomal machinery (SM) in *G. sulphuraria*, a toolkit that has been greatly reduced in complexity in many of its sister red algal lineages, and 2) the coincidence of high SM retention and intron enrichment in *G. sulphuraria* that has resulted in extensive alternative splicing (AS) in this species. Given these unique features in *G. sulphuraria*, we discuss the possible implications of AS in red algal evolution.

Results

Pattern of spliceosome machinery gene loss in red algae

Using a BLASTp search-based method (see Methods) with 215 non-redundant human SM-associated proteins [14] as the query, we identified homologs in red algae and their putative sister lineage, the Viridiplantae (Additional file 1: Table S1). Consistent with the fundamental function of the SM, a majority of these proteins have detectable homologs in red algae and Viridiplantae (Fig. 1a), with generally more genes found in the latter phylum (Fig. 1a). Substantial variation in SM gene number was found among red algal lineages with *Galdieria* species (*G. sulphuraria* and *G. phlegrea*) containing the largest number of genes and *C. merolae* the smallest (Fig. 1a); the latter result has previously been described [15]. The observed SM gene distribution among red algal species could have resulted from independent, recent gene losses in multiple lineages or from extensive gene acquisition via horizontal gene transfer (HGT; e.g., in *G. sulphuraria* [8]). To distinguish between these two scenarios, we used phylogenetics to study the origin of red algal SM genes (see Methods) and estimated the timing of SM gene losses using a robust red algal tree of life [16]. Most individual SM gene phylogenies suggest vertical transmission because of the shared common ancestry of red algae with a variety of other eukaryotes (e.g., Metazoa in Additional file 2: Figure S1A; see Additional file 3 for all of the phylogenies). No clear evidence was found for the HGT of SM genes in *Galdieria* and other red algal species (Additional file 3). Using Dollo parsimony [17], we reconstructed the evolutionary history of SM genes in red algae. A total of 155 SM associated genes was likely present in the stem lineage of red algae, most of which (150) are preserved in *Galdieria* species (Fig. 1b). In contrast, extensive SM gene losses occurred independently in other red algal lineages such as *C. merolae* (currently 53 SM genes), Bangiophyceae (*Porphyra yezoensis* + *Porphyra umbilicalis*, 91 SM genes), and *Porphyridium* species (*P. purpureum* and *P. aerugineum*, 110 SM genes) (Fig. 1a). *Rhodosorus marinus* (145) contains a SM gene number similar to that in *Galdieria* species (Fig. 1b). Using 303 highly conserved gene families in eukaryotes as reference, we assessed the completeness of each red algal protein dataset with BUSCO. Most species showed a high coverage (< 8% missing genes), except *Chondrus crispus* (16% missing) and Bangiophyceae (19% missing) (Additional file 4: Table S2). Because *C. crispus* contains slightly more SM-genes than its sister lineage

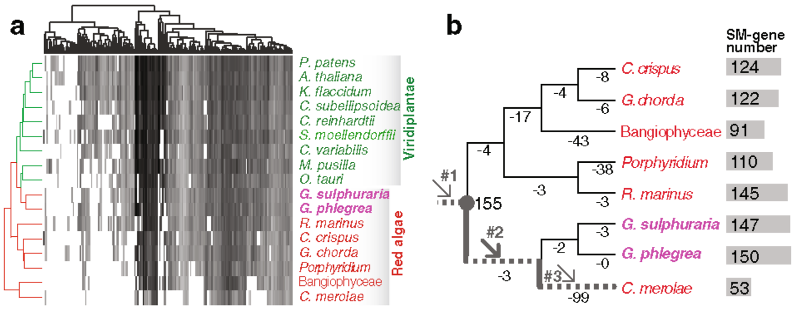


Fig. 1 The evolution of SM genes in red algae. **a** Distribution of 215 spliceosomal (SM) protein homologs in 17 red algal (in red color) and Viridiplantae lineages (in green color). *Galdieria* species are shown in the purple color. The heatmap is based on the sequence identities between human SM proteins and their homologs in the corresponding species. **b** The putative losses of SM genes in the red algal tree of life. The extent of SM gene loss is indicated by the numbers under branches. The grey circle indicates the most recent common ancestor of red algae. The dashed branches and arrows indicate the three phases of red algal genome reduction

Gracilariopsis chorda, this result suggests that the estimate of SM-gene number in most species was robust except for the Bangiophyceae. Whereas the extensive SM gene loss in *C. merolae* is likely explained by recent genome reduction specific to this extremophilic lineage [9] (arrow #3 in Fig. 1b), the underlying reasons for SM gene loss in *Porphyridium* and other mesophilic species are unclear. In contrast, the significant retention of SM genes at the split of extremophilic and mesophilic red algae (and maintenance in *Galdieria*) in the face of genome reduction, specific to this group (arrow #2 in Fig. 1b) is a surprising result.

Spliceosome composition and conservation in red algae

Based on human SM protein expression data [14] (Additional file 5: Table S3), we found that highly expressed SM proteins (79%) are twice as likely to be retained in red algae as those expressed at low levels (43%). Nearly all of the core proteins that directly bind small nucleolar (sn) RNAs to form small nuclear ribonucleoproteins (i.e., Sm, U1, U2, U5, U4/U6 and U5/U4/U6) are conserved in *Galdieria* species and have the lowest degree of loss in mesophilic red algae (Additional file 5: Table S3). All five snRNAs (U1-2 and U4-6) are found in *G. sulphuraria* (Additional file 6: Figure S2) and in five other red algal genomes (Additional file 7: Table S4), except for U1 snRNA that was most likely lost in *C. merolae* [15] (Additional file 7: Table S4). Whereas we included *G. phlegrea* to fully capture the SM gene inventory (Fig. 1), this taxon is not included in downstream analyses because of its close phylogenetic relationship to *G. sulphuraria* and the relatively low quality of intron annotation due to a lack of transcriptome data. In contrast, the remaining auxiliary SM proteins that generally perform peripheral or modulatory functions underwent more frequent loss (Additional file 5: Table S3). These results suggest that the red algal common ancestor contained 155 SM proteins that comprised the complete core of the spliceosomal machinery that was largely maintained in some lineages (such as *Galdieria*). It is noteworthy that although *C. merolae* and *G. sulphuraria* are both extremophiles that inhabit areas surrounding volcanic hot springs, these two species differ dramatically in lifestyle and metabolic capacity [18] which presumably is reflected by the additional phase of genome reduction in *C. merolae* (arrow #3 in Fig. 1b) [9]. Whereas the extremely reduced SM does perform RNA-splicing functions in *C. merolae*, it likely has a highly compromised efficiency given the minimal number of introns (only 27) present in this genome [7]. A highly reduced SM has also been found in several parasitic eukaryotes [19]. These species invariably show a paucity of introns including the kinetoplastid *Trypanosoma brucei* (13 introns in 8747 genes) [20], the microsporidian *Encephalitozoon cuniculi* (7 introns in 1996 genes) [21], and the diplomonad *Giardia lamblia* (6 introns in 7364 genes) [22].

Consistent with a previous study [15], we found that many red algal SM proteins are distantly related to reference sequences (i.e., human SM proteins) and have extremely long branches in phylogenies (Additional file 8: Figure S3). This pattern of evolution is common in red algal species with reduced SM gene sets such as *C. merolae*, Bangiophyceae, and *Porphyridium*, and is largely absent in *G. sulphuraria* and *R. marinus*. Using the BLASTp bit score as the metric, we found *G. sulphuraria* SM proteins to be generally more conserved than their orthologs in other red algal species (except *R. marinus*) at the primary sequence level (Additional file 8: Figure S3). Similarly, *G. sulphuraria* also shows the strongest overall sequence conservation among snRNAs, as reflected by their high alignment scores (Additional file 7: Table S4). *C. merolae* has the least conserved snRNAs (Additional file 7: Table S4). Whereas the fast evolution of SM proteins and snRNAs might reflect a genome-wide feature in *C. merolae* [16], the highly derived SM proteins and snRNAs in other red algal species likely resulted from the acquisition of novel functions or relaxed functional constraints. This result suggests that some apparent cases of gene loss in SM gene-poor red algal species (e.g., *C. merolae*) might instead be explained by high divergence; i.e., beyond sequence similarity-based recognition. In summary, our results demonstrate the conservation of the *G. sulphuraria* SM with respect to both gene inventory and protein similarity.

Enrichment of introns in the *G. sulphuraria* genome

Among the six red algal species with completed or draft genomes, intron numbers vary substantially, ranging from 27 in *C. merolae*, 245 in *P. purpureum*, to 13,245 in *G. sulphuraria* (Additional file 9: Table S5). The most highly conserved red algal SM in *G. sulphuraria* (Fig. 1) coincides with the enrichment of introns and multiple-exon genes in this species (Fig. 2a). Conversely, *C. merolae* that has the most reduced SM (Fig. 1) possesses the smallest number of introns (Fig. 2a and Additional file 9: Table S5). On average, *G. sulphuraria* genes are interrupted by two introns, whereas the corresponding numbers in other four red algal genomes are markedly smaller (0.005 – 0.3 intron/gene, Fig. 2a). The number of genes with one or more introns in *G. sulphuraria* greatly exceeds that in the other five studied red algal species (Fig. 2a). Whereas intron number is likely underestimated in the *P. yezoensis* genome because of its highly fragmented assembly [12], our conclusion does not change when the intron estimate is derived from a set of 'complete' *P. yezoensis* genes (i.e., 60% single-exon gene and 0.7 intron/gene on average [12]). Although in need of validation with additional genome data, these results suggest that the extent of SM conservation is likely associated with intron density in red algal genomes (Additional file 10: Figure S4A). A high number of auxiliary SM genes in *G. sulphuraria* likely

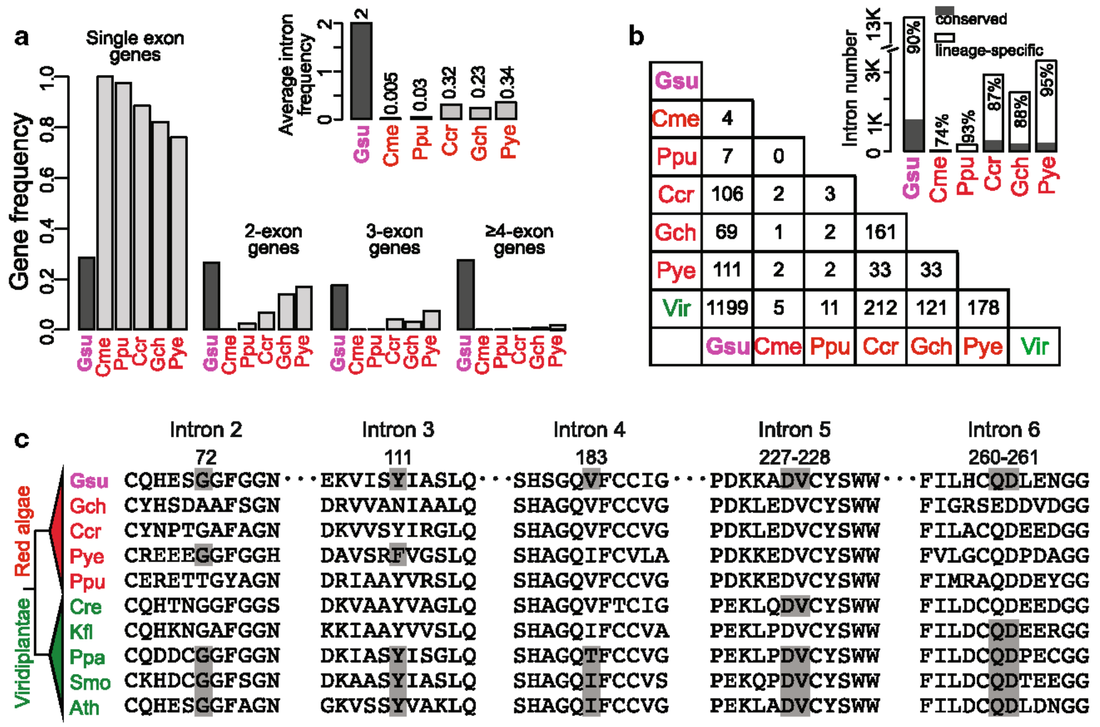


Fig. 2 Evolution of red algal introns. **a** The frequencies of single-exon genes and multi-exon genes in six red algal nuclear genomes. The bars in dark grey indicate *G. sulphuraria*. Abbreviations: Gsu (*G. sulphuraria*), Cme (*C. merolae*), Ppu (*P. purpureum*), Ccr (*C. crispus*), Gch (*G. chorda*), Pye (*P. yezoensis*) and Vir (Viridiplantae). The color scheme is the same as in Fig. 1. **b** Shared intron positions across algal lineages. The frequencies of lineage-specific intron for each red algal lineage are shown. **c** Conserved intron positions in *G. sulphuraria* geranylgeranyl transferase beta-subunit gene. The residues interrupted by intron (phase 2 and 3) or flanking intron (phase 1) are shadowed. The alignment is shown only for the 11-12 residue segments spanning the conserved intron positions. The positions of amino acids refer to those in *G. sulphuraria* proteins. Introns found only in Viridiplantae are not shown. Full alignment is provided in Additional file 11: Figure S5

results in an efficient SM that is able to process the relatively large number of introns in this species. Notably, *G. sulphuraria* has an exceptionally low GC content among red algae (Additional file 10: Figure S4B). Additional red algal genomic data are required to test the correlation between GC content and intron density in these taxa (Additional file 10: Figure S4C).

Origin of *G. sulphuraria* introns

To study the origin of *G. sulphuraria* spliceosomal introns, we compared their positions within homologous genes across six red algal species (Fig. 2b) and between them and five Viridiplantae lineages (Additional file 9: Table S5). Most of the *G. sulphuraria* intron positions (90%) appear to be lineage-specific (Fig. 2b), likely resulting from recent intron insertions. Based on a self-BLASTn search (*e*-value cutoff = 1e-5), only 3.6% (478/13,245) of introns share sequence similarity (query coverage ≥0.5) with one or more (up to 9) other introns. This result does not support the idea that the majority of *G. sulphuraria* introns resulted from recent intron duplications. Regarding intron positions that are shared with Viridiplantae, *G. sulphuraria* (1199) contains > 4-fold more ancestral introns than do other red algal lineages, such as *C. crispus* (212) and *G. chorda* (121) (Fig. 2b). This result suggests that many anciently derived introns were retained in *G. sulphuraria* and lost in other red algal lineages. Examples include the intron-rich *G. sulphuraria* gene encoding geranylgeranyl transferase beta-subunit (NCBI GeneID: 17088310). This gene contains six introns, of which five (from the 2nd to the 6th) have conserved positions in Viridiplantae homologs (Fig. 2c and Additional file 11: Figure S5). In contrast, all of these introns underwent losses in mesophilic red algae, resulting in a 3-exon gene in *P. yezoensis*, single-exon genes in *P. purpureum*, *C. crispus*, and *G. chorda* (Fig. 2c and Additional file 11: Figure S5). When assuming a simple evolutionary scenario (i.e., Dollo parsimony [17]), about 1700 introns are estimated to have been present in the red algal stem lineage, followed by significant losses in mesophilic red algae and in *C. merolae* (Additional file 12: Figure

S6). These results suggest that red algal introns have a high turnover rate, that is a common feature of many eukaryotes [23]. The relatively large number of introns in *G. sulphuraria* resulted from both lineage-specific intron gains and retention of ancestral introns. Notably, *G. sulphuraria* introns are much smaller in size (50 bp, on average) than in other red algal genomes (Additional file 13: Figure S7A). This is consistent with a strong size constraint that has resulted in the compact genome of *G. sulphuraria* (Additional file 13: Figure S7B) [8].

Alternative mRNA splicing in *G. sulphuraria*

Why would SM genes and a relatively more complex intron-exon structure be preserved in the compact *G. sulphuraria* genome? The answer to this question may lie in the fact that intron-exon structure provides the foundation in eukaryotes for generating multiple transcripts via alternative splicing. This is true in *G. sulphuraria*, as demonstrated by previous analysis of transcriptome data derived from Sanger sequences and 454 long-reads that revealed alternatively spliced isoforms for about 500 genes [8]. To test if AS in *G. sulphuraria* responds to environmental changes, we generated and analyzed extensive RNA-seq data from this alga under two arbitrary different temperature conditions: 'heat' (42 °C and 46 °C; non-stressed, because this alga normally lives at temperatures between 35 and 56 °C [24]) and 'cold' (28 °C; stressed) (see Methods). A total of 1766 introns were identified as being alternatively spliced (mostly via intron retention) under one or both temperature conditions (Additional file 14: Table S6). A total of 1397 of these alternatively spliced introns were located within 1027 known *G. sulphuraria* genes, including 12 genes derived via HGT (Additional file 14: Table S6). Among these 1766 introns, 1152 are identical with the annotated *G. sulphuraria* introns, accounting for 10.2% of the latter (13,245 introns). We predicted the impact of retention of these 1152 introns in the encoded transcripts and found that 792 (68. 7%) lead to frame-shifts (i.e., in lengths not divisible by 3) (Fig. 3a). When translated in the reading frame of preceding exons; i.e., 875 (75.9%) introns encode premature stop codons that lead to truncated proteins. Only 50 (4.3%) of retained introns do not cause these two types of changes in the inferred proteins. An example of intron retention is provided by the phosphoribosylformylglycinamidine cyclo-ligase gene (NCBI Gene ID: 17089374). The maintenance of its second intron introduces a stop codon (TAG) and leads to a truncated protein with a fragmented AIR synthase-like C-terminal domain (Fig. 3b). In a FDA synthase gene that contains two introns (NCBI Gene ID: 17086779), the retention of the first intron introduces a 1-bp frame-shift resulting in a ~ 150 amino acid novel peptide downstream of the protein encoded by the first exon (Fig. 3b). These two examples are supported by long reads generated by Sanger or 454 sequencing [8]. The functional implication of these protein variants is not yet known.

Differential intron splicing in *G. sulphuraria*

To test if AS in *G. sulphuraria* responds to temperature fluctuations, we searched and identified 212 introns that were (statistically significantly) differentially spliced between the heat and cold conditions (Additional file 15: Table S7). One example is a novel *Galdieria*-specific gene that is comprised of three exons (Fig. 3c). The splicing of the first intron is largely restricted to the heat condition, whereas the second intron is spliced in both conditions with apparently more extensive intron retention under heat than under cold (Fig. 3c). A second example is the gene encoding the RNA polymerase primary sigma factor (NCBI Gene ID: 17087802), with the first intron being significantly retained (i.e., 1/4~ 1/3 of total transcripts) under the heat condition. The same intron is rarely retained under the cold condition (Fig. 3c). This intron is located within a Sigma70-r3 domain (pfam04539) that is involved in the binding of core RNA polymerase. Retention of this intron leads to stop codons and a truncated Sigma70-r3 domain (Additional file 16: Figure S8). We also found many differentially spliced introns that are not located within any annotated genic region, which likely reflects splicing of non-coding RNA transcripts (Additional file 15: Table S7). Under the same conditions, 178 genes were differentially expressed (> 2.5 fold down- or up-regulated in terms of overall expression abundance) including only two SM genes (Additional file 17: Table S8). Because of splicing variation within a group (i.e., heat) due to differences in treatment (42 °C and 46 °C), our result represents a conservative estimate of the extent of differential splicing between the overall 'heat' and 'cold' conditions. Although a comprehensive analysis of temperature-dependent gene expression and the functional consequences of differential splicing are not within the scope of this paper, our results suggest that *G. sulphuraria* is able to respond to temperature changes (and likely other stimuli) using differential mRNA-splicing. These results explain the considerable fluctuations in microenvironments where many of these species live (e.g., non-thermophilic *Galdieria soos* [25]). Given the overall reduced SM in red algae, compared to humans and Viridiplantae, the prevalence of intron retention might also have resulted from compromised splicing efficiency in *G. sulphuraria*. How this possible scenario contributes to overall intron retention in *G. sulphuraria* is unknown.

Discussion

We show here that components of the spliceosomal machinery have undergone recent gene losses and accelerated evolution among different red algal lineages. In this context, the high conservation of the SM in *Galdieria* is

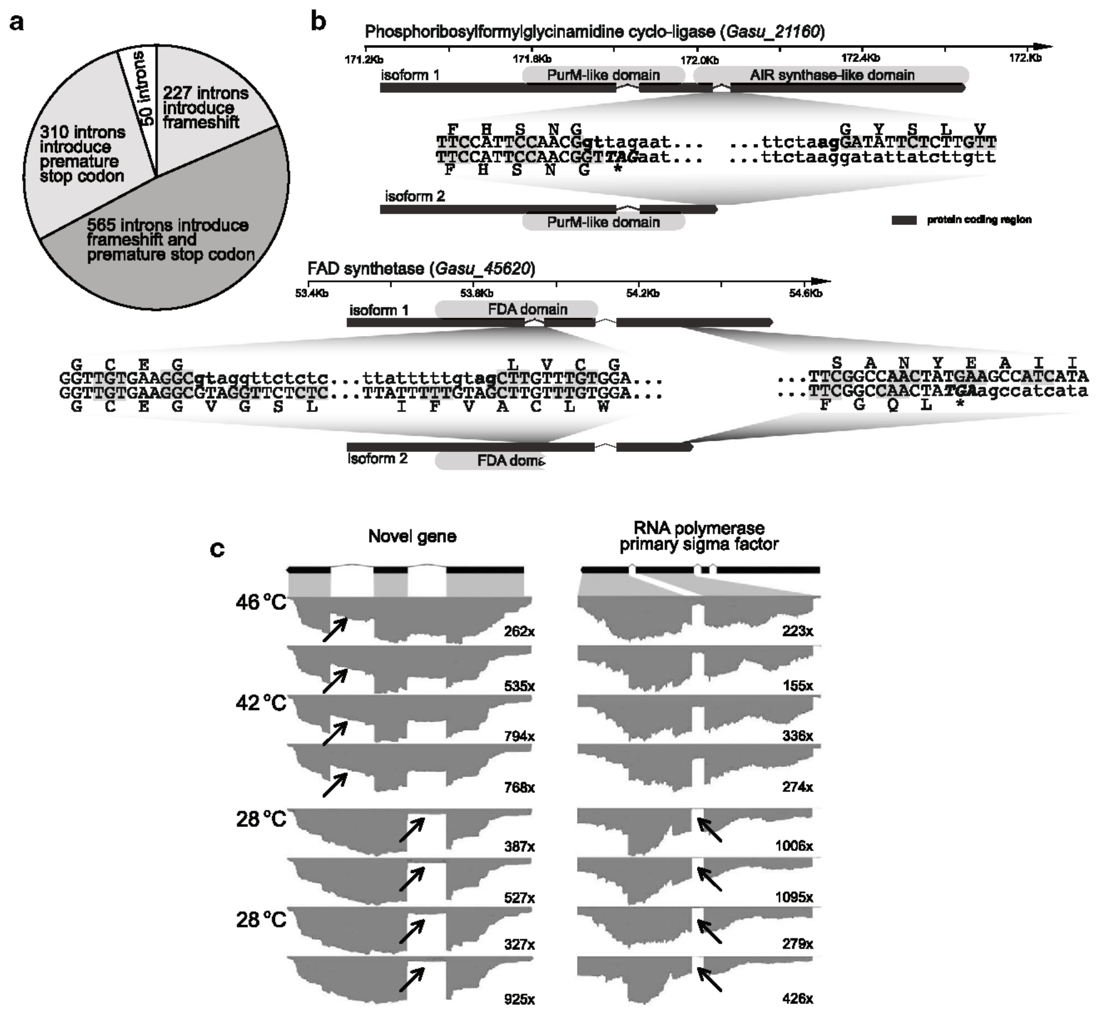


Fig. 3 Intron retention in *G. sulphuraria*. **a** Predicted changes resulting from the retention of 1152 alternatively spliced introns identified in this study. **b** Examples of intron retention in two *G. sulphuraria* genes. The sequences flanking the splicing signals (GT-AG) are illustrated. Coding nucleotides are shown in uppercase and non-coding nucleotides in lower case. Gray rectangular boxes are aligned with every other codon triplet. Gene models comprise thin black boxes (exons) connected with angled lines (introns). **c** Two examples of differential intron splicing between the heat and cold conditions in *G. sulphuraria*. Gene structures are shown at the top with the black boxes indicating the exons. For each gene, the read coverage along the displayed regions is shown for a total of 8 samples, with 2 replicates for each at the 4 temperature treatments (see Methods for details). The maximum read coverage is shown for each sample. For each intron, regions that are significantly more spliced than the compared temperature group are indicated with arrows

counterintuitive, given its relatively more reduced gene inventory resulting from lineage-specific genome reduction (Fig. 1b, arrow #2). In addition, *G. sulphuraria* contains a relatively large number of introns via ancient intron preservation and novel insertions, in spite of its compact genome size. This unexpected evolutionary trajectory allows alternative mRNA splicing that generates transcriptomic (and likely proteomic) diversity [26, 27] in this lineage for about a quarter of the alternatively spliced transcripts that do not encode premature stop codons (including 4.3% leading to insertion of amino acids and 19.7% resulting in novel peptides due to frame-shifts, e.g., Fig. 3b). The functional impact of alternative splicing on *G. sulphuraria* biology remains to be investigated.

It is noteworthy that of the 1152 well annotated, alternatively spliced introns we identified in *G. sulphuraria*, most cases (68.7%) of intron retention lead to a truncated protein (Fig. 3a). Why might this be tolerated? The most likely reason is nonsense-mediated decay (NMD), a process that is widespread in eukaryotes for regulating post-transcriptional gene expression [28]. NMD allows for the targeted degradation of alternatively spliced isoforms that would result in truncated proteins due to the introduction of premature termination codons (PTCs; e.g., due to intron retention), as described previously [29]. NMD is not however completely effective and PTCs persist in the transcript pools of many eukaryotes, suggesting a functional role [30]. Soergel et al. [31] postulated an evolutionary interaction between AS and NMD, that allows the rise of alternative, beneficial splice forms (i.e., under the umbrella of a well-established surveillance system) that can ultimately be fixed in the population. This scenario may provide an explanation for the extensive AS-derived PTCs we found in the *G. sulphuraria* RNA-seq data. Several key genes in NMD (i.e., UPF1-3) are present in *G. sulphuraria* and other red algal species (Additional file 18: Figure S9). Our findings with *G. sulphuraria* are generally in line with existing data from other systems. In *Arabidopsis thaliana*, about 13% of intron-containing genes are potentially regulated by AS/ NMD [29]. In the unicellular green alga, *Chlamydomonas reinhardtii*, there are 611 AS events that impact 3% of all genes with intron retention being the most common outcome leading to many PTCs [32]. The AS-derived variants may enhance gene regulation in *G. sulphuraria*. This idea is consistent with the reduced intergenic regions (i.e., that could encode *cis*-regulatory elements) that has resulted from genome streamlining (Additional file 13: Figure S7B).

An additional possible explanation for our results comes from a recent study of the yeast UV stress response where genes associated with transcription are regulated by non-coding RNAs derived from alternatively spliced, short transcripts of the same gene (i.e., alternative last exons (ALEs) [33, 34]). In the case of the *ASCC3* gene that represses RNA polymerase II transcription after UV irradiation, transcription of the complete gene (i.e., full-length protein) is de-repressed by an ALE derived from the same gene that acts as a non-coding regulatory RNA. Therefore, it is possible that stress pathways (i.e., not UV irradiation) impacted by our heat and cold treatments of *G. sulphuraria* may lead to the generation of shorter non-coding RNAs via AS that play a role in regulating the stress response. Thermal stress is clearly a major factor in the ecology of *G. sulphuraria*, therefore AS may produce both novel protein isoforms as well as regulatory RNAs (perhaps like ALEs) that play roles in responding to this stress. More generally, our results suggest that strong constraints that exist on the growth of gene numbers (and functions) due to genome reduction can be ameliorated at the transcriptome level. This insight required the analysis of free-living organisms that have relatively complex genomes (i.e., containing introns) and a history of ancient genome reduction, together with recent lineage-specific gene losses. In this regard, our results underline the utility of free-living taxa such as red algae as models for studying eukaryote genome reduction.

Conclusions

Our results revealed an unexpected aspect of *Galdieria* genome evolution. Whereas the correlation between SM gene number and spliceosomal intron density within red algae remains to be validated with more genomic data, our findings lead to several hypotheses that can be tested in this unique model to understand genome reduction in free-living organisms.

Methods

Detection of spliceosomal proteins in red algae

The culture of *G. sulphuraria* used in this study is the strain with the completed nuclear genome sequence (i.e., 074 W) and was isolated from a site near Reykjavik Island [8]. Using the 215 non-redundant human SM proteins as queries (Additional file 19: Supplementary Methods), we searched the proteomes from eight red algal species and nine Viridiplantae (Fig. 1a and Additional file 1: Table S1) using BLASTp (*e*-value cutoff = 1e-5). Significant hits that led to a reasonable alignment length (query coverage > 30%) and had the highest hit-query identities were recorded for each query versus each search species. The resulting data were clustered and visualized with the heatmap function in the R language. Genes and taxa were clustered using Euclidean distances between all gene (or taxon) pairs and the complete-linkage clustering method.

To search red algal SM with higher stringencies and examine their origins, we adopted a phylogenetic-based method [6]. We generated a proteome data comprising SM proteins derived from homology-based gene predictions (using human proteins as reference; Additional file 19: Supplementary Methods) and protein models annotated in existing studies (Table S1). To identify *G. sulphuraria* SM proteins, we searched the *G. sulphuraria* proteome data with BLASTp (*e*-value cutoff = 1e-3) using human SM proteins as queries. The top three *G. sulphuraria* hits according to bit-score (by default) and the top three hits according to query-hit alignment identity were recorded for further validation. To differentiate between orthologous and paralogous relationships between human SM queries and their *G. sulphuraria* homologs, these proteins were used as queries to search (BLASTp *e*-value cutoff = 1e-5) against a comprehensive local protein database [6]. The significant hits were recorded for each SM query and the representative sequences were selected with up to 8 sequences for each phylum in the default order sorted by bit-score. A second set of representative sequences was selected after re-sorting the BLASTp hits according to the query-hit sequence identity. The two sets of representative sequences (by bit-score and alignment identity) for all the SM queries (human and *G. sulphuraria* homologs potentially corresponding to the same SM gene) were then combined, aligned using MUSCLE (v3.8.31) [35] and trimmed using TrimAl (version 1.2) [36] in automated mode (–automated1). The phylogenetic tree was constructed using FastTree (version 2.1.7) [37] under the 'WAG+CAT' model with 4 rounds of minimum evolution SPR moves (–psr 4) and exhaustive ML nearest-neighbor interchanges (–mlacc 2, –slownni). Branch support was derived from the Shimodaira-Hasegawa test [38]. We examine the resulting phylogenies manually. A SM gene was regarded to exist in *G. sulphuraria*, if at least one of the *G. sulphuraria* gene candidates appeared in the same orthologous group as the human SM gene (see Additional file 2: Figure S1A for an example of this approach). The SM gene was considered to be absent if no *G. sulphuraria* candidates were found in the orthologous group with the human SM gene (see Additional file 2: Figure S1B for an example of this approach). HGT was inferred when the candidate red algal sequences were nested within multiple sequences from prokaryotic and/or fungal taxa. Following the same procedure as described above, the presence and absence of SM genes were determined in other red algal lineages that are shown in Fig. 1b.

In addition, we used *Galdieria* SM proteins as references for homology-based gene prediction in genome or transcriptome data from the remaining six red algal species (Fig. 2) (Additional file 19: Supplementary Methods). The resulting SM proteins were incorporated into our comprehensive local protein database described above. Using the *G. sulphuraria* SM proteins (or *G. phlegrea* when the *G. sulphuraria* gene was missing) as queries, we carried out a BLASTp search, sorted the significant hits, selected representative sequences, and aligned and built phylogenetic trees following the procedures described above. The resulting trees were manually inspected to identify additional red algal SM sequences that were monophyletic with the *Galdieria* queries.

Assessing completeness of the protein data

We used BUSCO (version 3) under the default settings to estimate the overall completeness of protein data (equivalent to genomic coverage) for each red algal species [39]. The 'Eukaryota sets' that contained 303 conserved gene families in eukaryotes were used as the reference for this analysis.

Detection of snRNAs

We downloaded snRNA alignments for U1 (RF00003), U2 (RF00004), U4 (RF00015), U5 (RF00020) and U6 (RF00026) from the Rfam database [40]. The alignments were calibrated (using cmcalibrate) and then used for snRNA searches in red algal genomes (using cmsearch) using Infernal (v1.1.2) with the default settings [41].

Intron analysis

We downloaded the genome and coding DNA sequences (CDSs) from six red algal species that have high-quality whole genome sequences: *G. sulphuraria* [8], *C. merolae* [7], *P. purpureum* [10], *C. crispus* [11], *P. yezoensis* [12] and *G. chorda* (unpublished data), and from five Viridiplantae species (Table S4). The CDSs were mapped to the corresponding genome sequences using BLAT [42] under the default settings. The non-specific alignments were removed and the positions of introns (in genomes) and exon junctions (in CDSs) were then subtracted from the BLAT output using custom scripts. Most of the introns (94-99%) were flanked by the canonical splicing signal (GT-AG) (Table S4).

To identify *G. sulphuraria* intron positions that are shared with *C. crispus*, we searched *G. sulphuraria* proteins against the *C. crispus* proteome using BLASTp (*e*-value cutoff = 1e-10) and retrieved information about the top 10 BLASTp hits. Because the original intron positions along the protein primary sequences are not comparable across sequences due to variable lengths of N′-terminal domains, insertions, and deletions, we built alignments for each query protein and its corresponding *C. crispus* hit(s) using MUSCLE (v3.8.31) [35]. With gaps being introduced during the alignment procedure, the intron positions in the original sequences (without gaps) were converted into column numbers for each sequence in their respective alignments. A *G. sulphuraria* intron position was considered as being conserved if it was located at the same column position in an alignment with one or more introns of the same phase in *C. crispus*. The same method was used to identify intron positions that were shared between any two of the species included in this study (Fig. 2b).

Galdieria sulphuraria cell cultures

Biological replicate cultures of *G. sulphuraria* 074W were grown separately at 42 °C, constant illumination (90 μE), and constant shaking (160 rpm) in photoautotrophic conditions using 2xGS Medium [43]. The experimental design followed a temperature shift timeline: after two weeks of cultivation at stated conditions the first sampling took place (H-42) and the cultures were swiftly moved to 28 °C. After cold treatment at 28 °C for 48 h, a second sampling was performed (C-28.1). The *G. sulphuraria* was then switched to 46 °C for 48 h, at the end of which a third sample was retrieved (H-46). It again was followed by a cold treatment at 28 °C for 48 h when a fourth sample was retrieved (C-28.2). Altogether, two time-points from high temperatures (42 °C and 46 ° C) and two time-points from cold temperatures (28 °C) were targeted for sampling.

Sequencing of the *Galdieria sulphuraria* transcriptome

RNA was extracted using Roboklon's Universal RNA Purification Kit by following the "plant tissue samples" protocol (Roboklon, Berlin - Germany). RNA quality and concentration was assessed using a Nanodrop photospectrometer ND-1000 (Peqlab Biotechnologie GmbH, Erlangen -Germany). The samples were synthesized to be compatible with Illumina HiSeq2000 RNA-seq libraries strictly following the "Illumina TruSeq RNA Sample Prep v2 LS Protocol" (Illumina, San Diego - USA). All reagents were scaled by 2/3 to the volume proposed in the protocol. The quality of all libraries was assessed using the Bioanalyzer (Agilent Technologies, Santa Clara - USA). The libraries were sequenced in paired-end mode (2x100bp) on two lanes with an Illumina HiSeq2000 sequencer at the BMFZ (Biologisch-Medizinisches Forschungszentrum, Düsseldorf, Germany). The resulting RNA-seq data were deposited in the NCBI Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE89169.

Detection of differentially spliced introns

The *G. sulphuraria* RNA-seq data were cleaned in paired-end mode using Trimmomatic (v0.36) [44] to remove contaminated adaptor sequences and low quality regions (SLIDINGWINDOW:6:13). Short reads (< 75 bp) were discarded. The cleaned sequence data were then mapped to *G. sulphuraria* genome sequences using STAR (v2.5.2a) [45]. The genome index was generated taking into account the small size of *G. sulphuraria* genome (−-genomeSAindexNbases 11). The reads were mapped to the genome assembly with an allowed maximum intron size (1000 bp) and maximum mate-pair distance (500 bp). Reads that mapped to more than one region were removed and broken pairs were discarded. The mapping results (in SAM format) were then used as input to search for alternatively spliced modules [46] that were differentially expressed across samples using DiffSplice [46] under the default setting, with the following modifications. We required a splice junction to be considered if the mean coverage across all samples was > 10× and RNA-splicing at the junction was found in at least four out of the eight different samples. The expression thresholds for exons and introns were specified to be 16× and 8× coverage, respectively. For the test of differential splicing, the minimal value for square root of JSD [46] was set to be 0.25 with false discovery rate threshold (=0.01). The minimum fold change (2.5) was required to call gene differential expression (down- or up-regulation). Because we aimed to test the existence of differential intron splicing in response to temperature changes (instead of global gene expression change across samples), we regarded the two samples (H-42 and H-46) as biological replicates from high temperature, and the other two samples (C-28.1 and C-28.2) as biological replicates from low temperature. This practice maximized statistical power to detect splicing events that were shared within groups (e.g., high temperature samples including H-42 and H-46) and differed between the groups (high versus low temperature samples). The examples of alternatively spliced modules that showed statistically significant difference between high and low temperatures (Fig. 3c) were visualized using CLC workbench (v8) (http://www.clcbio.com/products/clc-main-workbench/).

Additional files

Additional file 1: Table S1. Algal genome and transcriptome data used in this study. (PDF 96 kb)

Additional file 2: Figure S1. Two examples of spliceosomal single-gene phylogeny that show different ancestries of red algal spliceosomal genes. (PDF 137 kb)

Additional file 3: Phylogenies of red algal SM-genes. (TXT 5414 kb)

Additional file 4: Table S2. Completeness of proteomic data estimated using 303 BUSCO gene families that are evolutionarily conserved among eukaryotes. (PDF 119 kb)

Additional file 5: Table S3. Presence and absence of human spliceosomal machinery-associated proteins in red algae. (PDF 134 kb)

Additional file 6: Figure S2. The search results for snRNA component of the spliceosome in *Galdieria sulphuraria*. (PDF 91 kb)

Additional file 7: Table S4. INFERNAL scores and e-values for red algal snRNA genes. (PDF 64 kb)

Additional file 8: Figure S3. Sequence conservation in *Galdieria sulphuraria* genes. (PDF 181 kb)

Additional file 9: Table S5. The intron statistics in red algal and Viridiplantae genomes. (PDF 71 kb)

Additional file 10: Figure S4. GC content and intron density in red algae. (PDF 92 kb)

Additional file 11: Figure S5. Conservation of intron positions in the *Galdieria sulphuraria* geranylgeranyl transferase beta-subunit gene. (PDF 97 kb)

Additional file 12: Figure S6. Estimation of gains and losses of conserved introns in red algal phylogeny. (PDF 92 kb)

Additional file 13: Figure S7. The distributions of intron lengths in five red algal species. (PDF 82 kb)

Additional file 14: Table S6. *Galdieria sulphuraria* introns that underwent alternative splicing in our studied samples. (PDF 981 kb)

Additional file 15: Table S7. *Galdieria sulphuraria* introns that were differentially spliced under the heat and cold conditions. (PDF 226 kb)

Additional file 16: Figure S8. Intron retention in a *Galdieria sulphuraria* gene. (PDF 74 kb)

Additional file 17: Table S8. *Galdieria sulphuraria* genes that were differentially expressed under the heat and cold conditions. (PDF 186 kb)

Additional file 18: Figure S9. Phylogenetic trees of UPF1, UPF2, and UPF3. (PDF 97 kb)

Additional file 19: Supplementary Methods. (PDF 109 kb)

Abbreviations

ALE: Alternative last exons; AS: Alternative splicing; CAZyme: Carbohydrate-active enzyme; NMD: Nonsense-mediated decay; PTC: Premature termination codons; SM: Spliceosomal machinery

Acknowledgements

Not applicable.

Funding

HQ is partially supported by a grant from the National Science Foundation of the United States (EF1416785). HSY and DB acknowledge the support from the Collaborative Genome Program for Fostering New Post-Genome Industry (20140428) funded by the Ministry of Oceans and Fisheries of Korea. We gratefully acknowledge the BMFZ (Biologisch-Medizinisches Forschungszentrum, Düsseldorf, Germany) for conducting the RNA sequencing reported here. The funders had no role in study design, data collection, analysis and interpretation, or preparation of the manuscript.

Availability of data and materials

The datasets generated and analysed during the current study are available in the NCBI Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE89169.

Authors' contributions

HQ and DB designed the study. HSY contributed novel red algal genome data. APW conceived and designed the RNA-seq analysis of *G. sulphuraria* under temperature stress. AR extracted RNA from *G. sulphuraria*, prepared RNA-seq libraries, and contributed transcriptome data. HQ and AR performed data analyses. HQ and AR drafted the manuscript. APW, HSY, and DB revised the manuscript with each author making important intellectual contributions. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]Department of Ecology, Evolution and Natural Resources, Rutgers University, New Brunswick, NJ 08901, USA. [2]Institute for Plant Biochemistry, Cluster of Excellence on Plant Sciences (CEPLAS), Heinrich-Heine-University, D-40225 Düsseldorf, Germany. [3]Department of Biological Sciences, Sungkyunkwan University, Suwon 16419, Korea. [4]Department of Biochemistry and Microbiology, Rutgers University, New Brunswick, NJ 08901, USA.

Received: 28 July 2017 Accepted: 22 March 2018
Published online: 02 April 2018

References

1. Van de Peer Y, Maere S, Meyer A. The evolutionary significance of ancient genome duplications. Nat Rev Genet. 2009;10:725–32.
2. Fedoroff NV. Transposable elements, epigenetics, and genome evolution. Science. 2012;338:758–67.
3. Wolf YI, Koonin EV. Genome reduction as the dominant mode of evolution. BioEssays. 2013;35:829–37.
4. McCutcheon JP, Moran NA. Extreme genome reduction in symbiotic bacteria. Nat Rev Microbiol. 2012;10:13–26.
5. Guiry MD. How many species of algae are there? J Phycol. 2012;48:1057–63.
6. Qiu H, Price DC, Yang EC, Yoon HS, Bhattacharya D. Evidence of ancient genome reduction in red algae (Rhodophyta). J Phycol. 2015;51:624–36.
7. Matsuzaki M, Misumi O, Shin-I T, Maruyama S, Takahara M, Miyagishima S-Y, et al. Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. Nature. 2004;428:653–7.
8. Schönknecht G, Chen W-H, Ternes CM, Barbier GG, Shrestha RP, Stanke M, et al. Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote. Science. 2013;339:1207–10.
9. Qiu H, Price DC, Weber APM, Reeb V, Yang EC, Lee JM, et al. Adaptation through horizontal gene transfer in the cryptoendolithic red alga *Galdieria phlegrea*. Curr Biol. 2013;23:R865–6.
10. Bhattacharya D, Price DC, Chan CX, Qiu H, Rose N, Ball S, et al. Genome of the red alga *Porphyridium purpureum*. Nat Commun. 2013;4:1941.
11. Collén J, Porcel B, Carré W, Ball SG, Chaparro C, Tonon T, et al. Genome structure and metabolic features in the red seaweed *Chondrus crispus* shed light on evolution of the Archaeplastida. Proc Natl Acad Sci U S A. 2013;110:5247–52.
12. Nakamura Y, Sasaki N, Kobayashi M, Ojima N, Yasuike M, Shigenobu Y, et al. The first symbiont-free genome sequence of marine red alga, Susabi-nori (*Pyropia yezoensis*). PLoS One. 2013;8:e57122.
13. Shang X, Cao Y, Ma L. Alternative splicing in plant genes: a means of regulating the environmental fitness of plants. Int J Mol Sci. 2017;18:432.
14. Hegele A, Kamburov A, Grossmann A, Sourlis C, Wowro S, Weimann M, et al. Dynamic protein-protein interaction wiring of the human spliceosome. Mol Cell. 2012;45:567–80.
15. Stark MR, Dunn EA, Dunn WSC, Grisdale CJ, Daniele AR, Halstead MRG, et al. Dramatically reduced spliceosome in *Cyanidioschyzon merolae*. Proc Natl Acad Sci U S A. 2015;112:E1191–200.
16. Qiu H, Yoon HS, Bhattacharya D. Red algal phylogenomics provides a robust framework for inferring evolution of key metabolic pathways. PLoS Curr. 2016; 8. https://doi.org/10.1371/currents.tol.7b037376e6d84a1be34af756a4d90846.
17. Farris J. Phylogenetic analysis under Dollo's law. Syst Zool. 1977;26:77–88.
18. Barbier G, Oesterhelt C, Larson MD, Halgren RG, Wilkerson C, Garavito RM, et al. Comparative genomics of two closely related unicellular thermo-acidophilic red algae, *Galdieria sulphuraria* and *Cyanidioschyzon merolae*, reveals the molecular basis of the metabolic flexibility of *Galdieria sulphuraria* and significant differences in carbohydrate metabolism of both algae. Plant Physiol. 2005;137:460–74.
19. Hudson AJ, Stark MR, Fast NM, Russell AG, Rader SD. Splicing diversity revealed by reduced spliceosomes in *C. merolae* and other organisms. RNA Biol. 2015;12:1–8.
20. Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, et al. The genome of the African trypanosome *Trypanosoma brucei*. Science. 2005;309:416–22.
21. Katinka MD, Duprat S, Cornillot E, Méténier G, Thomarat F, Prensier G, et al. Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. Nature. 2001;414:450–3.
22. Morrison HG, McArthur AG, Gillin FD, Aley SB, Adam RD, Olsen GJ, et al. Genomic minimalism in the early diverging intestinal parasite *Giardia lamblia*. Science. 2007;317:1921–6.
23. Rogozin IB, Carmel L, Csuros M, Koonin EV. Origin and evolution of spliceosomal introns. Biol Direct. 2012;7:11.
24. Hirooka S, Miyagishima S-Y. Cultivation of acidophilic algae *Galdieria sulphuraria* and *Pseudochlorella* sp. YKT1 in media derived from acidic hot springs. Front Microbiol. 2016;7:2022.
25. Gross W, Oesterhelt C, Tischendorf G, Lederer F. Characterization of a non-thermophilic strain of the red algal genus *Galdieria* isolated from Soos (Czech Republic). Eur J Phycol. 2002;37:477–82.
26. Severing EI, van Dijk ADJ, van Ham RCHJ. Assessing the contribution of alternative splicing to proteome diversity in *Arabidopsis thaliana* using proteomics data. BMC Plant Biol. 2011;11:82.
27. Tress ML, Bodenmiller B, Aebersold R, Valencia A. Proteomics studies confirm the presence of alternative protein isoforms on a large scale. Genome Biol. 2008;9:R162.
28. Chang Y-F, Imam JS, Wilkinson MF. The nonsense-mediated decay RNA surveillance pathway. Annu Rev Biochem. 2007;76:51–74.
29. Kalyna M, Simpson CG, Syed NH, Lewandowska D, Marquez Y, Kusenda B, et al. Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in *Arabidopsis*. Nucleic Acids Res. 2012;40:2454–69.
30. Baek D, Green P. Sequence conservation, relative isoform frequencies, and nonsense-mediated decay in evolutionarily conserved alternative splicing. Proc Natl Acad Sci U S A. 2005;102:12813–8.
31. DAW S, Lareau LF, Brenner SE. Regulation of gene expression by coupling of alternative splicing and NMD. Nonsense-Mediat. MRNA decay. Georgetown: Landes Bioscience; 2006. p. 175–96.

32. Labadorf A, Link A, Rogers MF, Thomas J, Reddy AS, Ben-Hur A. Genome-wide analysis of alternative splicing in *Chlamydomonas reinhardtii*. BMC Genomics. 2010;11:114.
33. Conconi A, Bell B. Molecular biology: the long and short of a DNA-damage response. Nature. 2017;545:165–6.
34. Williamson L, Saponaro M, Boeing S, East P, Mitter R, Kantidakis T, et al. UV irradiation induces a non-coding RNA that functionally opposes the protein encoded by the same gene. Cell. 2017;168:843–855.e13.
35. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 2004;5:113.
36. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinforma Oxf Engl. 2009;25:1972–3.
37. Price MN, Dehal PS, Arkin AP. FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS One. 2010;5:e9490.
38. Shimodaira H, Hasegawa M. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol. 1999;16:1114–6.
39. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinforma. Oxf. Engl. 2015;31:3210–2.
40. Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, et al. Rfam 12.0: updates to the RNA families database. Nucleic Acids Res. 2015;43:D130–7.
41. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinforma. Oxf. Engl. 2013;29:2933–5.
42. Kent WJ. BLAT–the BLAST-like alignment tool. Genome Res. 2002;12:656–64.
43. Allen MB. Studies with *Cyanidium caldarium*, an anomalously pigmented chlorophyte. Arch Für Mikrobiol. 1959;32:270–7.
44. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinforma Oxf Engl. 2014;30:2114–20.
45. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinforma. Oxf. Engl. 2013;29:15–21.
46. Hu Y, Huang Y, Du Y, Orellana CF, Singh D, Johnson AR, et al. DiffSplice: the genome-wide detection of differential splicing events with RNA-seq. Nucleic Acids Res. 2013;41:e39.

Manuscript 2

The genomes of polyextremophilic cyanidiales contain 1% horizontally transferred genes with diverse adaptive functions

eLIFE
elifesciences.org

RESEARCH ARTICLE

The genomes of polyextremophilic cyanidiales contain 1% horizontally transferred genes with diverse adaptive functions

Alessandro W Rossoni[1], Dana C Price[2], Mark Seger[3], Dagmar Lyska[1], Peter Lammers[3], Debashish Bhattacharya[4], Andreas PM Weber[1]*

[1]Institute of Plant Biochemistry, Cluster of Excellence on Plant Sciences (CEPLAS), Heinrich Heine University, Düsseldorf, Germany; [2]Department of Plant Biology, Rutgers University, New Brunswick, United States; [3]Arizona Center for Algae Technology and Innovation, Arizona State University, Mesa, United States; [4]Department of Biochemistry and Microbiology, Rutgers University, New Brunswick, United States

Abstract The role and extent of horizontal gene transfer (HGT) in eukaryotes are hotly disputed topics that impact our understanding of the origin of metabolic processes and the role of organelles in cellular evolution. We addressed this issue by analyzing 10 novel Cyanidiales genomes and determined that 1% of their gene inventory is HGT-derived. Numerous HGT candidates share a close phylogenetic relationship with prokaryotes that live in similar habitats as the Cyanidiales and encode functions related to polyextremophily. HGT candidates differ from native genes in GC-content, number of splice sites, and gene expression. HGT candidates are more prone to loss, which may explain the absence of a eukaryotic pan-genome. Therefore, the lack of a pan-genome and cumulative effects fail to provide substantive arguments against our hypothesis of recurring HGT followed by differential loss in eukaryotes. The maintenance of 1% HGTs, even under selection for genome reduction, underlines the importance of non-endosymbiosis related foreign gene acquisition.
DOI: https://doi.org/10.7554/eLife.45017.001

*For correspondence: andreas.weber@uni-duesseldorf.de

Competing interests: The authors declare that no competing interests exist.

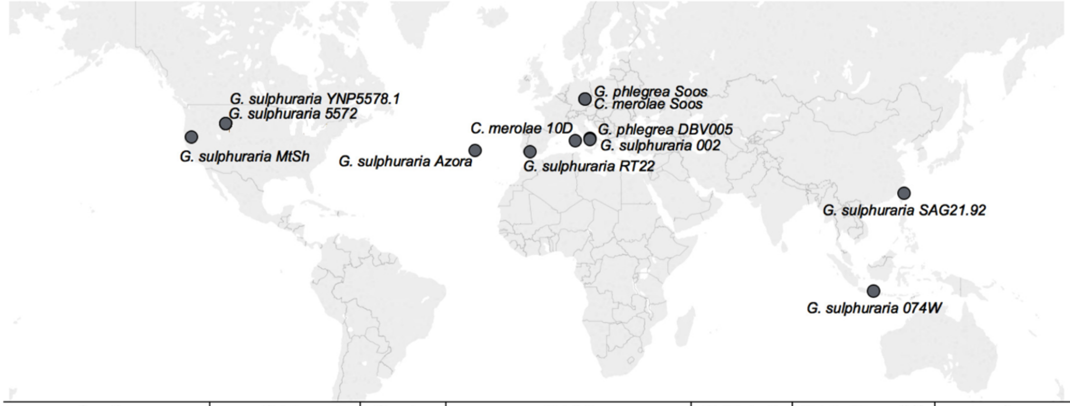
Funding: See page 28

Received: 10 January 2019
Accepted: 30 May 2019
Published: 31 May 2019

Reviewing editor: Paul B Rainey, Max Planck Institute for Evolutionary Biology, Germany

Introduction

Eukaryotes transmit their nuclear and organellar genomes from one generation to the next in a vertical manner. As such, eukaryotic evolution is primarily driven by the accumulation, divergence (e.g., due to mutation, insertion, duplication), fixation, and loss of gene variants over time. In contrast, horizontal (also referred to as lateral) gene transfer (HGT) is the inter- and intraspecific transmission of genes from parents to their offspring. HGT in Bacteria (***Doolittle, 1999***; ***Ochman et al., 2000***; ***Boucher et al., 2003***) and Archaea (***Nelson-Sathi et al., 2012***) is widely accepted and recognized as an important driver of evolution leading to the formation of pan-genomes (***Tettelin et al., 2005***; ***Vernikos et al., 2015***). A pan-genome comprises all genes shared by any defined phylogenetic clade and includes the so-called core (shared) genes associated with central metabolic processes, dispensable genes present in a subset of lineages often associated with the origin of adaptive traits, and lineage-specific genes (***Vernikos et al., 2015***). This phenomenon is so pervasive that it has been questioned whether prokaryotic genealogies can be reconstructed with any confidence using standard phylogenetic methods (***Philippe and Douady, 2003***; ***Doolittle and Brunet, 2016***). In contrast, as eukaryote genome sequencing has advanced, an increasing body of data has pointed towards



Species	Origin	Country	Habitat	Habitat pH	Habitat Temp (°C)	Source
C. merolae 10D*	Sardinia	Italy	Acidic Hot Spring	1.5	Up to 45°C	ATCC®, T. Kuroiwa
C. merolae Soos	Soos National Park	CZ	Diatom field	0.8 - 2	< 0° - 30°C	W. Gross, M. Seger
G. phlegrea DBV009*	Nepi	Italy	Sulphuar Spring	0.8	12°C	G. Pinto, ACUF
G. phlegrea Soos	Soos National Park	CZ	Diatom field	0.8 - 2	< 0° - 30°C	W. Gross, M. Seger
G. sulphuraria 002 (S)	La Solfatara	Italy	na	1	36°C	G. Pinto
G. sulphuraria 074W*	Mount Lawu	Indonesia	Fumaroles	na	35°C	W. Gross, P. De Luca
G. sulphuraria 5572	Norris Basin, YNP	USA	Acidic soil	1	55°C	M. Seger, R. W. Castenholz
G. sulphuraria Azora	Azores	Portugal	Porous sandstone, endolithic	2.1	na	W. Gross, A. Flechner
G. sulphuraria MS1	Contaminant	USA	Ronust contaminant of YNP cultures	na	na	M. Seger, P. Lammers
G. sulphuraria MtSh	Mount Shasta	USA	Soil, close to mountain peak (4300m)	2.2	na	W. Gross, R. R. Pausewein
G. sulphuraria RT22	Rio Tinto, Berrocal	Spain	Riverbank, endolithic	2.5	na	W. Gross, R. R. Pausewein
G. sulphuraria SAG21.92	Yangmingshan	Taiwan	Hot spring	na	na	J. T.
G. sulphuraria YNP5578.1	Nymph Creek, YNP	USA	Acid stream	3	42°C	M. Seger, R. W. Castenholz

Figure 1. Geographic origin and habitat description of the analyzed Cyanidiales strains. Available reference genomes are marked with an asterisk (*), whereas 'na' indicates missing information.

DOI: https://doi.org/10.7554/eLife.45017.002

the existence of HGT in these taxa, but at much lower rates than in prokaryotes (***Danchin, 2016***). The frequency and impact of eukaryotic HGT outside the context of endosymbiosis and pathogenicity however, remain hotly debated topics in evolutionary biology. Opinions range from the existence of ubiquitous and regular occurrence of eukaryotic HGT (***Husnik and McCutcheon, 2018***) to the almost complete dismissal of any eukaryotic HGT outside the context of endosymbiosis as being Lamarckian, thus false, and resulting from analysis artefacts (***Martin, 2018***; ***Martin, 2017***). HGT skeptics favor the alternative hypothesis of differential loss (DL) to explain the current data. DL imposes strict vertical inheritance (eukaryotic origin) on all genes outside the context of pathogenicity and endosymbiosis, including putative HGTs. Therefore, all extant genes have their root in LECA, the last eukaryotic common ancestor. Patchy gene distributions are the result of multiple ancient paralogs in LECA that have been lost over time in some eukaryotic lineages but retained in others. Under this view, there is no eukaryotic pan-genome, there are no cumulative effects (e.g., the evolution of eukaryotic gene structures and accrual of divergence over time), and therefore, mechanisms for the uptake and integration of foreign DNA in eukaryotes are unnecessary.

A comprehensive analysis of the frequency of eukaryotic HGT was recently done by ***Ku and Martin (2016)***. These authors reported the absence of eukaryotic HGT candidates sharing over 70% protein identity with their putative non-eukaryotic donors (for very recent HGTs, this figure could be as high as 100%). Furthermore, no continuous sequence identity distribution was detected for HGT candidates across eukaryotes and the 'the 70% rule' was proposed ('*Coding sequences in eukaryotic genomes that share more than 70% amino acid sequence identity to prokaryotic homologs are most likely assembly or annotation artifacts.*') (***Ku and Martin, 2016***). However, as noted by others (***Richards and Monier, 2016***; ***Leger, 2018***), this result was obtained by categorically dismissing all eukaryotic HGT singletons located within non-eukaryotic branches as assembly/annotation artefacts, as well as those remaining that exceeded the 70% threshold. In addition, all genes that were presumed to be of organellar origin were excluded from the analysis, leaving a small dataset extracted from already under-sampled eukaryotic genomes.

Given these uncertainties, the aim of our work was to systematically analyze eukaryotic HGT using the Cyanidiales (known as Cyanidiophytina in some taxonomic schemes) as model organisms. The Cyanidiales comprise a monophyletic clade of polyextremophilic, unicellular red algae (Rhodophyta) that thrive in acidic and thermal habitats worldwide (e.g., volcanoes, geysers, acid mining sites, acid rivers, urban wastewaters, geothermal plants) (**Castenholz and McDermott, 2010**). With a divergence age estimated to be around 1.92–1.37 billion years (**Yoon et al., 2004**), the Cyanidiales are the earliest split within Rhodophyta and define one of the oldest surviving eukaryotic lineages. They are located near the root of the supergroup Archaeplastida, whose ancestor underwent the primary plastid endosymbiosis with a cyanobacterium that established photosynthesis in eukaryotes (**Reyes-Prieto et al., 2007**; **Price et al., 2012**). In the context of HGT, the Cyanidiales became more broadly known after publication of the genome sequences of *Cyanidioschyzon merolae* 10D (**Matsuzaki et al., 2004**; **Nozaki et al., 2007**), *Galdieria sulphuraria* 074W (**Schönknecht et al., 2013**), and *Galdieria phlegrea* DBV009 (**Qiu et al., 2013**). The majority of putative HGTs in these taxa was hypothesized to have provided selective advantages during the evolution of polyextremophily, contributing to the ability of *Galdieria*, *Cyanidioschyzon*, and *Cyanidium* to cope with extremely low pH values, temperatures up to 56˚C, as well as high salt and toxic heavy metal ion concentrations (**Castenholz and McDermott, 2010**; **Doemel and Brock, 1971**; **Reeb and Bhattacharya, 2010**; **Hsieh et al., 2018**). In such environments, they can represent up to 90% of the total biomass, competing with specialized Bacteria and Archaea (**Seckbach, 1972**), although some Cyanidiales strains also occur in more temperate environments (**Qiu et al., 2013**; **Gross et al., 2002**; **Ciniglia et al., 2004**; **Barcyté et al., 2018**; **Iovinella et al., 2018**). The integration and maintenance of HGT-derived genes, in spite of strong selection for genome reduction in these taxa (**Qiu et al., 2015**) underlines the potential ecological importance of this process to niche specialization (**Schönknecht et al., 2013**; **Qiu et al., 2013**; **Raymond and Kim, 2012**; **Bhattacharya et al., 2013**; **Foflonker et al., 2018**; **Schönknecht et al., 2014**). For this reason, we chose the Cyanidiales as a model lineage for studying eukaryotic HGT.

It should be appreciated that the correct identification of HGTs based on sequence similarity and phylogeny is rarely trivial and unambiguous, leaving much space for interpretation and erroneous assignments. In this context, previous findings regarding HGT in Cyanidiales were based on single genome analyses and have therefore been questioned (**Ku and Martin, 2016**).

Many potential sources of error need to be excluded during HGT analysis, such as possible bacterial contamination in the samples, algorithmic errors during genome assembly and annotation, phylogenetic model misspecification, and unaccounted for gene paralogy (**Richards and Monier, 2016**). In addition, eukaryotic HGT reports based on single gene tree analysis are prone to misinterpretation and may be a product of deep branching artefacts and low genome sampling. Indeed, false claims of prokaryote-to-eukaryote HGT have been published (**Boothby et al., 2015**; **Crisp et al., 2015**) which were later corrected (**Koutsovoulos et al., 2016**; **Salzberg, 2017**).

Here, we used multi-genomic analysis with 13 Cyanidiales lineages (including 10 novel, long-read, draft genome sequences) from nine geographically isolated habitats. This approach increased phylogenetic resolution within Cyanidiales to allow more accurate assessment of HGT while avoiding many of the above-mentioned sources of error. The following questions were addressed by our research: (i) did HGT have a significant impact on Cyanidiales evolution? (ii) Are previous HGT findings in the sequenced Cyanidiales genomes an artefact of short read assemblies, limited genome databases, and uncertainties associated with single gene trees, or do they hold up with more extensive sampling? (iii) And, assuming that evidence of eukaryotic HGT is found across multiple Cyanidiales species, are cumulative effects observable, or is DL the better explanation for these results?

Table 1. Summary of the 13 analyzed Cyanidiales genomes.

The existing genomes of *Galdieria sulphuraria* 074W, *Cyanidioschyzon merolae* 10D, and *Galdieria phlegrea* are marked with '#'. The remaining 10 genomes are novel. Genome Size (Mb): size of the genome assembly in Megabases. Contigs: number of contigs produced by the genome assembly. The contigs were polished with quiver Contig N50 (kb): Contig N50. %GC Content: GC content of the genome given in percent. Genes: transcriptome size of species. Orthogroups: All Cyanidiales genes were clustered into a total of 9075 OGs. Here we show how many OGs there are per species. HGT Orthogroups: Number of OGs derived from HGT. HGT Genes: Number of HGT gene candidates found in species. %GC Native: GC content of the native transcriptome given in percent. %GC HGT: GC content of the HGT gene candidates given in percent % Multiexon Native: % of multiallelic genes in the native transcriptome. % Multiexon HGT: percent of multiallelic genes in the HGT gene candidates. S/M Native: Ratio of Multiexonic vs Singleexonic genes in native transcriptome. S/M HGT: Ratio of Multiexonic vs Singleexonic genes in HGT candidates. Asterisks (*) denote a significant difference (p<=0.05) between native and HGT gene subsets. EC, PFAM, GO, KEGG: Number of species-specific annotations in EC, PFAM, GO, KEGG.

	Genome features				Gene and OG counts		HGTs		HGT vs native gene subsets						Annotations			
Strain	Genome Size (Mb)	Contigs	Contig N50 (kb)	%GC Content	Genes	Ortho groups	HGT ortho groups	HGT genes	%GC Native	%GC HGT	(%) Multi exon Native	(%) Multi exon HGT	Exon/ Gene Native	Exon/ Gene HGT	EC	PFAM	KEGG	GO
G. sulphuraria 074W#	13.78	433	172.3	36.89	7174	5265	51	55	38.99	39.62*	73.6	47.3*	2.25	3.2*	938	3073	3241	6572
G. sulphuraria MS1	14.89	129	172.1	37.62	7441	5389	54	58	39.59	40.79*	83.4	62.1*	2.5	3.88*	930	3077	3178	6564
G. sulphuraria RT22	15.62	118	172.9	37.43	6982	5186	51	54	39.54	40.85*	74.7	51.9*	2.63	3.95*	941	3118	3223	6504
G. sulphuraria SAG21	14.31	135	158.2	37.92	5956	4732	44	47	40.04	41.47*	84.8	83.0	4.02	5.03*	931	3047	3143	6422
G. sulphuraria MtSh	14.95	101	186.6	40.04	6160	4746	46	47	41.33	42.48*	79.7	63.8*	3.15	4.32*	939	3114	3244	6450
G. sulphuraria Azora	14.06	127	162.3	40.10	6305	4905	49	58	41.34	42.57*	84.5	75.9*	2.68	4.03*	934	3072	3181	6474
G. sulphuraria YNP5587.1	14.42	115	170.8	40.05	6118	4846	46	46	41.33	42.14*	74.5	54.3*	2.61	3.65*	938	3084	3206	6516
G. sulphuraria 5572	14.28	108	229.7	37.99	6472	5009	46	53	39.68	40.5*	78.4	45.3*	2.15	3.53*	936	3108	3252	6540
G. sulphuraria 002	14.11	107	189.3	39.16	5912	4701	46	52	40.76	41.35*	97.1	50.0*	2.37	3.73*	927	3060	3184	6505
G. phlegrea DBV009#	11.41	9311	2.0	37.86	7836	5562	54	62	39.97	40.58*	na	na	na	na	935	3018	3125	6512
G. phlegrea Soos	14.87	108	201.1	37.52	6125	4624	44	47	39.57	40.73*	77.5	43.2*	2.19	3.33*	929	3034	3197	6493
C. merolae 10D#	16.73	22	859.1	54.81	4803	3980	33	33	56.57	56.57	0.5	0.0	1	1.01	883	2811	2832	6213
C. merolae Soos	12.33	35	567.5	54.33	4406	3574	34	34	54.84	54.26	9.4	2.9	1.06	1.1	886	2787	2823	6188

DOI: https://doi.org/10.7554/eLife.45017.003

Results

Features of the newly sequenced cyanidiales genomes

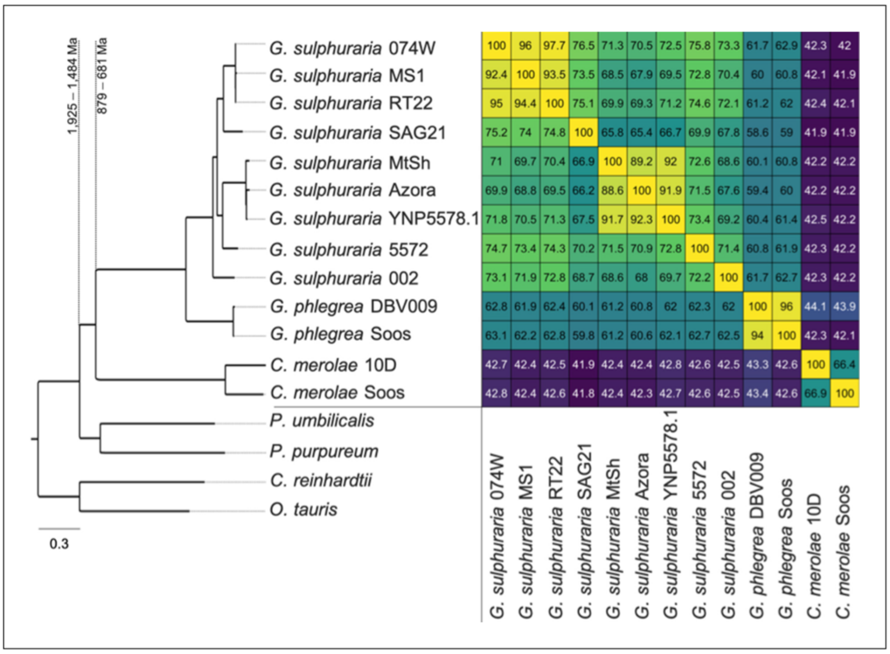
Genome sizes of the 10 targeted Cyanidiales (***Figure 1***) range from 12.33 Mbp - 15.62 Mbp, similar to other members of this red algal lineage (***Matsuzaki et al., 2004***; ***Schönknecht et al., 2013***; ***Qiu et al., 2013***) (***Table 1***). PacBio sequencing yielded 0.56 Gbp – 1.42 Gbp of raw sequence reads with raw read N50 ranging from 7.9 kbp – 14.4 kbp, which translated to a coverage of 28.91x – 70.99x at the unitigging stage (39.46x – 91.20x raw read coverage) (Appendix 1). We predicted a total of 61,869 novel protein coding sequences which, together with the protein data sets of the already published Cyanidiales species (total of 81,682 predicted protein sequences), capture 295/303 (97.4%) of the highly conserved eukaryotic BUSCO dataset. Each species, taken individually, scored an average of 92.7%. In spite of massive gene losses observed in the Cyanidiales (***Qiu et al., 2015***), these results corroborate previous observations that genome reduction has had a minor influence on the core eukaryotic gene inventory in free-living organisms (***Qiu et al., 2016***). Even *C. merolae* Soos, the species with the most limited coding capacity (4406 protein sequences), includes 89.5% of the eukaryotic BUSCO dataset. The number of contigs obtained from the *Galdieria* genomes ranged between 101–135. *G. sulphuraria* 17.91 (a strain different from the ones sequenced) was reported to have 40 chromosomes, and strains isolated from Rio Tinto (Spain), 47 or 57 chromosomes (***Moreira et al., 1994***). Pulsed-field gel electrophoresis indicates that *G. sulphuraria* 074W has approximately 42 chromosomes that are between 100 kbp and 1 Mbp in size (***Weber, 2007***). The genome assembly of *C. merolae* Soos produced 35 contigs, which approximates the 22 chromosomes (including plastid and mitochondrion) of the *C. merolae* 10D telomere-to-telomere assembly. Whole genome alignments indicate that a portion of the assembled contigs represent complete chromosomes.

Orthogroups and phylogeny

The 81,682 predicted protein sequences from all 13 genomes clustered into a total of 9075 orthogroups and phylogenetic trees were built for each orthogroup. The reference species tree was constructed using 2,090 OGs that contained a single-copy gene in at least 12 of the 17 taxa (*Porphyra umbilicalis* (***Brawley et al., 2017***), *Porphyridium purpureum* (***Bhattacharya et al., 2013***), *Ostreococcus tauri* RCC4221 (***Blanc-Mathieu et al., 2014***), and *Chlamydomonas reinhardtii* (***Merchant et al., 2007***) were added to the dataset as outgroups). As a result, the species previously named *G. sulphuraria* Soos and *C. merolae* MS1 were reannotated as *G. phlegrea* Soos and *G. sulphuraria* MS1. Given these results, we sequenced a second genome of *C. merolae* and a representative of the *G. phlegrea* lineage. The species tree reflects previous findings that suggest more biodiversity exists within the Cyanidiales (***Ciniglia et al., 2004***) (***Figure 2***).

Analysis of HGTs

The most commonly used approach to identify HGT candidates is to determine the position of eukaryotic and non-eukaryotic sequences in a maximum likelihood tree. Using this approach, 96 OGs were identified in which Cyanidiales genes shared a monophyletic descent with prokaryotes, representing 1.06% of all OGs. A total of 641 individual Cyanidiales sequences are considered as HGT candidates (***Table 1***). The amount of HGT per species varied considerably between members of the *Cyanidioschyzon* (33–34 HGT events, all single copy genes) and *Galdieria* lineages with 44–54 HGT events (52.6 HGT origins on average, 47–62 HGT gene candidates). In comparison to previous studies (***Schönknecht et al., 2013***; ***Qiu et al., 2013***), no evidence of massive gene family expansion regarding HGT genes was found because the maximum number of gene copies in HGT orthogroups was three. We note, however, that one large gene family of STAND-type ATPases that was previously reported to originate from an archaeal HGT (***Schönknecht et al., 2013***) did not meet the criteria used in our restrictive Blast searches; that is the 10 ^{5} *e*-value cut-off for consideration and a minimum of three different non-eukaryotic donors. This highly diverged family requires more sophisticated comparative analyses that were not done here (Appendix 2).



	G. sulphuraria 074W	G. sulphuraria MS1	G. sulphuraria RT22	G. sulphuraria SAG21	G. sulphuraria MtSh	G. sulphuraria Azora	G. sulphuraria YNP5578.1	G. sulphuraria 5572	G. sulphuraria 002	G. phlegrea DBV009	G. phlegrea Soos	C. merolae 10D	C. merolae Soos
G. sulphuraria 074W	100	96	97.7	76.5	71.3	70.5	72.5	75.8	73.3	61.7	62.9	42.3	42
G. sulphuraria MS1	92.4	100	93.5	73.5	68.5	67.9	69.5	72.8	70.4	60	60.8	42.1	41.9
G. sulphuraria RT22	95	94.4	100	75.1	69.9	69.3	71.2	74.6	72.1	61.2	62	42.4	42.1
G. sulphuraria SAG21	75.2	74	74.8	100	65.8	65.4	66.7	69.9	67.8	58.6	59	41.9	41.9
G. sulphuraria MtSh	71	69.7	70.4	66.9	100	89.2	92	72.6	68.6	60.1	60.8	42.2	42.2
G. sulphuraria Azora	69.9	68.8	69.5	66.2	88.6	100	91.9	71.5	67.6	59.4	60	42.2	42.2
G. sulphuraria YNP5578.1	71.8	70.5	71.3	67.5	91.7	92.3	100	73.4	69.2	60.4	61.4	42.5	42.2
G. sulphuraria 5572	74.7	73.4	74.3	70.2	71.5	70.9	72.8	100	71.4	60.8	61.9	42.3	42.2
G. sulphuraria 002	73.1	71.9	72.8	68.7	68.6	68	69.7	72.2	100	61.7	62.7	42.3	42.2
G. phlegrea DBV009	62.8	61.9	62.4	60.1	61.2	60.8	62	62.3	62	100	96	44.1	43.9
G. phlegrea Soos	63.1	62.2	62.8	59.8	61.2	60.6	62.1	62.7	62.5	94	100	42.3	42.1
C. merolae 10D	42.7	42.4	42.5	41.9	42.4	42.4	42.8	42.6	42.5	43.3	42.6	100	66.4
C. merolae Soos	42.8	42.4	42.6	41.8	42.4	42.3	42.7	42.6	42.5	43.4	42.6	66.9	100

Figure 2. Species tree of the 13 analyzed extremophilic Cyanidiales genomes using mesophilic red (*Porphyra umbilicalis, Porphyridium purpureum*) and green algae (*Ostreococcus tauri*, *Chlamydomonas reinhardtii*) as outgroups. IQTREE was used to infer a single maximum-likelihood phylogeny based on orthogroups containing single-copy representative proteins from at least 12 of the 17 taxa (13 Cyanidiales + 4 Others). Each orthogroup alignment represented one partition with unlinked models of protein evolution chosen by IQTREE. Consensus tree branch support was determined by 2000 rapid bootstraps. All nodes in this tree had 100% bootstrap support, and are therefore not shown. Divergence time estimates are taken from ***Yang et al. (2016)***. Similarity is derived from the average one-way best blast hit protein identity (minimum protein identity threshold = 30%). The minimal protein identity between two strains was 65.4%, measured between *g. sulphuraria* SAG21.92, which represent the second most distant sampling locations (12,350 km). Similar lineage boundaries were obtained for the *C. merolae* samples (66.4% protein identity), which are separated by only 1150 km.
DOI: https://doi.org/10.7554/eLife.45017.004

Gene co-localization on raw sequence reads

One major issue associated with previous HGT studies is the incorporation of contaminant DNA into the genome assembly, leading to incorrect results (***Boothby et al., 2015***; ***Crisp et al., 2015***; ***Koutsovoulos et al., 2016***; ***Salzberg, 2017***). Here, we screened for potential bacterial contamination in our tissue samples using PCR analysis of extracted DNA with the *rbcL* and 18S rRNA gene markers prior to sequencing. No instances of contamination were found. Furthermore, our work relied on PacBio RSII long-read sequencing technology, whereby single reads frequently exceed 10 kbp of DNA. Given these robust data, we also tested for co-occurrence of HGT gene candidates and 'native' genes in the same read. The protein sequences of each species were queried with tblastn (10^{-5} *e*-value, 75 bitscore) against a database consisting of the uncorrected PacBio RSII long reads. This analysis showed that 629/641 (98.12%) of the HGT candidates co-localize with native red algal genes on the same read (38,297 reads in total where co-localization of native genes and HGT candidates was observed). It should be noted that the 10 novel genomes we determined share HGT candidates with *C. merolae* 10D, *G. sulphuraria* 074W, and *G. phlegrea* DBV009, which were sequenced in different laboratories, at different points in time, using different technologies, and assembly pipelines. Hence, we consider it highly unlikely that these HGT candidates result from bacterial contamination. As the accuracy of long read sequencing technologies further increases, we believe this criterion for excluding bacterial contamination provides an additional piece of evidence that should be added to the guidelines for HGT discovery (***Richards and Monier, 2016***).

Differences in molecular features between native and HGT-derived genes

One of the main consequences of HGT is that horizontally acquired genes may have different structural characteristics when compared to native genes (cumulative effects). HGT-derived genes initially retain characteristics of the genome of the donor lineage. Consequently, the passage of time is required (and expected) to erase these differences. Therefore, we searched for differences in genomic features between HGT candidates and native Cyanidiales genes with regard to: (1) GC-content, (2) the number of spliceosomal introns and the exon/gene ratio, (3) differential transcription, (4) percent protein identity between HGT genes and their non-eukaryotic donors, and (5) cumulative effects as indicators of their non-eukaryotic origin (**Danchin, 2016**; **Ku and Martin, 2016**; **Schönknecht et al., 2013**).

GC-content

All 11 *Galdieria* species showed significant differences (GC-content of transcripts is normally distributed, Student's *t*-test, two-sided, $p \leq 0.05$) in percent GC-content between native sequences and HGT candidates (**Table 1**). Sequences belonging to the *Galdieria* lineage have an exceptionally low GC-content (39%–41%) in comparison to the majority of thermophilic organisms that exhibit higher values (~55%). On average, HGT candidates in *Galdieria* display 1% higher GC-content in comparison to their native counterparts. No significant differences were found for *C. merolae* 10D and *C. merolae* Soos in this respect. Because native *Cyanidioschyzon* genes have an elevated GC-content (54%–56%), this makes it difficult to distinguish between them and HGT-derived genes (Appendix 3).

Spliceosomal introns and exon/Gene

Bacterial genes lack spliceosomal introns and therefore the spliceosomal machinery. Consequently, genes acquired through HGT are initially single-exons and may acquire introns over time due to the invasion of existing intervening sequences. We detected significant discrepancies in the ratio of single-exon to multi-exon genes between HGT candidates and native genes in the *Galdieria* lineage. On average, 42% of the *Galdieria* HGT candidates are single-exon genes, whereas only 19.2% of the native gene set are comprised of single-exons. This difference is significant (categorical data, 'native' vs 'HGT' and 'single exon' vs. 'multiple exon', Fisher's exact test, $p \leq 0.05$) in all *Galdieria* species except *G. sulphuraria* SAG21.92 (**Table 1**). The *Cyanidioschyzon* lineage contains a highly reduced spliceosomal machinery (**Qiu et al., 2018**), therefore only ~10% of native genes are multi-exonic in *C. merolae* Soos and only 1/34 HGT candidates has gained an intron. *C. merolae* 10D has only 26 multi-exonic genes (~0.5% of all transcripts) and none of its HGT candidates has gained an intron. Enrichment testing is not possible with these small sample sizes (Appendix 4).

We analyzed the number of exons that are present in multi-exonic genes and obtained similar results for the *Galdieria* lineage (**Table 1**). All *Galdieria* species show significant differences regarding the exon/gene ratio between native and HGT genes (non-normal distribution regarding the number of exons per gene, Wilcoxon-Mann-Whitney-Test, 1000 bootstraps, $p <= 0.05$). HGT candidates in *Galdieria* have 0.97–1.36 fewer exons per gene in comparison to their native counterparts. Because the multi-exonic HGT subset in both *Cyanidioschyzon* species combined includes only one multi-exonic HGT candidate, no further analysis was performed (Appendix 4).

Differential transcription

Several RNA-Seq datasets are publicly available for *G. sulphuraria* 074W (**Rossoni, 2018**) and *C. merolae* 10D (**Rademacher et al., 2016**). We aligned (**Kim et al., 2015**) the transcriptome reads to the respective genomes, using an identical data processing pipeline (**Robinson et al., 2010**) for both datasets to exclude potential algorithmic errors (**Figure 3**). The average read count per gene (measured as counts per million, CPM), of native genes was 154 CPM in *G. sulphuraria* 074W and 196 CPM *C. merolae* 10D. The average read counts for HGT candidates in *G. sulphuraria* 074W and *C. merolae* 10D were 130 CPM and 184 CPM, respectively. No significant differences in RNA abundance between native genes and HGT candidates were observed for these taxa (non-normal distribution of CPM, Wilcoxon-Mann-Whitney-Test, $p < 0.05$).

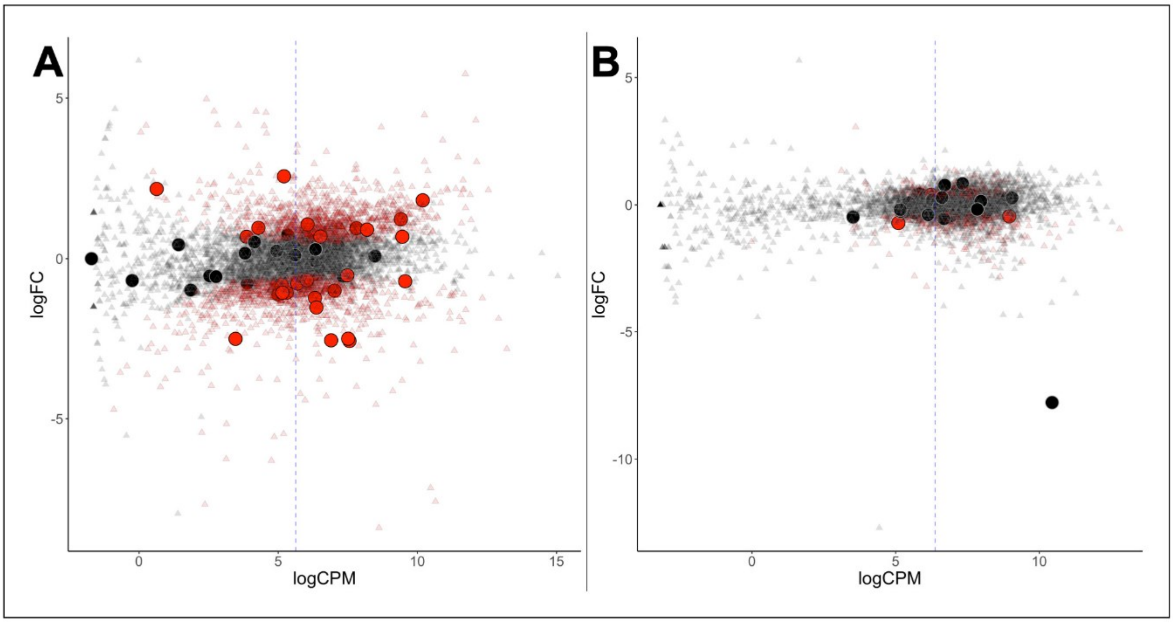


Figure 3. Differential gene expression of *G. sulphuraria* 074W. (**A**) and *C. merolae* 10D (**B**), here measured as log fold change (logFC) vs transcription rate (logCPM). Differentially expressed genes are colored red (quasi-likelihood (QL) F-test, Benjamini-Hochberg, $p <= 0.01$). HGT candidates are shown as large circles. The blue dashes indicate the average logCPM of the dataset. Although HGT candidates are not significantly more or less expressed than native genes, they react significantly stronger to temperature changes in *G. sulphuraria* 074W ('more red than black dots'). This is not the case in high CO_2 treated *C. merolae* 10D.

DOI: https://doi.org/10.7554/eLife.45017.005

Gene function – not passage of time – explains percent protein identity (PID) between Cyanidiales HGT candidates and their non-eukaryotic donors

Once acquired, any HGT-derived gene may be fixed in the genome and propagated across the lineage. The PID data can be further divided into different subsets depending on species composition of the OG. Of the total 96 OGs putatively derived from HGT events, 60 are exclusive to the *Galdieria* lineage (62.5%), 23 are exclusive to the *Cyanidioschyzon* lineage (24%), and 13 are shared by both lineages (13.5%) (***Figure 4A***). Consequently, either a strong prevalence for lineage specific DL exists, or both lineages underwent individual sets of HGT events because they share their habitat with other non-eukaryotic species (which is what the HGT theory would assume). The 96 OGs in question are affected by gene loss or partial fixation. Once acquired only 8/13 of the 'Cyanidiales' (including 'Multiple HGT' and 'Uncertain') OGs and 20/60 of the *Galdieria* specific OGs are encoded by all species. Once acquired by the *Cyanidioschyzon* ancestor, the HGT candidates were retained by both *C. merolae* Soos and *C. merolae* 10D in 22/23 *Cyanidioschyzon* specific OGs. It is not possible to verify whether the only *Cyanidioschyzon* OG containing one HGT candidate is the result of gene loss, individual acquisition, or due to erroneously missing this gene model during gene prediction. The average percent PID between HGT gene candidates of the 13 OGs shared by all Cyanidiales and their non-eukaryotic donors is 41.2% (min = 24.4%; max = 65.4%) (***Figure 4B***). From the HGT perspective, these OGs are derived from ancient HGT events that occurred at the root of the Cyanidiales, well before the split of the *Galdieria* and *Cyanidioschyzon* lineages. The OGs were retained over time in all Cyanidiales, although evidence of subsequent gene loss is observed. Under the DL hypothesis, this group of OGs contains genes that have been lost in all other eukaryotic lineages except the Cyanidiales. Similarly, the average PID between HGT candidates their non-eukaryotic donors in OGs

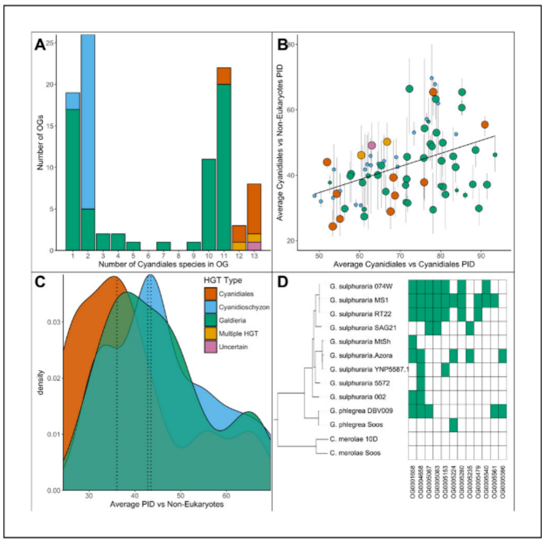


Figure 4. Comparative analysis of the 96 OGs potentially derived from HGT. (**A**) OG count vs. the number of Cyanidiales species contained in an OG (=OG size). Only genes from the sequenced genomes were considered (13 species). A total of 60 OGs are exclusive to the *Galdieria* lineage (11 species), 23 OGs are exclusive to the *Cyanidioschyzon* lineage (two species), and 13 OGs are shared by both lineages. A total of 46/96 HGT events seem to be affected by later gene erosion/partial fixation. (**B**) OG-wise PID between HGT candidates vs. their potential non-eukaryotic donors. Point size represents the number of sequenced species contained in each OG. Because only two genomes of *Cyanidioschyzon* were sequenced, the maximum point size for this lineage is 2. The whiskers span minimum and maximum shared PID of each OG. The PID within Cyanidiales HGTs vs. PID between Cyanidiales HGTs and their potential non-eukaryotic donors is positively correlated (Kendall's tau coefficient, p=0.000747), showing evolutionary constraints that are gene function dependent, rather than time-dependent. (**C**) Density curve of average PID towards potential non-eukaryotic donors. The area under each curve is equal to 1. The average PID of HGT candidates found in both lineages ('ancient HGT', left dotted line) is ~5% lower than the average PID of HGT candidates exclusive to *Galdieria* or *Cyandioschyzon* ('recent HGT', right dotted lines). This difference is not significant (pairwise Wilcoxon rank-sum test, Benjamini-Hochberg, p>0.05). (**D**) Presence/Absence pattern (green/white) of Cyanidiales species in HGT OGs. Some patterns strictly follow the branching structure of the species tree. They represent either recent HGTs that affect a monophyletic subset of the *Galdieria* lineage, or are the last eukaryotic remnants of an ancient gene that was eroded through differential loss. In other cases, the presence/absence pattern of *Galdieria* species is random and conflicts with the *Galdieria* lineage phylogeny. HGT would assume either multiple independent acquisitions of the same HGT candidate, or a partial fixation of the HGT candidate in the lineage, while still allowing for gene erosion. According to DL, these are the last existing paralogs of an ancient gene, whose erosion within the eukaryotic kingdom is nearly complete.

DOI: https://doi.org/10.7554/eLife.45017.006

exclusive to the *Cyanidioschyzon* lineage is 46.4% (min = 30.8%; max = 69.7%) and 45.1% (min = 27.4%; max = 69.5%) for those OGs exclusive to the *Galdieria* lineage. According to the HGT view, these subsets of candidates were horizontally acquired either in the *Cyanidioschyzon* lineage, or in the *Galdieria* lineage after the split between *Galdieria* and *Cyanidioschyzon*. DL would impose gene loss on all other eukaryotic lineages except *Galdieria* or *Cyanidioschyzon*. Over time, sequence similarity between the HGT candidate and the non-eukaryotic donor is expected to decrease at a rate that reflects the level of functional constraint. The average PID of 'ancient' HGT candidates shared by both lineages (before the split into *Galdieria* and *Cyanidioschyzon* approximately 800 Ma years ago [***Yang et al., 2016***]) is ~5% lower than the average PID of HGT candidates exclusive to one lineage which, according to HGT would represent more recent HGT events because their acquisition occurred only after the split (thus lower divergence) (***Figure 4C***). However, no significant difference between *Galdieria*-exclusive HGTs, *Cyandioschyzon*-exclusive HGTs, and HGTs shared by both lineages was found (non-normal distribution of percent protein identity, Shapiro-Wilk normality test, W = 0.95, p=0.002; Pairwise Wilcoxon rank-sum test, Benjamini-Hochberg, all comparisons p>0.05). Therefore, neither *Cyanidioschyzon* nor *Galdieria* specific HGTs, or HGTs shared by all Cyanidiales, are significantly more, or less, similar to their potential prokaryotic donors. We also addressed the differences in PID within the three groups. The average PID within HGT gene candidates of the 13 OGs shared by all Cyanidiales is 75.0% (min = 51.9%; max = 90.9%) (**Figure 4B**). Similarly, the average PID within HGT candidates in OGs exclusive to the *Cyanidioschyzon* lineage is 65.1% (min = 48.9%; max = 83.8%) and 75.0% (min = 52.6%; max = 93.4%) for those OGs exclusive to the *Galdieria* lineage. Because we sampled only two *Cyanidioschyzon* species in comparison to 11 *Galdieria* lineages that are also much more closely related (**Figure 2A**), a comparison between these two groups was not done. However, a significant positive correlation (non-normal distribution of PID across all OGs, Kendall's tau coefficient, p=0.000747) exists between the PID within Cyanidiales HGTs versus PID between Cyanidiales HGTs and their non-eukaryotic donors (**Figure 4B**). Hence, the more similar Cyanidiales sequences are to each other, the more similar they are to their non-eukaryotic donors, showing gene function dependent evolutionary constraints.

Complex origins of HGT-impacted orthogroups

While comparing the phylogenies of HGT candidates, we also noted that not all Cyanidiales genes within one OG necessarily originate via HGT (phylogenetic trees of each HGT-OG are included in **Figure 5—figure supplements 1–96**). Among the 13 OGs that contain HGT candidates present in both *Galdieria* and *Cyanidioschyzon*, we found two cases (**Figure 4A**, 'Multiple HGT'), OG0002305 and OG0003085, in which *Galdieria* and *Cyanidioschyzon* HGT candidates cluster in the same orthogroup. However, these have different non-eukaryotic donors and are located on distinct phylogenetic branches that do not share a monophyletic descent (**Figure 5A**). This is potentially the case for OG0002483 as well, but we were uncertain due to low bootstrap values (**Figure 4A**, 'Uncertain'). These OGs either represent two independent acquisitions of the same function or, according to DL, the LECA encoded three paralogs of the same gene which were propagated through evolutionary time. One of these was retained by the *Galdieria* lineage (and shares sequence similarity with one group of prokaryotes), the second was retained by *Cyanidioschyzon* (and shares sequence similarity with a different group of prokaryotes), and a third paralog was retained by all other eukaryotes. It should be noted that the 'other eukaryotes' do not always cluster in one uniformly eukaryotic clade which increases the number of required paralogs in LECA to explain the current pattern. Furthermore, some paralogs could also have already been completely eroded and do not exist in extant eukaryotes. Similarly, 6/60 *Galdieria* specific OGs also contain *Cyanidioschyzon* genes (OG0001929, OG0001938, OG0002191, OG0002574, OG0002785 and OG0003367). Here, they are nested within other eukaryote lineages and would not be derived from HGT (**Figure 5B**). Also, eight of the 23 *Cyanidioschyzon* specific HGT OGs contain genes from *Galdieria* species (OG0001807, OG0001810, OG0001994, OG0002727, OG0002871, OG0003539, OG0003929 and OG0004405) which cluster within the eukaryotic branch and are not monophyletic with *Cyanidioschyzon* HGT candidates (**Figure 5C**). According to the HGT view, this subset of candidates was horizontally acquired in either the *Cyanidioschyzon* lineage, or the *Galdieria* lineage only after the split between *Galdieria* and *Cyanidioschyzon*, possibly replacing the ancestral gene or functionally complementing a function that was lost due to genome reduction. According to DL, the LECA would have encoded two paralogs of the same gene. One was retained by all eukaryotes, red algae, and *Galdieria* or *Cyanidioschyzon*, the other exclusively by *Cyanidioschyzon* or *Galdieria* together with non-eukaryotes.

Stronger erosion of HGT genes impedes assignment to HGT or DL

As already noted above, only 50/96 of the sampled HGT-impacted OGs do not appear to be affected by erosion. Dense sampling of 11 taxa within the *Galdieria* lineage allowed a more in-depth analysis of this issue. Here, a bimodal distribution is observed regarding the number of species per OG in the native and HGT dataset (**Figure 6C**). Only 52.5% of the native gene set is present in all *Galdieria* strains (defined as 10 and 11 strains in order to account for potential misassemblies and missed gene models during prediction). Approximately 1/3 of the native OGs (36.1%) has been affected by gene erosion to such a degree that it is present in only one, or two *Galdieria* strains. In comparison, 26.7% of the candidate HGT-impacted OGs are encoded in >10 *Galdieria* strains,

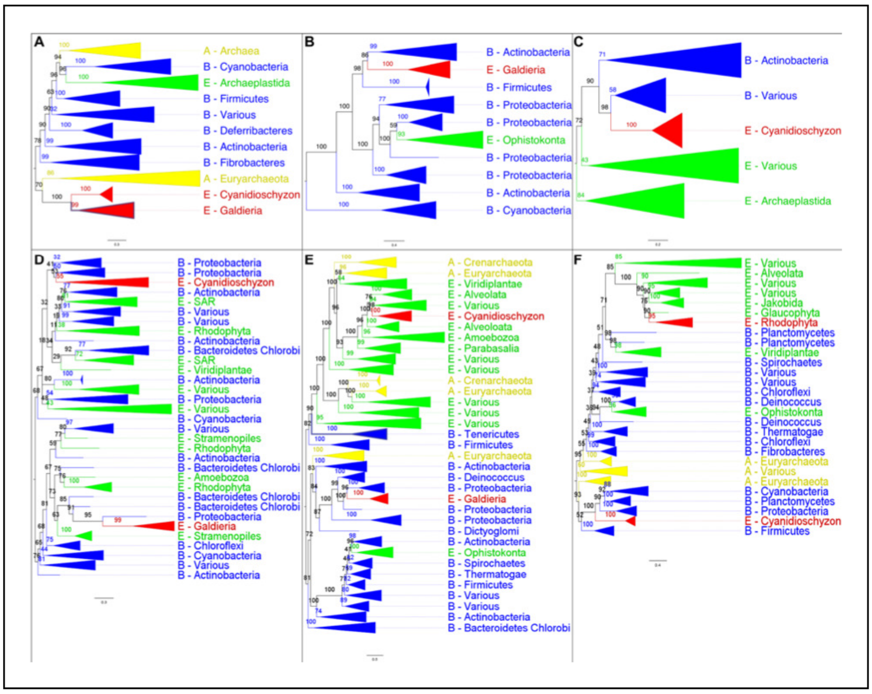


Figure 5. The analysis of OGs containing HGT candidates revealed different patterns of HGT acquisition. Some OGs contain genes that are shared by all Cyanidiales, whereas others are unique to the *Galdieria* or *Cyanidioschyzon* lineage. In some cases, HGT appears to have replaced the eukaryotic genes in one lineage, whereas the other lineage maintained the eukaryotic ortholog. Here, some examples of OG phylogenies are shown, which were simplified for ease of presentation. The first letter of the tip labels indicates the kingdom. A = Archaea (yellow), B = Bacteria (blue), E = Eukaryota (green). Branches containing Cyanidiales sequences are highlited in red. (**A**) Example of an ancient HGT that occurred before *Galdieria* and *Cyanidioschyzon* split into separate lineages. As such, both lineages are monophyletic (e.g., OG0001476). (**B**) HGT candidates are unique to the *Galdieria* lineage (e.g. OG0001760). (**C**) HGT candidates are unique to the *Cyanidioschyzon* lineage (e.g. OG0005738). (**D**) *Galdieria* and *Cyanidioschyzon* HGT candidates are derived from different HGT events and share monophyly with different non-eukaryotic organisms (e.g., OG0003085). (**E**) *Galdieria* HGT candidates cluster with non-eukaryotes, whereas the *Cyanidioschyzon* lineage clusters with eukaryotes (e.g., OG0001542). (**F**) *Cyanidioschyzon* HGT candidates cluster with non-eukaryotes, whereas the *Galdieria* lineage clusters with eukaryotes (e.g., OG0006136).

DOI: https://doi.org/10.7554/eLife.45017.007

The following figure supplements are available for figure 5:

Figure supplement 1. Sequence tree of orthogroup OG0001476.

DOI: https://doi.org/10.7554/eLife.45017.008

Figure supplement 2. Sequence tree of orthogroup OG0001486.

DOI: https://doi.org/10.7554/eLife.45017.009

Figure supplement 3. Sequence tree of orthogroup OG0001509.

DOI: https://doi.org/10.7554/eLife.45017.010

Figure supplement 4. Sequence tree of orthogroup OG0001513.

DOI: https://doi.org/10.7554/eLife.45017.011

Figure supplement 5. Sequence tree of orthogroup OG0001542.

DOI: https://doi.org/10.7554/eLife.45017.012

Figure supplement 6. Sequence tree of orthogroup OG0001613.

DOI: https://doi.org/10.7554/eLife.45017.013

Figure supplement 7. Sequence tree of orthogroup OG0001658.

Figure 5 continued on next page

Figure 5 continued

DOI: https://doi.org/10.7554/eLife.45017.014

Figure supplement 8. Sequence tree of orthogroup OG0001760.

DOI: https://doi.org/10.7554/eLife.45017.015

Figure supplement 9. Sequence tree of orthogroup OG0001807.

DOI: https://doi.org/10.7554/eLife.45017.016

Figure supplement 10. Sequence tree of orthogroup OG0001810.

DOI: https://doi.org/10.7554/eLife.45017.017

Figure supplement 11. Sequence tree of orthogroup OG0001929.

DOI: https://doi.org/10.7554/eLife.45017.018

Figure supplement 12. Sequence tree of orthogroup OG0001938.

DOI: https://doi.org/10.7554/eLife.45017.019

Figure supplement 13. Sequence tree of orthogroup OG0001955.

DOI: https://doi.org/10.7554/eLife.45017.020

Figure supplement 14. Sequence tree of orthogroup OG0001976.

DOI: https://doi.org/10.7554/eLife.45017.021

Figure supplement 15. Sequence tree of orthogroup OG0001994.

DOI: https://doi.org/10.7554/eLife.45017.022

Figure supplement 16. Sequence tree of orthogroup OG0002036.

DOI: https://doi.org/10.7554/eLife.45017.023

Figure supplement 17. Sequence tree of orthogroup OG0002051.

DOI: https://doi.org/10.7554/eLife.45017.024

Figure supplement 18. Sequence tree of orthogroup OG0002191.

DOI: https://doi.org/10.7554/eLife.45017.025

Figure supplement 19. Sequence tree of orthogroup OG0002305.

DOI: https://doi.org/10.7554/eLife.45017.026

Figure supplement 20. Sequence tree of orthogroup OG0002337.

DOI: https://doi.org/10.7554/eLife.45017.027

Figure supplement 21. Sequence tree of orthogroup OG0002431.

DOI: https://doi.org/10.7554/eLife.45017.028

Figure supplement 22. Sequence tree of orthogroup OG0002483.

DOI: https://doi.org/10.7554/eLife.45017.029

Figure supplement 23. Sequence tree of orthogroup OG0002574.

DOI: https://doi.org/10.7554/eLife.45017.030

Figure supplement 24. Sequence tree of orthogroup OG0002578.

DOI: https://doi.org/10.7554/eLife.45017.031

Figure supplement 25. Sequence tree of orthogroup OG0002609.

DOI: https://doi.org/10.7554/eLife.45017.032

Figure supplement 26. Sequence tree of orthogroup OG0002676.

DOI: https://doi.org/10.7554/eLife.45017.033

Figure supplement 27. Sequence tree of orthogroup OG0002727.

DOI: https://doi.org/10.7554/eLife.45017.034

Figure supplement 28. Sequence tree of orthogroup OG0002785.

DOI: https://doi.org/10.7554/eLife.45017.035

Figure supplement 29. Sequence tree of orthogroup OG0002871.

DOI: https://doi.org/10.7554/eLife.45017.036

Figure supplement 30. Sequence tree of orthogroup OG0002896.

DOI: https://doi.org/10.7554/eLife.45017.037

Figure supplement 31. Sequence tree of orthogroup OG0002999.

DOI: https://doi.org/10.7554/eLife.45017.038

Figure supplement 32. Sequence tree of orthogroup OG0003085.

DOI: https://doi.org/10.7554/eLife.45017.039

Figure supplement 33. Sequence tree of orthogroup OG0003250.

DOI: https://doi.org/10.7554/eLife.45017.040

Figure supplement 34. Sequence tree of orthogroup OG0003367.

DOI: https://doi.org/10.7554/eLife.45017.041

Figure supplement 35. Sequence tree of orthogroup OG0003441.

Figure 5 continued on next page

Figure 5 continued

DOI: https://doi.org/10.7554/eLife.45017.042

Figure supplement 36. Sequence tree of orthogroup OG0003539.

DOI: https://doi.org/10.7554/eLife.45017.043

Figure supplement 37. Sequence tree of orthogroup OG0003777.

DOI: https://doi.org/10.7554/eLife.45017.044

Figure supplement 38. Sequence tree of orthogroup OG0003782.

DOI: https://doi.org/10.7554/eLife.45017.045

Figure supplement 39. Sequence tree of orthogroup OG0003834.

DOI: https://doi.org/10.7554/eLife.45017.046

Figure supplement 40. Sequence tree of orthogroup OG0003846.

DOI: https://doi.org/10.7554/eLife.45017.047

Figure supplement 41. Sequence tree of orthogroup OG0003856.

DOI: https://doi.org/10.7554/eLife.45017.048

Figure supplement 42. Sequence tree of orthogroup OG0003901.

DOI: https://doi.org/10.7554/eLife.45017.049

Figure supplement 43. Sequence tree of orthogroup OG0003905.

DOI: https://doi.org/10.7554/eLife.45017.050

Figure supplement 44. Sequence tree of orthogroup OG0003907.

DOI: https://doi.org/10.7554/eLife.45017.051

Figure supplement 45. Sequence tree of orthogroup OG0003929.

DOI: https://doi.org/10.7554/eLife.45017.052

Figure supplement 46. Sequence tree of orthogroup OG0003954.

DOI: https://doi.org/10.7554/eLife.45017.053

Figure supplement 47. Sequence tree of orthogroup OG0004030.

DOI: https://doi.org/10.7554/eLife.45017.054

Figure supplement 48. Sequence tree of orthogroup OG0004102.

DOI: https://doi.org/10.7554/eLife.45017.055

Figure supplement 49. Sequence tree of orthogroup OG0004142.

DOI: https://doi.org/10.7554/eLife.45017.056

Figure supplement 50. Sequence tree of orthogroup OG0004203.

DOI: https://doi.org/10.7554/eLife.45017.057

Figure supplement 51. Sequence tree of orthogroup OG0004258.

DOI: https://doi.org/10.7554/eLife.45017.058

Figure supplement 52. Sequence tree of orthogroup OG0004339.

DOI: https://doi.org/10.7554/eLife.45017.059

Figure supplement 53. Sequence tree of orthogroup OG0004392.

DOI: https://doi.org/10.7554/eLife.45017.060

Figure supplement 54. Sequence tree of orthogroup OG0004405.

DOI: https://doi.org/10.7554/eLife.45017.061

Figure supplement 55. Sequence tree of orthogroup OG0004486.

DOI: https://doi.org/10.7554/eLife.45017.062

Figure supplement 56. Sequence tree of orthogroup OG0004658.

DOI: https://doi.org/10.7554/eLife.45017.063

Figure supplement 57. Sequence tree of orthogroup OG0005083.

DOI: https://doi.org/10.7554/eLife.45017.064

Figure supplement 58. Sequence tree of orthogroup OG0005087.

DOI: https://doi.org/10.7554/eLife.45017.065

Figure supplement 59. Sequence tree of orthogroup OG0005153.

DOI: https://doi.org/10.7554/eLife.45017.066

Figure supplement 60. Sequence tree of orthogroup OG0005224.

DOI: https://doi.org/10.7554/eLife.45017.067

Figure supplement 61. Sequence tree of orthogroup OG0005235.

DOI: https://doi.org/10.7554/eLife.45017.068

Figure supplement 62. Sequence tree of orthogroup OG0005280.

DOI: https://doi.org/10.7554/eLife.45017.069

Figure supplement 63. Sequence tree of orthogroup OG0005479.

Figure 5 continued on next page

Figure 5 continued

DOI: https://doi.org/10.7554/eLife.45017.070
Figure supplement 64. Sequence tree of orthogroup OG0005540.
DOI: https://doi.org/10.7554/eLife.45017.071
Figure supplement 65. Sequence tree of orthogroup OG0005561.
DOI: https://doi.org/10.7554/eLife.45017.072
Figure supplement 66. Sequence tree of orthogroup OG0005596.
DOI: https://doi.org/10.7554/eLife.45017.073
Figure supplement 67. Sequence tree of orthogroup OG0005683.
DOI: https://doi.org/10.7554/eLife.45017.074
Figure supplement 68. Sequence tree of orthogroup OG0005694.
DOI: https://doi.org/10.7554/eLife.45017.075
Figure supplement 69. Sequence tree of orthogroup OG0005738.
DOI: https://doi.org/10.7554/eLife.45017.076
Figure supplement 70. Sequence tree of orthogroup OG0005963.
DOI: https://doi.org/10.7554/eLife.45017.077
Figure supplement 71. Sequence tree of orthogroup OG0005984.
DOI: https://doi.org/10.7554/eLife.45017.078
Figure supplement 72. Sequence tree of orthogroup OG0006136.
DOI: https://doi.org/10.7554/eLife.45017.079
Figure supplement 73. Sequence tree of orthogroup OG0006143.
DOI: https://doi.org/10.7554/eLife.45017.080
Figure supplement 74. Sequence tree of orthogroup OG0006191.
DOI: https://doi.org/10.7554/eLife.45017.081
Figure supplement 75. Sequence tree of orthogroup OG0006251.
DOI: https://doi.org/10.7554/eLife.45017.082
Figure supplement 76. Sequence tree of orthogroup OG0006252.
DOI: https://doi.org/10.7554/eLife.45017.083
Figure supplement 77. Sequence tree of orthogroup OG0006435.
DOI: https://doi.org/10.7554/eLife.45017.084
Figure supplement 78. Sequence tree of orthogroup OG0006482.
DOI: https://doi.org/10.7554/eLife.45017.085
Figure supplement 79. Sequence tree of orthogroup OG0006498.
DOI: https://doi.org/10.7554/eLife.45017.086
Figure supplement 80. Sequence tree of orthogroup OG0006623.
DOI: https://doi.org/10.7554/eLife.45017.087
Figure supplement 81. Sequence tree of orthogroup OG0006670.
DOI: https://doi.org/10.7554/eLife.45017.088
Figure supplement 82. Sequence tree of orthogroup OG0007051.
DOI: https://doi.org/10.7554/eLife.45017.089
Figure supplement 83. Sequence tree of orthogroup OG0007123.
DOI: https://doi.org/10.7554/eLife.45017.090
Figure supplement 84. Sequence tree of orthogroup OG0007346.
DOI: https://doi.org/10.7554/eLife.45017.091
Figure supplement 85. Sequence tree of orthogroup OG0007383.
DOI: https://doi.org/10.7554/eLife.45017.092
Figure supplement 86. Sequence tree of orthogroup OG0007550.
DOI: https://doi.org/10.7554/eLife.45017.093
Figure supplement 87. Sequence tree of orthogroup OG0007551.
DOI: https://doi.org/10.7554/eLife.45017.094
Figure supplement 88. Sequence tree of orthogroup OG0007596.
DOI: https://doi.org/10.7554/eLife.45017.095
Figure supplement 89. Sequence tree of orthogroup OG0008189.
DOI: https://doi.org/10.7554/eLife.45017.096
Figure supplement 90. Sequence tree of orthogroup OG0008334.
DOI: https://doi.org/10.7554/eLife.45017.097
Figure supplement 91. Sequence tree of orthogroup OG0008335.

Figure 5 continued on next page

Figure 5 continued

DOI: https://doi.org/10.7554/eLife.45017.098

Figure supplement 92. Sequence tree of orthogroup OG0008579.

DOI: https://doi.org/10.7554/eLife.45017.099

Figure supplement 93. Sequence tree of orthogroup OG0008680.

DOI: https://doi.org/10.7554/eLife.45017.100

Figure supplement 94. Sequence tree of orthogroup OG0008822.

DOI: https://doi.org/10.7554/eLife.45017.101

Figure supplement 95. Sequence tree of orthogroup OG0008898.

DOI: https://doi.org/10.7554/eLife.45017.102

Figure supplement 96. Sequence tree of orthogroup OG0008996.

DOI: https://doi.org/10.7554/eLife.45017.103

whereas 53.0% are present in less than three. The latter number might be an underestimation due to the strict threshold for HGT discovery which led to the removal of HGT candidates that were singletons. The HGT distribution is therefore skewed towards OGs containing only a few or one *Galdieria* species as the result of recent HGT events that occurred; for example after the split of *G. sulphuraria* and *G. phlegrea*. In spite of the strong erosion which would also lead to partial fixation of presumably recent HGT events, we analyzed whether the distribution patterns of HGT candidates across the sequenced genomes reflect the branching pattern of the species trees (**Figure 4C**). This is true for all HGT candidates that are exclusive to the *Cyanidioschyzon* or *Galdieria* lineage. Either the HGT candidates were acquired after the split of the two lineages (according to HGT), or differentially lost in one of the two lineages (according to DL). In the 60 *Galdieria* specific OGs we found 12 OGs containing less than 10 and more than one *Galdieria* species (**Figure 4C**). In 5/12 of the cases, the presence absence pattern reflects the species tree (OG0005087, OG0005083, GO0005479, OG0005540). Here, the potential HGT candidates are not found in any other eukaryotic species. According to HGT, they were acquired by a monophyletic sub-clade of the *Galdieria* lineage. According to DL, they were lost in all eukaryotes with the exception of this subset of the *Galdieria* lineage (e.g., OG0005280 and OG0005083 were potentially acquired or maintained exclusively by the last common ancestor of *G. sulphuraria* 074W, *G. sulphuraria* MS1, *G. sulphuraria* RT22, and *G. sulphuraria* SAG21). In the remaining OGs, the HGT gene candidate is distributed across the *Galdieria* lineage and conflicts with the branching pattern of the species tree. HGT would assume either multiple independent acquisitions of the same HGT candidate, or partial fixation of the HGT candidate in the lineage, while still allowing for gene erosion. According to DL, these are the last existing paralogs of an ancient gene, whose erosion within the eukaryotic kingdom is nearly complete. However, it must be considered that in some cases, DL must have occurred independently across multiple species in a brief of time after the gene was maintained for hundreds of millions of years across the lineage (e.g., OG0005224 contains *G. phlegrea* Soos, *G. sulphuraria* Azora and *G. sulphuraria* MS1). This implies that the gene was present in the ancestor of the *Galdieria* lineage and also in the last common ancestor of closely related *G. sulphuraria* MS1, *G. sulphuraria* 074W and *G. sulphuraria* RT22 (as well as *G. sulphuraria* SAG21) and the last common ancestor of closely related *G. sulphuraria* MtSh, *G. sulphuraria* Azora and *G. sulphuraria* YNP5587.1 (as well as *G. sulphuraria* 5572). A gene that was encoded and maintained since LECA, was lost independently in 6/8 species within the past few million years.

The seventy percent rule

In their analysis of eukaryotic HGT (**Ku and Martin, 2016**), Ku and co-authors reach the conclusion that prokaryotic homologs of genes in eukaryotic genomes that share >70% PID are not found outside individual genome assemblies (unless derived from endosymbiotic gene transfer, EGT). Hence, they are considered as assembly artifacts. We analyzed whether our dataset supports this rule, or alternatively, it is arbitrary and a byproduct of the analysis approach used, combined with low eukaryotic sampling (**Richards and Monier, 2016**; **Leger, 2018**). In addition to the 96 OGs potentially acquired through HGT, 2134 of the 9075 total OGs contained non-eukaryotic sequences, in which the Cyanidiales sequences cluster within the eukaryotic kingdom, but are similar enough to non-eukaryotic species to produce blast hits. Based on the average PID, no OG contains HGT

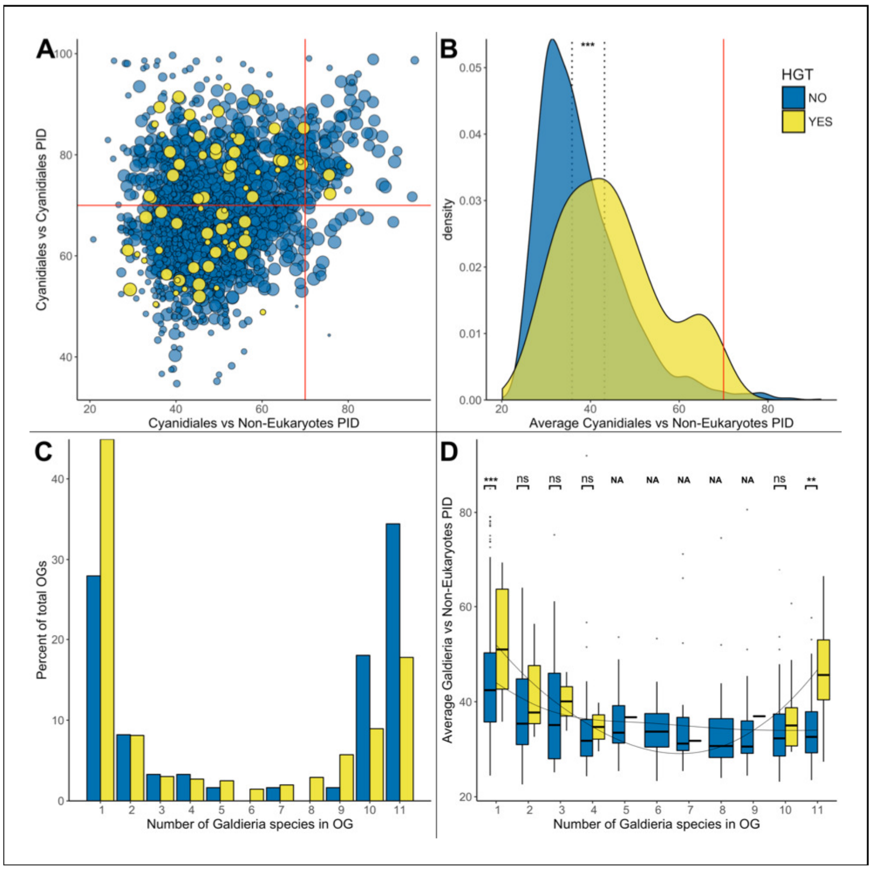


Figure 6. HGT vs. non-HGT orthogroup comparisons. (A) Maximum PID of Cyanidiales genes in native (blue) and HGT (yellow) orthogroups when compared to non-eukaryotic sequences in each OG. The red lines denote the 70% PID threshold for assembly artifacts according to 'the 70% rule'. Dots located in the top-right corner depict the 73 OGs that appear to contradict this rule, plus the 5 HGT candidates that score higher than 70%. 18/73 of those OGs are not derived from EGT or contamination within eukaryotic assemblies. (B) Density curve of average PID towards non-eukaryotic species in the same orthogroup (potential non-eukaryotic donors in case of HGT candidates). The area under each curve is equal to 1. The average PID of HGT candidates (left dotted line) is 6.1% higher than the average PID of native OGs also containing non-eukaryotic species (right dotted line). This difference is significant (Wilcoxon rank-sum test, $p>0.01$). (C) Distribution of OG-sizes (=number of *Galdieria* species present in each OG) between the native and HGT dataset. A total of 80% of the HGT OGs and 89% of the native OGs are present in either ≤10 species, or ≤2 species. Whereas 52.5% of the native gene set is conserved in ≤10 *Galdieria* strains, only 36.1% of the HGT candidates are conserved. In contrast, about 50% of the HGT candidates are present in only one *Galdieria* strain. (D) Pairwise OG-size comparison between HGT OGs and native OGs. A significantly higher PID when compared to non-eukaryotic sequences was measured in the HGT OGs at OG-sizes of 1 and 11 (Wilcoxon rank-sum test, BH, $p<0.01$). No evidence of cumulative effects was detected in the HGT dataset. However, the fewer *Galdieria* species that are contained in one OG, the higher the average PID when compared to non-eukaryotic species in the same tree (Jonckheere-Terpstra, $p<0.01$) in the native dataset.

DOI: https://doi.org/10.7554/eLife.45017.104

candidates that share over 70% PID to their non-eukaryotic donors with OG0006191 having the highest average PID (69.68%). However, 5/96 HGT-impacted OGs contain one or more individual HGT candidates that exceed this threshold (5.2% of the HGT OGs) (***Figure 6A***). These sequences are found in OG0001929 (75.56% PID, 11 *Galdieria* species), OG0002676 (75.76% PID, 11 *Galdieria* species), OG0006191 (80.00% PID, both *Cyanidioschyzon* species), OG0008680 (72.37% PID, 1 *Galdieria* species), and OG0008822 (71.17% PID, 1 *Galdieria* species). Moreover, we find 73 OGs with eukaryotes as sisters sharing over 70% PID to non-eukaryotic sequences (0.8% of the native OGs) (***Figure 6A***). On closer inspection, the majority are derived from endosymbiotic gene transfer (EGT): 16/73 of the OGs are of proteobacterial descent and 33/73 OGs are phylogenies with gene origin in Cyanobacteria and/or Chlamydia. These annotations generally encompass mitochondrial/plastid components and reactions, as well as components of the phycobilisome, which is exclusive to Cyanobacteria, red algae, and red algal derived plastids. Of the remaining 24 OGs, 18 cannot be explained through EGT or artifacts alone unless multiple eukaryotic genomes would share the same artifact (and also assuming all gene transfers from Cyanobacteria, Chlamydia, and Proteobacteria are derived from EGT). A total of 6/24 OGs are clearly cases of contamination within the eukaryotic assemblies. Although 'the 70% rule' captures a large proportion of the dataset, increasing the sampling resolution within eukaryotes increased the number of exceptions to the rule. This number is likely to increase as more high-quality eukaryote nuclear genomes are determined. Considering the paucity of these data across the eukaryotic tree of life and the rarity of eukaryotic HGT, the systematic dismissal of eukaryotic singletons located within non-eukaryotic branches as assembly/annotation artifacts (or contamination) may come at the cost of removing true positives.

Cumulative effects

We assessed our dataset for evidence of cumulative effects within the candidate HGT-derived OGs. If cumulative effects were present, then recent HGT candidates would share higher similarity to their non-eukaryotic ancestors than genes resulting from more ancient HGT. Hence, the fewer species that are present in an OG, the higher likelihood of a recent HGT (unless the tree branching pattern contradicts this hypothesis, such as in OG 0005224, which is limited to 3 *Galdieria* species, but is ancient due to its presence in *G. sulphuraria* and *G. phlegrea*). In the case of DL, no cumulative effects as well as no differences between the HGT and native dataset are expected because the PID between eukaryotes and non-eukaryotes is irrelevant to this issue because all genes are native and occurred in the LECA. According to DL, the monophyletic position of Cyanidiales HGT candidates with non-eukaryotes is determined by the absence of other eukaryotic orthologs (given the limited current data) and may be the product of deep branching effects.

First, we tested for general differences in PID with regard to non-eukaryotic sequences between the native and HGT datasets (***Figure 6B***). Neither the PID with non-eukaryotic species in the same OG for the native dataset, nor the PID with potential non-eukaryotic donors in the same OG for the HGT dataset was normally distributed (Shapiro-Wilk normality test, p=2.2e-16/0.00765). Consequently, exploratory analysis was performed using non-parametric testing. On average, the PID with non-eukaryotic species in OGs containing HGT candidates is higher by 6.1% in comparison to OGs with eukaryotic descent. This difference is significant (Wilcoxon rank-sum test, p=0.000008).

Second, we assessed if OGs containing fewer *Galdieria* species would have a higher PID with their potential non-eukaryotic donors in the HGT dataset. We expected a lack of correlation with OG size in the native dataset because the presence/absence pattern of HGT candidates within the *Galdieria* lineage is dictated by gene erosion and thus independent of which non-eukaryotic sequences also cluster in the same phylogeny. Jonckheere's test for trends revealed a significant trend within the native subset: the fewer *Galdieria* species are contained in one OG, the higher the average PID with non-eukaryotic species in the same tree (Jonckheere-Terpstra, p=0.002). This was not the case in the 'HGT' subset. Here, no general trend was observed (Jonckheere-Terpstra, p=0.424).

Third, we compared the PID between HGT-impacted OGs and native OGs of the same size (OGs containing the same number of *Galdieria* species). This analysis revealed a significantly higher PID with non-eukaryotic sequences in favor of the HGT subset in OGs containing either one *Galdieria* sequence, or all 11 *Galdieria* sequences (Wilcoxon rank-sum test, Benjamini-Hochberg, p=2.52e-08| 3.39e-03) (***Figure 6D***). Hence, the 'most recent' and 'most ancient' HGT candidates share the highest identity with their non-eukaryotic donors, which is also significantly higher when compared to native genes in OGs of the same size.

Natural habitat of extant prokaryotes with closely related orthologs

We next set out to explore the natural habitats of extant prokaryotes that harbor the closest orthologs with candidate HGTs in the Cyanidiales. To this end, we counted the frequency at which any non-eukaryotic species shared monophyly with Cyanidiales (*Table 2*). A total of 568 non-eukaryotic species (19 Archaea, 549 Bacteria), from 365 different genera representing 24 divisions share monophyly with the 96 OGs containing HGT candidates. Most prominent are Proteobacteria that are sister phyla to 53/96 OGs. This group is followed by Firmicutes (28), Actinobacteria (19), Chloroflexi (12), and Bacteroidetes/Chlorobi (10). The only frequently occurring archaeal orthologs were found in Euryarchaeota (6 OGs). Interestingly, the closest orthologs often occurred in extremophilic prokaryotes that share similar (current) habitats with Cyanidiales. We hypothesize that potential non-eukaryotic HGT donors might share similar habitats because proximity is thought to favor HGT. However, we have no direct evidence of what the environment might have been at the time of HGT, or whether a third organism acted as the vector and has not been sampled in our analyses. It is worth noting that the phylogenetic data clearly demonstrate that Cyanidiales have been extremophiles for hundreds

Table 2. Natural habitats of extant prokaryotes harboring the closest orthologs to Cyanidiales HGTs.

Numbers in brackets represent how many times HGT candidates from Cyanidiales shared monophyly with non-eukaryotic organisms; for example Proteobacteria were found in 53/96 of the OG monophylies. **Kingdom**: Taxon at kingdom level. **Species**: Scientific species name. **Habitat**: habitat description of the original sampling site. **pH**: pH of the original sampling site. **Temp**: Temperature in Celsius of the sampling site. **Salt**: Ion concentration of the original sampling site. **na**: no information available.

	Phylogeny		Natural habitat of closest non-eukaryotic ortholog			
Kingdom	**Division**	**Species**	**Habitat description**	**pH**	**Max. temp**	**Salt**
Bacteria	Proteobacteria (53)	*Acidithiobacillus thiooxidans (4)*	Mine drainage/Mineral ores	2.0–2.5	30°C	'hypersaline'
		Carnimonas nigrificans (4)	Raw cured meat	3.0	35°C	8% NaCl
		Methylosarcina fibrata (4)	Landfill	5.0–9.0	37°C	1% NaCl
		Sphingomonas phyllosphaerae (3)	Phyllosphere of Acacia caven	na	28°C	na
		Gluconacetobacter diazotrophicus (3)	Symbiont of various plant species	2.0–6.0	na	'high salt'
		Gluconobacter frateurii (3)	na	na	na	na
		Luteibacter yeojuensis (3)	River	na	na	na
		Thioalkalivibrio sulfidiphilus (3)	Soda lake	8.0–10.5	40°C	15% total salts
		Thiomonas arsenitoxydans (3)	Disused mine site	3.0–8.0	30°C	'halophilic'
	Firmicutes (28)	*Sulfobacillus thermosulfidooxidans (6)*	Copper mining	2.0–2.5	45°C	'salt tolerant'
		Alicyclobacillus acidoterrestris (4)	Soil sample	2.0–6.0	53°C	5% NaCl
		Gracilibacillus lacisalsi (3)	Salt lake	7.2–7.6	50°C	25% total salts
	Actinobacteria (19)	*Amycolatopsis halophila (3)*	Salt lake	6.0–8.0	45°C	15% NaCl
		Rubrobacter xylanophilus (3)	Thermal industrial runoff	6.0–8.0	60°C	6.0% NaCl
	Chloroflexi (12)	*Caldilinea aerophila (4)*	Thermophilic granular sludge	6.0–8.0	65°C	3% NaCl
		Ardenticatena maritima (3)	Coastal hydrothermal field	5.5–8.0	70°C	6% NaCl
		Ktedonobacter racemifer (3)	Soil sample	4.8–6.8	33°C	>3% NaCl
	Bacteroidetes Chlorobi (10)	*Salinibacter ruber (4)*	Saltern crystallizer ponds	6.5–8.0	52°C	30% total salts
		Salisaeta longa (3)	Experimental mesocosm (Salt)	6.5–8.5	46°C	20% NaCl
	Nitrospirae (7)	*Leptospirillum ferriphilum (4)*	Arsenopyrite biooxidation tank	0–3.0	40°C	2% NaCl
	Fibrobacteres (6)	*Acidobacteriaceae bacterium TAA166 (3)*	na	na	na	na
	Deinococcus (5)	*Truepera radiovictrix (3)*	Hot spring runoffs	7.5–9.5	na	6% NaCl
Archaea	Euryarchaeota (6)	*Ferroplasma acidarmanus (3)*	Acid mine drainage	0–2.5	40°C	'halophilic'

DOI: https://doi.org/10.7554/eLife.45017.105

of millions of years. It is however conceivable that the HGTs may have occurred when these cells were being dispersed (they have a worldwide distribution) from one extreme site to another and would have encountered mesophilic donors at these times. Given these caveats, it is interesting to note that *Sulfobacillus thermosulfidooxidans* (Firmicutes), a mixotrophic, acidophilic (pH 2.0), and moderately thermophilic (45˚C) bacterium that was isolated from acid mining environments in northern Chile (where *Galdieria* is also present) was most prominent amongst the prokaryotic orthogroups. *Sulfobacillus thermosulfidooxidans* shares monophyly in 6/96 HGT-derived OGs and is followed in frequency by several species that are either thermophiles, acidophiles, or halophiles and share habitats in common with Cyanidiales (***Table 2***).

Functions of horizontally acquired genes in cyanidiales

We analyzed the putative molecular functions and processes acquired through HGT. Annotations were curated using information gathered from blast, GO-terms, PFAM, KEGG, and EC. A total of 72 GO annotations occurred more than once within the 96 HGT-impacted OGs. Furthermore, 37/72 GO annotations are significantly enriched (categorical data, 'native' vs 'HGT', Fisher's exact test, Benjamini-Hochberg, $p \leq 0.05$). The most frequent terms were: 'decanoate-CoA ligase activity' (5/72 GOs, p=0), 'oxidation-reduction process' (16/72 GOs, p=0.001), 'transferase activity' (14/72 GOs, p=0.009), 'carbohydrate metabolic process' (5/72 GOs, p=0.01), 'oxidoreductase activity' (9/72 GOs, p=0.012), 'methylation '(6/72 GOs, p=0.013), 'methyltransferase activity' (5/72 GOs, p=0.023), 'transmembrane transporter activity' (4/72 GOs, p=0.043), and 'hydrolase activity' (9/72 GOs, p=0.048). In comparison to previous studies, our analysis did not report a significant enrichment of membrane proteins in the HGT dataset ('membrane', 11/72 OGs, p=0.699; 'integral component of membrane', 22/72 GOs, p=0.416. The GO annotation 'extracellular region' was absent in the HGT dataset) (***Schönknecht et al., 2013***). As such, we report a strong bias for metabolic functions among HGT candidates (***Figure 7***).

Metal and xenobiotic resistance/detoxification

Geothermal environments often contain high arsenic (Ar) concentrations, up to a several g/L as well as high levels of mercury (Hg), such as >200 μg/g in soils of the Norris Geyser Basin (Yellowstone National Park) and volcanic waters in southern Italy (***Stauffer and Thompson, 1984***; ***Aiuppa, 2003***), both known Cyanidiales habitats (***Castenholz and McDermott, 2010***; ***Ciniglia et al., 2004***; ***Toplin et al., 2008***; ***Pinto, 1975***). Studies with *G. sulphuraria* have shown an increased efficiency and speed regarding the biotransformation of $HgCl_2$ compared to eukaryotic algae (***Kelly et al., 2007***). Orthologs of OG0002305, which are present in all 13 Cyanidiales genomes, encode mercuric reductase that catalyzes the critical step in Hg^{2+} detoxification, converting cytotoxic Hg^{2+} into the less toxic metallic mercury, Hg^0. Arsenate (As(V)) is imported into the cell by high-affinity P_i transport systems (***Meharg and MacNair, 1992***; ***Catarecha et al., 2007***), whereas aquaporins regulate arsenite (As(III)) uptake (***Zhao et al., 2010***). *Galdieria* and *Cyanidioschyzon* possess a eukaryotic gene-set for the chemical detoxification and extrusion of As through biotransformation and direct efflux (***Schönknecht et al., 2013***). Arsenic tolerance was expanded in the *Galdieria* lineage through the acquisition (OG0001513) of a bacterial ***arsC*** gene, thus enabling the reduction of As(V) to As(III) using thioredoxin as the electron acceptor. It is known that As(III) can be converted into volatile dimethylarsine and trimethylarsine through a series of reactions, exported, or transported to the vacuole in conjugation with glutathione. Two separate acquisitions of a transporter annotated as ArsB are present in *G. sulphuraria* RT22 and *G. sulphuraria* 5572 (OG0006498, OG0006670), as well as a putative cytoplasmic heavy metal binding protein (OG0006191) in the *Cyanidioschyzon* lineage.

In the context of xenobiotic detoxification, we found an aliphatic nitrilase (OG0001760) involved in styrene degradation and three (OG0003250, OG0005087, OG0005479) *Galdieria* specific 4-nitrophenylphosphatases likely involved in the bioremediation of highly toxic hexachlorocyclohexane (HCH) (***van Doesburg et al., 2005***), or more generally other cyclohexyl compounds, such as cyclohexylamine. In this case, bioremediation can be achieved through the hydrolysis of 4-nitrophenol to 4-nitrophenyl phosphate coupled with phosophoesterase/metallophosphatase activity. The resulting cyclohexyl compounds serve as multifunctional intermediates in the biosynthesis of various heterocyclic and aromatic metabolites. A similar function in the *Cyanidioschyzon* lineage could be taken up by OG0006252, a cyclohexanone monooxygenase (***Chen et al., 1988***) oxidizing phenylacetone to

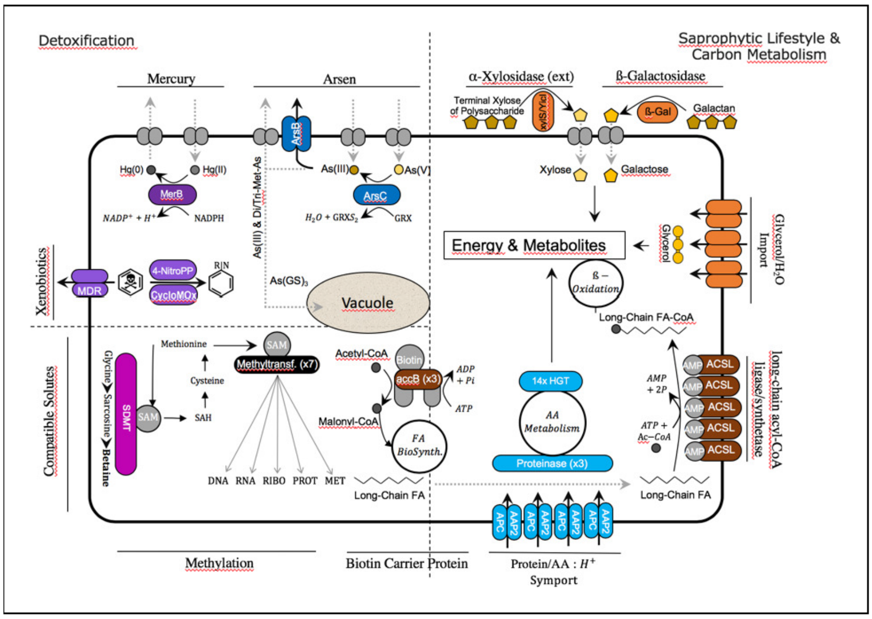


Figure 7. Cyanidiales live in hostile habitats, necessitating a broad range of adaptations to polyextremophily. The majority of the 96 HGT-impacted OGs were annotated and putative functions identified (in the image, colored fields are from HGT, whereas gray fields are native functions). The largest number of HGT candidates is involved in carbon and amino acid metabolism, especially in the *Galdieria* lineage. The excretion of lytic enzymes and the high number of importers (protein/AA symporter, glycerol/H_2O symporter) within the HGT dataset suggest a preference for import and catabolic function.

DOI: https://doi.org/10.7554/eLife.45017.106

benzyl acetate that can also oxidize various aromatic ketones, aliphatic ketones (e.g., dodecan- 2-one) and sulfides (e.g., 1-methyl-4-(methylsulfanyl)benzene). In this context, a probable multidrug-resistance/quaternary ammonium compound exporter (OG0002896), which is present in all Cyanidiales, may control relevant efflux functions whereas a phosphatidylethanolamine (penicillin?) binding protein (OG0004486) could increase the stability of altered peptidoglycan cell walls. If these annotations are correct, then *Galdieria* is an even more promising target for industrial bioremediation applications than previously thought (***Henkanatte-Gedera et al., 2017***; ***Fukuda et al., 2018***).

Cellular oxidant reduction

Increased temperature leads to a higher metabolic rate and an increase in the production of endogenous free radicals (FR), such as reactive oxygen species (ROS) and reactive nitrogen species (RNS), for example during cellular respiration (***Phaniendra et al., 2015***). Furthermore, heavy metals such as lead and mercury, as well as halogens (fluorine, chlorine, bromine, iodine) stimulate formation of FR (***Dietz et al., 1999***). FR are highly biohazard and cause damage to lipids (***Ylä-Herttuala, 1999***), proteins (***Stadtman and Levine, 2000***) and DNA (***Marnett, 2000***). In the case of the superoxide radical ($^{\bullet}O^{2-}$), enzymes such as superoxide dismutase enhance the conversion of 2 x $^{\bullet}O^{2-}$, into hydrogen peroxide (H_2O_2) which is in turn reduced to H_2O through the glutathione-ascorbate cycle. Other toxic hydroperoxides (R-OOH) can be decomposed various peroxidases to H_2O and alcohols (R-OH) at the cost of oxidizing the enzyme, which is later recycled (re-reduced) through oxidation of thioredoxin (***Rouhier et al., 2008***). The glutathione and thioredoxin pools and their related enzymes are thus factors contributing to a successful adaptation to geothermal environments. Here, we found a cytosolic and/or extracellular peroxiredoxin-6 (OG0005984) specific to the *Cyanidioschyzon* lineage and two peroxidase-related enzymes (probable alkyl hydroperoxide reductases acting on carboxymuconolactone) in the *Galdieria* lineage (OG0004203, OG0004392) (**Chae et al., 1994**). In addition, a thioredoxin oxidoreductase related to alkyl hydroperoxide reductases (OG0001486) as well as a putative glutathione-specific gamma-glutamylcyclotransferase 2 (OG0003929) are present in all Cyanidiales. The latter has been experimentally linked to the process of heavy metal detoxification in *Arabidopsis thaliana* (**Paulose et al., 2013**).

Carbon metabolism

G. sulphuraria is able to grow heterotrophically using a large variety of different carbon sources and compounds released from dying cells (**Gross et al., 1995**; **Gross, 1998**). In contrast, *C. merolae* is strictly photoautotrophic (**De Luca et al., 1978**). *G. sulphuraria* can be maintained on glycerol as the sole carbon source (**Gross et al., 1995**) making use of a family of glycerol uptake transporters likely acquired via HGT (**Schönknecht et al., 2013**). We confirm the lateral acquisition of glycerol transporters in *G. sulphuraria* RT22 (OG0006482), *G. sulphuraria* Azora and *G. sulphuraria* SAG21 (OG0005235). The putative HGT glycerol transporters found in *G. sulphuraria* 074W did not meet the required threshold of two Cyanidiales sequences (from different strains) in one OG. In addition, another MIP family aquaporin, permeable to H_2O, glycerol and other small uncharged molecules (**Liu et al., 2007**) is encoded by *G. sulphuraria* Azora (OG0007123). This could be an indication of a very diverse horizontal acquisition pattern regarding transporters. OG0003954 is the only exception to this rule, because it is present in all *Galdieria* lineages and is orthologous to AcpA|SatP acetate permeases involved with the uptake of acetate and succinate (**Robellet et al., 2008**; **Sá-Pessoa et al., 2013**).

We found evidence of saprophytic adaptations in *Galdieria* through the potential horizontal acquisition of an extracellular beta-galactosidase enzyme (**Rojas et al., 2004**; **Rico-Díaz et al., 2014**). This enzyme contains all five bacterial beta-galactosidase domains (OG0003441) involved in the catabolism of glycosaminoglycans, a polysaccharide deacetylase/peptidoglycan-N-acetylglucosamine deacetylase (OG0004030) acting on glucosidic (but note peptide bonds) that may degrade chitooligosaccharides, chitin, and/or xylan (**Psylinakis et al., 2005**; **Lee et al., 2002**) as well as an α-amylase (OG0004658) converting starch/glycogen to dextrin/maltose (**Diderichsen and Christiansen, 1988**) which is missing only in *G. sulphuraria* SAG21. All other HGT OGs involved in sugar metabolism are involved in the intercellular breakdown and interconversions of sugar carbohydrates. OG0006623 contains a non-phosphorylating glyceraldehyde-3-phosphate dehydrogenase found in hyperthermophile archaea (**Ettema et al., 2008**) (*G. sulphuraria* 002). The OG0005153 encodes a glycosyl transferase family one protein involved in carbon metabolism (*G. sulphuraria* 074W, *G. sulphuraria* MS1, *G. sulphuraria* RT22, *G. sulphuraria* YNP5587.1). All *Galdieria* have an alpha-xylosidase resembling an extremely thermo-active and thermostable α-galactosidase (OG0001542) (**van Lieshout et al., 2003**; **Okuyama et al., 2004**). The only horizontal acquisition in this category present in all Cyanidiales is a cytoplasmic ribokinase involved in the D-ribose catabolic process (OG0001613).

The irreversible synthesis of malonyl-CoA from acetyl-CoA through acetyl-CoA carboxylase (ACCase) is the rate limiting and step in fatty acid biosynthesis. The bacterial ACCase complex consists of three separate subunits, whereas the eukaryotic ACCase is composed of a single multifunctional protein. Plants contain both ACCase isozymes. The eukaryotic enzyme is located in the cytosol and a bacterial-type enzyme consisting of four subunits is plastid localized. Three of the HGT orthogroups (OG0002051, OG0007550 and OG0007551) were annotated as bacterial biotin carboxyl carrier proteins (AbbB/BCCP), which carry biotin and carboxybiotin during the critical and highly regulated carboxylation of acetyl-CoA to form malonyl-CoA [ATP +Acetyl CoA + HCO^{3-} ⇌ ADP + Orthophosphate + Malonyl-CoA]. Whereas OG0002051 is present in all Cyanidiales and located in the cytoplasm, OG0007550 and OG0007551 are unique to *C. merolae* Soos and annotated as 'chloroplastic'. Prior to fatty acid (FA) beta-oxidation, FAs need to be transformed to a FA-CoA before entering cellular metabolism as an exogenous or endogenous carbon source (eicosanoid metabolism is the exception). This process is initiated by long-chain-fatty-acid-CoA ligases/acyl-CoA synthetases (ACSL) (**Mashek et al., 2007**) [ATP + long-chain carboxylate + CoA ⇌ AMP + diphosphate + Acyl-CoA]. Five general non-eukaryotic ACSL candidates were found (OG0001476, OG0002999, OG0005540, OG0008579, OG0008822). Only OG0001476 is present in all species, whereas OG0002999 is present in all *Galdieria,* OG0005540 in *G. sulphuraria* 074W and *G.*

sulphuraria MS1, and OG0008579 and OG0008822 are unique to *G. phlegrea* DBV009. The GO annotation suggests moderate specificity to decanoate-CoA. However, OG0002999 also indicates involvement in the metabolism of linoleic acid, a $C_{18}H_{32}O_2$ polyunsaturated acid found in plant glycosides. ACSL enzymes share significant sequence identity but show partially overlapping substrate preferences in terms of length and saturation as well as unique transcription patterns. Furthermore, ACSL proteins play a role in channeling FA degradation to various pathways, as well as enhancing FA uptake and FA cellular retention. Although an annotation of the different ACSL to their specific functions was not possible, their involvement in the saprophytic adaptation of *Cyanidioschyzon* and especially *Galdieria* appears to be plausible.

Amino acid metabolism

Oxidation of amino acids (AA) can be used as an energy source. Once AAs are deaminated, the resulting α-ketoacids ('carbon backbone') can be used in the tricarboxylic acid cycle for energy generation, whereas the remaining NH_4^+ can be used for the biosynthesis of novel AAs, nucleotides, and ammonium containing compounds, or dissipated through the urea cycle. In this context, we confirm previous observations regarding a horizontal origin of the urease accessory protein UreE (OG0003777) present in the *Galdieria* lineage (**Qiu et al., 2013**) (the other urease genes reported in *G. phlegrea* DBV009 appear to be unique to this species and were thus removed from this analysis as singletons; for example *ureG*, OG0008984). AAs are continuously synthesized, interconverted, and degraded using a complex network of balanced enzymatic reactions (e.g., peptidases, lyases, transferases, isomerases). Plants maintain a functioning AA catabolism that is primarily used for the interconversion of metabolites because photosynthesis is the primary source of energy. The Cyanidiales, and particularly the *Galdieria* lineage is known for its heterotrophic lifestyle. We assigned 19/96 HGT-impacted OGs to this category. In this context, horizontal acquisition of protein|AA:proton symporter AA permeases (OG0001658, OG0005224, OG0005596, OG0007051) may be the first indication of adaptation to a heterotrophic lifestyle in *Galdieria*. Once a protein is imported, peptidases cleave single AAs by hydrolyzing the peptide bonds. Although no AA permeases were found in the *Cyanidioschyzon* lineage, a cytoplasmic threonine-type endopeptidase (OG0001994) and a cytosolic proline iminopeptidase involved in arginine and proline metabolism (OG0006143) were potentially acquired through HGT. At the same time, the *Galdieria* lineage acquired a Clp protease (OG0007596). The remaining HGT candidates are involved in various amino acid metabolic pathways. The first subset is shared by all Cyanidiales, such as a cytoplasmic imidazoleglycerol-phosphate synthase involved in the biosynthetic process of histidine (OG0002036), a phosphoribosyltransferase involved in phenylalanine/tryptophan/tyrosine biosynthesis (OG0001509) and a peptydilproline peptidyl-prolyl cis-trans isomerase acting on proline (OG0001938) (**Dilworth et al., 2012**). The second subset is specific to the *Cyanidium* lineage. It contains a glutamine/leucine/phenylalanine/valine dehydrogenase (OG0006136) (**Kloosterman et al., 2006**), a glutamine cyclotransferase (OG0006251) (**Dahl et al., 2000**), a cytidine deaminase (OG0003539) as well as an adenine deaminase (OG0005683) and a protein binding hydrolase containing a NUDIX domain (OG0005694). The third subset is specific to the *Galdieria* lineage and contains an ornithine deaminase, a glutaryl-CoA dehydrogenase (OG0007383) involved in the oxidation of lysine, tryptophan, and hydroxylysine (**Rao et al., 2006**), as well as an ornithine cyclodeaminase (OG0004258) involved in arginine and/or proline metabolism. Finally, a lysine decarboxylase (OG0007346), a bifunctional ornithine acetyltransferase/N-acetylglutamate synthase (**Martin and Mulks, 1992**) involved in the arginine biosynthesis (OG0008898) and an aminoacetone oxidase family FAD-binding enzyme (OG0007383), probably catalytic activity against several different L-amino acids were found as unique acquisitions in *G. sulphuraria* SAG21, *G. phlegrea* DBV009 and *G. sulphuraria* YNP5587.1 respectively.

One carbon metabolism and methylation

One-carbon (1C) metabolism based on folate describes a broad set of reactions involved in the activation and transfer C1 units in various processes including the synthesis of purine, thymidine, methionine, and homocysteine re-methylation. C1 units can be mobilized using tetrahydrofolate (THF) as a cofactor in enzymatic reactions, vitamin B12 (cobalamin) as a co-enzyme in methylation/rearrangement reactions and S-adenosylmethionine (SAM) (**Ducker and Rabinowitz, 2017**). In terms of purine biosynthesis, OG0005280 encodes an ortholog of a bacterial FAD-dependent thymidylate (dTMP) synthase converting dUMP to dTMP by oxidizing THF present in *G. sulphuraria* 074W, *G. sulphuraria* MS1, and *G. sulphuraria* RT22. In terms of vitamin B12 biosynthesis, an ortholog of the cobalamin biosynthesis protein CobW was found in the *Cyanidioschyzon* lineage (OG0002609). Much of the methionine generated through C1 metabolism is converted to SAM, the second most abundant cofactor after ATP, which is a universal donor of methyl (-CH_3) groups in the synthesis and modification of DNA, RNA, hormones, neurotransmitters, membrane lipids, proteins and also play a central role in epigenetics and posttranslational modifications. Within the 96 HGT-impacted dataset we found a total of 9 methyltransferases (OG0003901, OG0003905, OG0002191, OG0002431, OG0002727, OG0003907, OG0005083 and OG0005561) with diverse functions, 8 of which are SAM-dependent methyltransferases. OG0002431 (Cyanidiales), OG0005561 (*G. sulphuraria* MS1 and *G. phlegrea* DBV009) and OG0005083 (*G. sulphuraria* SAG21) encompass rather unspecific SAM-dependent methyltransferases with a broad range of possible methylation targets. OG0002727, which is exclusive to *Cyanidioschyzon*, and OG0002191, which is exclusive to *Galdieria*, both methylate rRNA. OG0002727 belongs to the Erm rRNA methyltransferase family that methylate adenine on 23S ribosomal RNA (***Yu et al., 1997***). Whether it confers macrolide-lincosamide-streptogramin (MLS) resistance, or shares only adenine methylating properties remains unclear. The OG0002191 is a 16S rRNA (cytidine1402-2'-O)-methyltransferase involved the modulation of translational fidelity (***Kimura and Suzuki, 2010***).

Osmotic resistance and salt tolerance

Cyanidiales withstand salt concentrations up to 10% NaCl (***Albertano, 2000***). The two main strategies to prevent the accumulation of cytotoxic salt concentrations and to withstand low water potential are the active removal of salt from the cytosol and the production of compatible solutes. Compatible solutes are small metabolites that can accumulate to very high concentrations in the cytosol without negatively affecting vital cell functions while keeping the water potential more negative in relation to the saline environment, thereby avoiding loss of water. The *G. sulphuraria* lineage produces glycine/betaine as compatible solutes under salt stress in the same manner as halophilic bacteria (***Imhoff and Rodriguez-Valera, 1984***) through the successive methylation of glycine via sarcosine and dimethylglycine to yield betaine using S-adenosyl methionine (SAM) as a cofactor (***Lu et al., 2006***; ***Waditee et al., 2003***; ***Nyyssola et al., 2000***). This reaction is catalyzed by the enzyme sarcosine dimethylglycine methyltransferase (SDMT), which has already been characterized in *Galdieria* (***McCoy et al., 2009***). Our results corroborate the HGT origin of this gene, supporting two separate acquisitions of this function (OG0003901, OG0003905). In this context, a inositol 2-dehydrogenase possibly involved in osmoprotective functions (***Kingston et al., 1996***) in *G. phlegrea* DBV009 was also found in the HGT dataset (OG0008335).

Non-Metabolic functions

Outside the context of HGT involving enzymes that perform metabolism related functions, we found 6/96 OGs that are annotated as transcription factors, ribosomal components, rRNA, or fulfilling functions not directly involved in metabolic fluxes. Specifically, two OGs associated with the bacterial 30S ribosomal subunit were found, whereas OG0002627 (*Galdieria*) is orthologous to the tRNA binding translation initiation factor eIF1a which binds the fMet-tRNA(fMet) start site to the ribosomal 30S subunit and defines the reading frame for mRNA translation (***Simonetti et al., 2009***), and OG0004339 (*Galdieria*) encodes the S4 structural component of the S30 subunit. Three genes functioning as regulators were found in *Cyanidioschyzon*, a low molecular weight phosphotyrosine protein phosphatase with an unknown regulator function (OG0002785), a SfsA nuclease (***Takeda et al., 2001***), similar to the sugar fermentation stimulation protein A and (OG0002871) a MRP family multidrug resistance transporter connected to parA plasmid partition protein, or generally involved in chromosome partitioning (mrp). Additionally, we found a *Cyanidioschyzon*-specific RuvX ortholog (OG0002578) involved in chromosomal crossovers with endonucleolytic activity (***Nautiyal et al., 2016***) as well as a likely Hsp20 heat shock protein ortholog (OG0004102) unique to the *Galdieria* lineage.

Various functions and uncertain annotations

The remaining OGs were annotated with a broad variety of functions. For example, OG0001929, OG0001810, OG0004405, and OG0001087 are possibly connected to the metabolism of cell wall precursors and components and OG0001929 (*Galdieria*) is an isomerizing glutamine-fructose-6-phosphate transaminase most likely involved in regulating the availability of precursors for N- and O-linked glycosylation of proteins, such as for peptidoglycan. In contrast, OG0004405 (*Cyanidioschyzon*) synthesizes exopolysaccharides on the plasma membrane and OG0001087 (*Cyanidiales*) and OG0001810 (*Cyanidioschyzon*) are putative undecaprenyl transferases (UPP) which function as lipid carrier for glycosyl transfer in the biosynthesis of cell wall polysaccharide components in bacteria (**Apfel et al., 1999**). The OGs OG0002483 and OG0001955 are involved in purine nucleobase metabolic processes, probably in cAMP biosynthesis (**Galperin, 2005**) and IMP biosynthesis (**Schrimsher et al., 1986**). A *Cyanidioschyzon* specific 9,15,9'-tri-cis-zeta-carotene isomerase (OG0002574) may be involved in the biosynthesis of carotene (**Chen et al., 2010**). Two of the 96 HGT OGs obtained the tag 'hypothetical protein' and could not be further annotated. Others had non-specific annotations, such as 'selenium binding protein' (OG0003856) or contained conflicting annotations.

Discussion

Making an argument for the importance of HGT in eukaryote (specifically, Cyanidiales) evolution, as we do here, requires that three major issues are addressed: a mechanism for foreign gene uptake and integration, the apparent absence of eukaryotic pan-genomes, and the lack of evidence for cumulative effects (**Martin, 2017**). The latter two arguments are dealt with below but the first concern no longer exists. For example, recent work has shown that red algae harbor naturally occurring plasmids, regions of which are integrated into the plastid DNA of a taxonomically wide array of species (**Lee et al., 2016**). Genetic transformation of the unicellular red alga *Porphyridium purpureum* has demonstrated that introduced plasmids accumulate episomally in the nucleus and are recognized and replicated by the eukaryotic DNA synthesis machinery (**Li and Bock, 2018**). These results suggest that a connection can be made between the observation of bacterium-derived HGTs in *P. purpureum* (**Bhattacharya et al., 2013**) and a putative mechanism of bacterial gene origin *via* long-term plasmid maintenance. Other proposed mechanisms for the uptake and integration of foreign DNA in eukaryotes are well-studied, observed in nature, and can be successfully recreated in the lab (**Leger, 2018**; **Li and Bock, 2018**).

HGT- the eukaryotic pan-genome

Eukaryotic HGT is rare and affected by gene erosion. Within the 13 analyzed genomes of the polyextremophilic Cyanidiales (**Foflonker et al., 2018**; **Schönknecht et al., 2014**), we identified and annotated 96 OGs containing 641 single HGT candidates. Given an approximate age of 1,400 Ma years and ignoring gene erosion, on average, one HGT event occurs every 14.6 Ma years in Cyanidiales. This figure ranges from one HGT every 33.3 Ma years in *Cyanidioschyzon* and one HGT every 13.3 Ma in *Galdieria*. Still, one may ask, given that eukaryotic HGT exists, what comprises the eukaryotic pan-genome and why does it not increase in size as a function of time due to HGT accumulation? In response, it should be noted that evolution is 'blind' to the sources of genes and selection does not act upon native genes in a manner different from those derived from HGT. In our study, we report examples of genes derived from HGT that are affected by gene erosion and/or partial fixation (**Figure 4A**). As such, only 8/96 of the HGT-impacted OGs (8.3%) are encoded by all 13 Cyanidiales species. Looking at the *Galdieria* lineage alone (**Figure 6C**), 28 of the 60 lineage-specific OGs (47.5%) show clear signs of erosion (HGT orthologs are present in ≤10 *Galdieria* species), to the point where a single ortholog of an ancient HGT event may remain.

When considering HGT in the Cyanidiales it is important to keep in mind the ecological boundaries of this group, the distance between habitats, the species composition of habitats, and the mobility of Cyanidiales within those borders that control HGT. Hence, we would not expect the same HGT candidates derived from the same non-eukaryotic donors to be shared between Cyanidiales and marine/freshwater red algae (unless they predate the split between Cyanidiales and other red algae), but rather between Cyanidiales and other polyextremophilic organisms. In this context, inspection of the habitats and physiology of potential HGT donors revealed that the vast majority is extremophilic and, in some cases, shares the same habitat as Cyanidiales (**Table 2**). A total of 84/96 of the inherited gene functions could be connected to ecologically important traits such as heavy metal detoxification, xenobiotic detoxification, ROS scavenging, and metabolic functions related to carbon, fatty acid, and amino acid turnover. In contrast, only 6/96 OGs are related to methylation and ribosomal functions. We did not find HGTs contributing other traits such as ultrastructure, development, or behavior (**Figure 7**). If cultures were exposed to abiotic stress, the HGT candidates were significantly enriched within the set of differentially expressed genes (**Figure 3**). These results not only provide evidence of successful integration into the transcriptional circuit of the host, but also support an adaptive role of HGT as a mechanism to acquire beneficial traits. Because eukaryotic HGT is the exception rather than the rule, its number in eukaryotic genomes does not need to increase as a function of time and may have reached equilibrium in the distant past between acquisition and erosion.

HGT vs. DL

Ignoring the cumulative evidence from this and many other studies, one may still dismiss the phylogenetic inference as mere assembly artefact and overlook all the significant differences and trends between native genes and HGT candidates. This could be done by superimposing vertical inheritance (and thus eukaryotic origin) on all HGT events outside the context of pathogenicity and endosymbiosis. Under this extreme view, all extant genes would have their roots in LECA. Consequently, patchy phylogenetic distributions are the result of multiple putative ancient paralogs existing in the LECA followed by mutation, gene duplication, and gene loss. Following this line of reasoning, all HGT candidates in the Cyanidiales would be the product of DL acting on all other eukaryotic species, with the exception of the Cyanidiales, *Galdieria* and/or *Cyanidioschyzon* (**Figure 5A–C**). However, we found cases where either *Galdieria* HGT candidates (six orthogroups), or *Cyanidioschyzon* HGT candidates (eight orthogroups) show non-eukaryotic origin, whereas the others cluster within the eukaryotic branch (**Figure 5E–F**). In addition, we find two cases in which *Galdieria* and *Cyanidioschyzon* HGT candidates are located in different non-eukaryotic branches (**Figure 5D**). DL would require LECA to have encoded three paralogs of the same gene, one of which was retained by *Cyanidioschyzon*, another by *Galdieria*, whereas the third by all other eukaryotes. The number of required paralogs in the LECA would be further increased when taking into consideration that some ancient paralogs of LECA may have been eroded in all eukaryotes and that eukaryote phylogenies are not always monophyletic which would additionally increase the number of required paralogs in the LECA in order to explain the current pattern. The strict superimposition of vertical inheritance would thus require a complex LECA, an issue known as 'the genome of Eden'.

Cumulative effects are observed when genes derived from HGT increasingly diverge as a function of time. Hence, a gradual increase in protein identity towards their non-eukaryotic donor species is expected the more recent an individual HGT event is. The absence of cumulative effects in eukaryotic HGT studies has been used as argument in favor of strict vertical inheritance followed by DL. Here, we also did not find evidence for cumulative effects in the HGT dataset. 'Recent' HGT events that are exclusive to either the *Cyanidioschyzon* or *Galdieria* lineage shared 5% higher PID with their potential non-eukaryotic donors in comparison to ancient HGT candidates that predate the split, but this difference was not significant (**Figure 4C**). We also tested for cumulative effects between the number of species contained in orthogroups compared to the percent protein identity shared with potential non-eukaryotic donors under the assumption that recent HGT events would be present in fewer species in comparison to ancient HGT events that occurred at the root of *Galdieria* (**Figure 6D**). Neither a gradual increase in protein identity for potentially recent HGT events, nor a general trend could be determined. Only orthogroups containing one *Galdieria* species reported a statistically significant higher protein identity to their potential non-eukaryotic donors which could be an indication of 'most recent' HGT.

Whereas the absence of cumulative effects may speak against HGT, this does not automatically argue in favor of strict vertical inheritance followed by DL. Here, the null hypothesis would be that no differences exist between HGT genes and native genes because all genes are descendants of LECA. This null hypothesis is rejected on multiple levels. At the molecular level, the HGT subset differs significantly from native genes with respect to various genomic and molecular features (e.g., GC-content, frequency of multiexonic genes, number of exons per gene, responsiveness to temperature stress) (**Table 1**, **Figure 3**). Furthermore, HGT candidates in *Galdieria* are significantly more similar (6.1% average PID) to their potential non-eukaryotic donors when compared to native genes and non-eukaryotic sequences in the same orthogroup (**Figure 6B**). This difference cannot be explained by the absence of eukaryotic orthologs. We also find significant differences in PID with regard to non-eukaryotic sequences between HGT and native genes in orthogroups containing either one *Galdieria* sequence, or all eleven *Galdieria* sequences regarding (**Figure 6D**). Hence, the 'most recent' and 'most ancient' HGT candidates share the highest resemblance to their non-eukaryotic donors, which is also significantly higher when compared to native genes in OGs of the same size. Intriguingly, a general trend towards 'cumulative effects' could be observed for native genes, highlighting the differences between these two gene sources in Cyanidiales.

Given these results and interpretations, we advocate the following view of eukaryotic HGT. Specifically, two forces may act simultaneously on HGT candidates in eukaryotes. The first is strong evolutionary pressure for adaptation of eukaryotic genetic features and compatibility with native replication and transcriptional mechanisms to ensure integration into existing metabolic circuits (e.g., codon usage, splice sites, methylation, pH differences in the cytosol). The second however is that key structural aspects of HGT-derived sequence cannot be significantly altered by the first process because they ensure function of the transferred gene (e.g., protein domain conservation, three-dimensional structure, ligand interaction). Consequently, HGT candidates may suffer more markedly from gene erosion than native genes due to these countervailing forces, in spite of potentially providing beneficial adaptive traits. This view suggests that we need to think about eukaryotic HGT in fundamentally different ways than is the case for prokaryotes, necessitating a taxonomically broad genome-based approach that is slowly taking hold.

In summary, we do not discount the importance of DL in eukaryotic evolution because it can impact ca. 99% of the gene inventory in Cyanidiales. What we strongly espouse is that strict vertical inheritance in combination with DL cannot explain all the data. HGTs in Cyanidiales are significant because the 1% (values will vary across different eukaryotic lineages) helps explain the remarkable evolutionary history of these extremophiles. Lastly, we question the validity of the premise regarding the applicability of cumulative effects in the prokaryotic sense to eukaryotic HGT. The absence of cumulative effects and a eukaryotic pan-genome are neither arguments in favor of HGT, nor DL.

Materials and methods

Cyanidiales strains used for draft genomic sequencing

Ten Cyanidiales strains (**Figure 1**) were sequenced in 2016/2017 using the PacBio RS2 (Pacific Biosciences Inc, Menlo Park, CA) technology (**Rhoads and Au, 2015**) and P6-C4 chemistry (the only exception being *C. merolae* Soos, which was sequenced as a pilot study using P4-C2 chemistry in 2014). Seven strains, namely *G. sulphuraria* 5572, *G. sulphuraria* 002, *G. sulphuraria* SAG21.92, *G. sulphuraria* Azora, *G. sulphuraria* MtSh, *G. sulphuraria* RT22 and *G. sulphuraria* MS1 were sequenced at the University of Maryland Institute for Genome Sciences (Baltimore, MD). The remaining three strains, *G. sulphuraria* YNP5587.1, *G. phlegrea* Soos, and *C. merolae* Soos were sequenced at the Max-Planck-Institut für Pflanzenzüchtungsforschung (Cologne, Germany). To obtain axenic and monoclonal genetic material for sequencing, single colonies of each strain were grown at 37˚C in the dark on plates containing glucose as the sole carbon source (1% Gelrite mixed 1:1 with 2x Allen medium [**Allen, 1959**], 50 µM Glucose). The purity of single colonies was assessed using microscopy (Zeiss Axio Imager 2, 1000x) and molecular markers (18S, *rbcL*). Long-read compatible DNA was extracted using a genomic-tip 20/G column following the steps of the 'YEAST' DNA extraction protocol (QIAGEN N.V., Hilden, Germany). The size and quality of DNA were assessed via gel electrophoresis and the Nanodrop instrument (Thermo Fisher Scientific Inc, Waltham, MA).

Assembly

All genomes (excluding the already published *G. sulphuraria* 074W, *G. phlegrea* DBV009 and *C. merolae* 10D) were assembled using canu version 1.5 (**Koren et al., 2017**). The genomic sequences were polished three times using the Quiver algorithm (**Chin et al., 2013**). Different versions of each genome were assessed using BUSCO v.3 (**Simão et al., 2015**) and the best performing genome was chosen as reference for gene model prediction. Each genome was queried against the National

Center for Biotechnology Information (NCBI) nr database (**Geer et al., 2010**) in order to detect contigs consisting exclusively of bacterial best blast hits (i.e., possible contamination). None were found.

Gene prediction

Gene and protein models for the 10 sequenced Cyanidiales were predicted using MAKER v3 beta (**Cantarel et al., 2008**). MAKER was trained using existing protein sequences from *Cyanidioschyzon merolae* 10D and *Galdieria sulphuraria* 074W, for which we used existing RNA-Seq (**Rossoni, 2018**) data with expression values > 10 FPKM (**Rademacher et al., 2016**) combined with protein sequences from the UniProtKB/Swiss-Prot protein database (**UniProt Consortium T, 2018**). Augustus (**Stanke and Morgenstern, 2005**), GeneMark ES (**Borodovsky and Lomsadze, 2011**), and EVM (**Haas et al., 2008**) were used for gene prediction. MAKER was run iteratively and using various options for each genome. The resulting gene models were again assessed using BUSCO v.3 (**Simão et al., 2015**) and PFAM 31.0 (**Finn et al., 2016**). The best performing set of gene models was chosen for each species.

Sequence annotation

The transcriptomes of all sequenced species and those of *Cyanidioschyzon merolae* 10D, *Galdieria sulphuraria* 074W, and *Galdieria phlegrea* DB10 were annotated (re-annotated) using BLAST2GO PRO v.5 (**Götz et al., 2008**) combined with INTERPROSCAN (**Jones et al., 2014**) in order to obtain the annotations, Gene Ontology (GO)-Terms (**Ashburner et al., 2000**), and Enzyme Commission (EC)-Numbers (**Bairoch, 2000**). KEGG orthology identifiers (KO-Terms) were obtained using KAAS (**Ogata et al., 1999**; **Moriya et al., 2007**) and PFAM annotations using PFAM 31.0 (**Finn et al., 2016**).

Orthogroups and phylogenetic analysis

The 81,682 predicted protein sequences derived from the 13 genomes listed in **Table 1** were clustered into orthogroups (OGs) using OrthoFinder v. 2.2 (**Emms and Kelly, 2015**). We queried each OG member using DIAMOND v. 0.9.22 (**Buchfink et al., 2015**) to an in-house database comprising NCBI RefSeq sequences with the addition of predicted algal proteomes available from the JGI Genome Portal (**Nordberg et al., 2014**), TBestDB (**O'Brien et al., 2007**), dbEST (**Boguski et al., 1993**), and the MMETSP (Moore Microbial Eukaryote Transcriptome Sequencing Project) (**Keeling et al., 2014**). The database was partitioned into four volumes: Bacteria, Metazoa, remaining taxa, and the MMETSP data. To avoid taxonomic sampling biases due to under or overabundance of particular lineages in the database, each volume was queried independently with an expect (*e*-value) of 1 × $10^{-5,}$ and the top 2000 hits were saved and combined into a single list that was then sorted by descending DIAMOND bitscore. Proteins containing one or more bacterial hits (and thus possible HGT candidates) were retained for further analysis, whereas those lacking bacterial hits were removed. A taxonomically broad list of hits was selected for each query (the maximum number of genera selected for each taxonomic phylum present in the DIAMOND output was equivalent to 180 divided by the number of unique phyla), and the corresponding sequences were extracted from the database and aligned using MAFFT v7.3 (**Katoh and Standley, 2013**) together with queries and hits selected in the same manner for remaining proteins assigned to the same OG (duplicate hits were removed). A maximum-likelihood phylogeny was then constructed for each alignment using IQTREE v7.3 (**Nguyen et al., 2015**) under automated model selection, with node support calculated using 2000 ultrafast bootstraps. Single-gene trees for the referenced HGT candidates from previous research regarding *G. sulphuraria* 074W (**Schönknecht et al., 2013**) and *G. phlegrea* DBV009 (**Qiu et al., 2013**) were constructed in the same manner, without assignment to OG. To create the algal species tree, the OG assignment was re-run with the addition of proteomes from outgroup taxa *Porphyra umbilicalis* (**Brawley et al., 2017**), *Porphyridium purpureum* (**Bhattacharya et al., 2013**), *Ostreococcus tauri* RCC4221 (**Blanc-Mathieu et al., 2014**), and *Chlamydomonas reinhardtii* (**Merchant et al., 2007**). Orthogroups were parsed and 2090 were selected that contained single-copy representative proteins from at least 12/17 taxa; those taxa with multi-copy representatives were removed entirely from the OG. The proteins for each OG were extracted and aligned with MAFFT, and IQTREE was used to construct a single maximum-likelihood phylogeny via a partitioned analysis in which each OG alignment represented one partition with unlinked models of protein evolution chosen by IQTREE. Consensus tree branch support was determined by 2,000 UF bootstraps.

Detection of HGTs

All phylogenies containing bacterial sequences were inspected manually. Only trees in which there were at least two different Cyanidiales sequences and at least three different non-eukaryotic donors were retained. The singleton HGT candidates in Cyanidiales are presented in the appendix (Appendix 5) and were not analyzed further here. Phylogenies with cyanobacteria and Chlamydiae as sisters were considered as EGT and excluded from the analysis. Genes that were potentially transferred from cyanobacteria were only accepted as HGT candidates when homologs were absent in other photosynthetic eukaryotes; that is the cyanobacterium was not the closest neighbor, and when the annotation did not include a photosynthetic function, to discriminate from EGT. Furthermore, phylogenies containing inconsistencies within the distribution patterns of species, especially at the root, or UF values below 70% spanning over multiple nodes, were excluded. Each orthogroup was queried against NCBI nr to detect eukaryotic homologs not present in our databases. The conservative approach to HGT assignment used here allowed identification of robust candidates for in-depth analysis. This may however have come at the cost of underestimating HGT at the single species level. Furthermore, some of the phylogenies that were rejected because <3 non-eukaryotic donors were found may have resulted from current incomplete sampling of prokaryotes. For example, OG0001817 is present in the sister species *G. sulphuraria* 074W and *G. sulphuraria* MS1 but has a single bacterial hit (*Acidobacteriaceae bacterium* URHE0068, CBS domain-containing protein, GI:651323331).

Data deposit

The nuclear, plastid, and mitochondrial sequences of the 10 novel genomes, as well as gene models, ESTs, protein sequences, protein alignments, orthogroup and single gene trees, and gene annotations are available at http://porphyra.rutgers.edu. Raw PacBio RSII reads, and also the genomic, chloroplast and mitochondrial sequences, have been submitted to the NCBI and are retrievable via BioProject ID PRJNA512382.

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (under Germany´s Excellence Strategy – EXC-2048/1 – Project ID: 390686111 and EXC 1028, and WE 2231/21–1 to APMW) and by the Heine Research Academy (AWR). We thank Dr. Luke Tallon and Dr. Bruno Hüttel for the excellent technical assistance for PacBio sequencing. Dr. Marion Eisenhut for her fantastic Power Point templates. DCP and DB are grateful to the New Jersey Agricultural Experiment Station and the Rutgers University School of Environmental and Biological Sciences Genome Cooperative for supporting this research.

Additional information

Funding

Funder	Grant reference number	Author
Deutsche Forschungsgemeinschaft	EXC 1028	Andreas P M Weber
Heinrich-Heine-Universität Düsseldorf		Andreas P M Weber
Deutsche Forschungsgemeinschaft	WE 2231/21-1	Andreas P M Weber

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Alessandro W Rossoni, Conceptualization, Resources, Data curation, Software, Formal analysis, Funding acquisition, Validation, Investigation, Visualization, Methodology, Writing—original draft, Project administration; Dana C Price, Resources, Data curation, Software, Formal analysis, Methodology, Writing—original draft; Mark Seger, Conceptualization, Resources, Investigation, Methodology, Writing—review and editing; Dagmar Lyska, Formal analysis, Validation, Methodology, Writing—original draft; Peter Lammers, Conceptualization, Resources, Funding acquisition, Writing—review and editing; Debashish Bhattacharya, Conceptualization, Resources, Data curation, Formal analysis, Supervision, Validation, Methodology, Writing—review and editing; Andreas PM Weber, Conceptualization, Resources, Data curation, Supervision, Funding acquisition, Methodology, Project administration, Writing—review and editing

Author ORCIDs

Alessandro W Rossoni https://orcid.org/0000-0003-4716-6799
Debashish Bhattacharya https://orcid.org/0000-0003-0611-1273
Andreas PM Weber https://orcid.org/0000-0003-0970-4672

Decision letter and Author response

Decision letter https://doi.org/10.7554/eLife.45017.151
Author response https://doi.org/10.7554/eLife.45017.152

Additional files

Supplementary files

- Transparent reporting form

DOI: https://doi.org/10.7554/eLife.45017.107

Data availability

The genomic, chloroplast and mitochondrial sequences of the 10 novel genomes, as well as gene models, ESTs, protein sequences, and gene annotations are available at http://porphyra.rutgers.edu. These data have also be uploaded to Dryad doi:10.5061/dryad.m06n200. Raw PacBio RSII reads, and also the genomic, chloroplast and mitochondrial sequences, have been submitted to the NCBI and are retrievable via BioProject ID PRJNA512382.

The following datasets were generated:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Alessandro W Rossoni, Dana C Price, Mark Seger, Dagmar Lyska, Peter Lammers, Debashish Bhattacharya, Andreas PM Weber	2019	Genome sequencing of 10 novel Cyanidiales strains	https://www.ncbi.nlm.nih.gov/bioproject/PRJNA512382/	NCBI Sequence Read Archive, PRJNA512382
Alessandro W Rossoni, Dana C Price, Mark Seger, Dagmar Lyska, Peter Lammers, Debashish Bhattacharya, Andreas PM Weber	2019	Data from: The genomes of polyextremophilic Cyanidiales contain 1% horizontally transferred genes with diverse adaptive functions	http://dx.doi.org/10.5061/dryad.m06n200	Dryad Digital Repository, 10.5061/dryad.m06n200
Alessandro W Rossoni, Dana C Price, Mark Seger, Dagmar Lyska, Peter Lammers, Debashish Bhattacharya, Andreas PM Weber	2019	Red Algal Resources to Promote Integrative Research in Algal Genomics	http://porphyra.rutgers.edu/bindex.php	Rutgers University, Red Algal

References

Aiuppa A. 2003. The aquatic geochemistry of arsenic in volcanic groundwaters from southern italy. *Science* **18**: 1283–1296. DOI: https://doi.org/10.1016/S0883-2927(03)00051-9

Albertano P. 2000. The taxonomic position of cyanidium, cyanidioschyzon and galdieria: an update. *Hydrobiologia* **433**:137–143. DOI: https://doi.org/10.1023/A:1004031123806

Allen MB. 1959. Studies with Cyanidium caldarium, *an anomalously pigmented chlorophyte.*In: *Archiv Für Mikrobiologie* **32**:270–277. DOI: https://doi.org/10.1007/BF00409348

Apfel CM, Takács B, Fountoulakis M, Stieger M, Keck W. 1999. Use of genomics to identify bacterial undecaprenyl pyrophosphate synthetase: cloning, expression, and characterization of the essential uppS gene. *Journal of Bacteriology* **181**:483–492. PMID: 9882662

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. 2000. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature Genetics* **25**:25–29. DOI: https://doi.org/10.1038/75556, PMID: 10802651

Bairoch A. 2000. The ENZYME database in 2000. *Nucleic Acids Research* **28**:304–305. DOI: https://doi.org/10.1093/nar/28.1.304, PMID: 10592255

Barcyté D, Elster J, Nedbalová L. 2018. Plastid-encoded *rbc* L phylogeny suggests widespread distribution of *galdieria phlegrea* (Cyanidiophyceae, rhodophyta). *Nordic Journal of Botany* **36**:e01794. DOI: https://doi.org/10.1111/njb.01794

Bhattacharya D, Price DC, Chan CX, Qiu H, Rose N, Ball S, Weber AP, Arias MC, Henrissat B, Coutinho PM, Krishnan A, Zäuner S, Morath S, Hilliou F, Egizi A, Perrineau MM, Yoon HS. 2013. Genome of the red alga porphyridium purpureum. *Nature Communications* **4**. DOI: https://doi.org/10.1038/ncomms2931, PMID: 23770768

Blanc-Mathieu R, Verhelst B, Derelle E, Rombauts S, Bouget FY, Carré I, Château A, Eyre-Walker A, Grimsley N, Moreau H, Piégu B, Rivals E, Schackwitz W, Van de Peer Y, Piganeau G. 2014. An improved genome of the model marine alga ostreococcus tauri unfolds by assessing Illumina de novo assemblies. *BMC Genomics* **15**: 1103. DOI: https://doi.org/10.1186/1471-2164-15-1103, PMID: 25494611

Boguski MS, Lowe TMJ, Tolstoshev CM. 1993. dbEST — database for "expressed sequence tags". *Nature Genetics* **4**:332–333. DOI: https://doi.org/10.1038/ng0893-332

Boothby TC, Tenlen JR, Smith FW, Wang JR, Patanella KA, Nishimura EO, Tintori SC, Li Q, Jones CD, Yandell M, Messina DN, Glasscock J, Goldstein B. 2015. Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. *PNAS* **112**:15976–15981. DOI: https://doi.org/10.1073/pnas.1510461112, PMID: 26598659

Borodovsky M, Lomsadze A. 2011. Eukaryotic gene prediction using GeneMark.hmm-E and GeneMark-ES. *Current Protocols in Bioinformatics* **35**:4.6.1–4.6.4. DOI: https://doi.org/10.1002/0471250953.bi0406s35

Boucher Y, Douady CJ, Papke RT, Walsh DA, Boudreau ME, Nesbø CL, Case RJ, Doolittle WF. 2003. Lateral gene transfer and the origins of prokaryotic groups. *Annual Review of Genetics* **37**:283–328. DOI: https://doi.org/10.1146/annurev.genet.37.050503.084247, PMID: 14616063

Brawley SH, Blouin NA, Ficko-Blean E, Wheeler GL, Lohr M, Goodson HV, Jenkins JW, Blaby-Haas CE, Helliwell KE, Chan CX, Marriage TN, Bhattacharya D, Klein AS, Badis Y, Brodie J, Cao Y, Collén J, Dittami SM, Gachon CMM, Green BR, et al. 2017. Insights into the red algae and eukaryotic evolution from the genome of *porphyra umbilicalis* (Bangiophyceae, Rhodophyta). *PNAS* **114**:E6361–E6370. DOI: https://doi.org/10.1073/pnas.1703088114, PMID: 28716924

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. *Nature Methods* **12**: 59–60. DOI: https://doi.org/10.1038/nmeth.3176

Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Sánchez Alvarado A, Yandell M. 2008. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research* **18**: 188–196. DOI: https://doi.org/10.1101/gr.6743907, PMID: 18025269

Castenholz RW, McDermott TR. 2010. The Cyanidiales: ecology, biodiversity, and biogeography. In: Seckbach J, Chapman D. J (Eds). *Red Algae in the Genomic Age*. Springer. p. 357–371. DOI: https://doi.org/10.1007/978-90-481-3795-4_19

Catarecha P, Segura MD, Franco-Zorrilla JM, Garcia-Ponce B, Lanza M, Solano R, Paz-Ares J, Leyva A. 2007. A mutant of the arabidopsis phosphate transporter PHT1;1 displays enhanced arsenic accumulation. *The Plant Cell Online* **19**:1123–1133. DOI: https://doi.org/10.1105/tpc.106.041871

Chae HZ, Robison K, Poole LB, Church G, Storz G, Rhee SG. 1994. Cloning and sequencing of thiol-specific antioxidant from mammalian brain: alkyl hydroperoxide reductase and thiol-specific antioxidant define a large family of antioxidant enzymes. *PNAS* **91**:7017–7021. DOI: https://doi.org/10.1073/pnas.91.15.7017, PMID: 8041738

Chen YC, Peoples OP, Walsh CT. 1988. Acinetobacter cyclohexanone monooxygenase: gene cloning and sequence determination. *Journal of Bacteriology* **170**:781–789. DOI: https://doi.org/10.1128/jb.170.2.781-789.1988, PMID: 3338974

Chen Y, Li F, Wurtzel ET. 2010. Isolation and characterization of the Z-ISO gene encoding a missing component of carotenoid biosynthesis in plants. *Plant Physiology* **153**:66–79. DOI: https://doi.org/10.1104/pp.110.153916, PMID: 20335404

Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods* **10**:563–569. DOI: https://doi.org/10.1038/nmeth.2474, PMID: 23644548

Ciniglia C, Yoon HS, Pollio A, Pinto G, Bhattacharya D. 2004. Hidden biodiversity of the extremophilic cyanidiales red algae. *Molecular Ecology* **13**:1827–1838. DOI: https://doi.org/10.1111/j.1365-294X.2004.02180.x, PMID: 15189206

Crisp A, Boschetti C, Perry M, Tunnacliffe A, Micklem G. 2015. Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes. *Genome Biology* **16**:50. DOI: https://doi.org/10.1186/s13059-015-0607-3, PMID: 25785303

Dahl SW, Slaughter C, Lauritzen C, Bateman RC, Connerton I, Pedersen J. 2000. Carica papaya glutamine cyclotransferase belongs to a novel plant enzyme subfamily: cloning and characterization of the recombinant enzyme. *Protein Expression and Purification* **20**:27–36. DOI: https://doi.org/10.1006/prep.2000.1273, PMID: 11035947

Danchin EG. 2016. Lateral gene transfer in eukaryotes: tip of the iceberg or of the ice cube? *BMC Biology* **14**: 101. DOI: https://doi.org/10.1186/s12915-016-0330-x, PMID: 27863503

De Luca P, Taddei R, Varano L. 1978. « Cyanidioschyzon merolae »: a new alga of thermal acidic environments. *Webbia* **33**:37–44. DOI: https://doi.org/10.1080/00837792.1978.10670110

Diderichsen BÃ¸rge, Christiansen L. 1988. Cloning of a maltogenic alpha-amylase from Bacillus stearothermophilus. *FEMS Microbiology Letters* **56**:53–60. DOI: https://doi.org/10.1111/j.1574-6968.1988.tb03149.x

Dietz K-J, Baier M, Krämer U. 1999. Free radicals and reactive oxygen species as mediators of heavy metal toxicity in plants. In: *Heavy Metal Stress in Plants*. Springer. p. 73–97. DOI: https://doi.org/10.1007/978-3-662-07745-0_4

Dilworth D, Gudavicius G, Leung A, Nelson CJ. 2012. The roles of peptidyl-proline isomerases in gene regulation. *Biochemistry and Cell Biology = Biochimie Et Biologie Cellulaire* **90**:55–69. DOI: https://doi.org/10.1139/o11-045, PMID: 21999350

Doemel WN, Brock T. 1971. The physiological ecology of *Cyanidium caldarium*. *Microbiology* **67**:17–32. DOI: https://doi.org/10.1099/00221287-67-1-17

Doolittle WF. 1999. Lateral genomics. *Trends in Cell Biology* **9**:M5–M8. DOI: https://doi.org/10.1016/S0962-8924(99)01664-5, PMID: 10611671

Doolittle WF, Brunet TD. 2016. What is the tree of life? *PLOS Genetics* **12**:e1005912. DOI: https://doi.org/10.1371/journal.pgen.1005912, PMID: 27078870

Ducker GS, Rabinowitz JD. 2017. One-Carbon metabolism in health and disease. *Cell Metabolism* **25**:27–42. DOI: https://doi.org/10.1016/j.cmet.2016.08.009, PMID: 27641100

Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* **16**:157. DOI: https://doi.org/10.1186/s13059-015-0721-2, PMID: 26243257

Ettema TJ, Ahmed H, Geerling AC, van der Oost J, Siebers B. 2008. The non-phosphorylating glyceraldehyde-3-phosphate dehydrogenase (GAPN) of sulfolobus solfataricus: a key-enzyme of the semi-phosphorylative branch of the Entner-Doudoroff pathway. *Extremophiles* **12**:75–88. DOI: https://doi.org/10.1007/s00792-007-0082-1, PMID: 17549431

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A. 2016. The pfam protein families database: towards a more sustainable future. *Nucleic Acids Research* **44**:D279–D285. DOI: https://doi.org/10.1093/nar/gkv1344, PMID: 26673716

Foflonker F, Mollegard D, Ong M, Yoon HS, Bhattacharya D. 2018. Genomic analysis of picochlorum species reveals how microalgae may adapt to variable environments. *Molecular Biology and Evolution* **408**. DOI: https://doi.org/10.1093/molbev/msy167

Fukuda S-ya, Yamamoto R, Iwamoto K, Minoda A. 2018. Cellular accumulation of cesium in the unicellular red alga *Galdieria sulphuraria* under mixotrophic conditions. *Journal of Applied Phycology* **30**:3057–3061. DOI: https://doi.org/10.1007/s10811-018-1525-z

Galperin MY. 2005. A census of membrane-bound and intracellular signal transduction proteins in bacteria: bacterial IQ, extroverts and introverts. *BMC Microbiology* **5**:35. DOI: https://doi.org/10.1186/1471-2180-5-35, PMID: 15955239

Geer LY, Marchler-Bauer A, Geer RC, Han L, He J, He S, Liu C, Shi W, Bryant SH. 2010. The NCBI BioSystems database. *Nucleic Acids Research* **38**:D492–D496. DOI: https://doi.org/10.1093/nar/gkp858, PMID: 19854944

Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talón M, Dopazo J, Conesa A. 2008. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research* **36**:3420–3435. DOI: https://doi.org/10.1093/nar/gkn176, PMID: 18445632

Gross W, Schnarrenberger CJP, Physiology C. 1995. Heterotrophic growth of two strains of the acido-thermophilic red alga galdieria sulphuraria. *Plant and Cell Physiology* **36**:633–638. DOI: https://doi.org/10.1093/oxfordjournals.pcp.a078803

Gross W. 1998. Cryptoendolithic growth of the red alga galdieria sulphuraria in volcanic areas. *European Journal of Phycology* **33**:25–31. DOI: https://doi.org/10.1017/S0967026297001467

Gross W, Oesterhelt C, Tischendorf G, Lederer F. 2002. Characterization of a non-thermophilic strain of the red algal genus *Galdieria* isolated from Soos (Czech Republic) . *European Journal of Phycology* **37**:477–482. DOI: https://doi.org/10.1017/S0967026202003773

Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biology* **9**:R7. DOI: https://doi.org/10.1186/gb-2008-9-1-r7, PMID: 18190707

Henkanatte-Gedera SM, Selvaratnam T, Karbakhshravari M, Myint M, Nirmalakhandan N, Van Voorhies W, Lammers PJ. 2017. Removal of dissolved organic carbon and nutrients from urban wastewaters by galdieria sulphuraria: laboratory to field scale demonstration. *Algal Research* **24**:450–456. DOI: https://doi.org/10.1016/j.algal.2016.08.001

Hsieh CJ, Zhan SH, Liao CP, Tang SL, Wang LC, Watanabe T, Geraldino PJL, Liu SL. 2018. The effects of contemporary selection and dispersal limitation on the community assembly of acidophilic microalgae. *Journal of Phycology* **54**:720–733. DOI: https://doi.org/10.1111/jpy.12771, PMID: 30055054

Husnik F, McCutcheon JP. 2018. Functional horizontal gene transfer from bacteria to eukaryotes. *Nature Reviews Microbiology* **16**:67–79. DOI: https://doi.org/10.1038/nrmicro.2017.137, PMID: 29176581

Imhoff JF, Rodriguez-Valera F. 1984. Betaine is the main compatible solute of halophilic eubacteria. *Journal of Bacteriology* **160**:478–479. PMID: 6148337

Iovinella M, Eren A, Pinto G, Pollio A, Davis SJ, Cennamo P, Ciniglia C. 2018. Cryptic dispersal of cyanidiophytina (Rhodophyta) in non-acidic environments from Turkey. *Extremophiles* **22**:713–723. DOI: https://doi.org/10.1007/s00792-018-1031-x, PMID: 29779132

Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong SY, Lopez R, Hunter S. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**:1236–1240. DOI: https://doi.org/10.1093/bioinformatics/btu031, PMID: 24451626

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30**:772–780. DOI: https://doi.org/10.1093/molbev/mst010, PMID: 23329690

Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, Armbrust EV, Archibald JM, Bharti AK, Bell CJ, Beszteri B, Bidle KD, Cameron CT, Campbell L, Caron DA, Cattolico RA, Collier JL, Coyne K, Davy SK, Deschamps P, et al. 2014. The marine microbial eukaryote transcriptome sequencing project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLOS Biology* **12**:e1001889. DOI: https://doi.org/10.1371/journal.pbio.1001889, PMID: 24959919

Kelly DJ, Budd K, Lefebvre DD. 2007. Biotransformation of mercury in pH-stat cultures of eukaryotic freshwater algae. *Archives of Microbiology* **187**:45–53. DOI: https://doi.org/10.1007/s00203-006-0170-0, PMID: 17031617

Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods* **12**:357–360. DOI: https://doi.org/10.1038/nmeth.3317, PMID: 25751142

Kimura S, Suzuki T. 2010. Fine-tuning of the ribosomal decoding center by conserved methyl-modifications in the Escherichia coli 16S rRNA. *Nucleic Acids Research* **38**:1341–1352. DOI: https://doi.org/10.1093/nar/gkp1073, PMID: 19965768

Kingston RL, Scopes RK, Baker EN. 1996. The structure of glucose-fructose oxidoreductase from zymomonas mobilis: an osmoprotective periplasmic enzyme containing non-dissociable NADP. *Structure* **4**:1413–1428. DOI: https://doi.org/10.1016/S0969-2126(96)00149-9, PMID: 8994968

Kloosterman TG, Hendriksen WT, Bijlsma JJ, Bootsma HJ, van Hijum SA, Kok J, Hermans PW, Kuipers OP. 2006. Regulation of glutamine and glutamate metabolism by GlnR and GlnA in streptococcus pneumoniae. *Journal of Biological Chemistry* **281**:25097–25109. DOI: https://doi.org/10.1074/jbc.M601661200, PMID: 16787930

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Research* **27**:722–736. DOI: https://doi.org/10.1101/gr.215087.116, PMID: 28298431

Koutsovoulos G, Kumar S, Laetsch DR, Stevens L, Daub J, Conlon C, Maroon H, Thomas F, Aboobaker AA, Blaxter M. 2016. No evidence for extensive horizontal gene transfer in the genome of the tardigrade *hypsibius dujardini*. *PNAS* **113**:5053–5058. DOI: https://doi.org/10.1073/pnas.1600338113, PMID: 27035985

Ku C, Martin WF. 2016. A natural barrier to lateral gene transfer from prokaryotes to eukaryotes revealed from genomes: the 70 % rule. *BMC Biology* **14**:89. DOI: https://doi.org/10.1186/s12915-016-0315-9, PMID: 27751184

Lee HS, Kim MS, Cho HS, Kim JI, Kim TJ, Choi JH, Park C, Lee HS, Oh BH, Park KH. 2002. Cyclomaltodextrinase, Neopullulanase, and maltogenic amylase are nearly indistinguishable from each other. *Journal of Biological Chemistry* **277**:21891–21897. DOI: https://doi.org/10.1074/jbc.M201623200, PMID: 11923309

Lee J, Kim KM, Yang EC, Miller KA, Boo SM, Bhattacharya D, Yoon HS. 2016. Reconstructing the complex evolutionary history of mobile plasmids in red algal genomes. *Scientific Reports* **6**:23744. DOI: https://doi.org/10.1038/srep23744, PMID: 27030297

Leger MM. 2018. Demystifying eukaryote lateral gene transfer (Response to martin 2017 DOI: 10.1002/bies.201700115). *BioEssays* **40**:e1700242. DOI: https://doi.org/10.1002/bies.201700242

Li Z, Bock R. 2018. Replication of bacterial plasmids in the nucleus of the red alga porphyridium purpureum. *Nature Communications* **9**:3451. DOI: https://doi.org/10.1038/s41467-018-05651-1, PMID: 30150628

Liu Y, Promeneur D, Rojek A, Kumar N, Frøkiaer J, Nielsen S, King LS, Agre P, Carbrey JM. 2007. Aquaporin 9 is the major pathway for glycerol uptake by mouse erythrocytes, with implications for malarial virulence. *PNAS* **104**:12560–12564. DOI: https://doi.org/10.1073/pnas.0705313104, PMID: 17636116

Lu WD, Chi ZM, Su CD. 2006. Identification of glycine betaine as compatible solute in synechococcus sp. WH8102 and characterization of its N-methyltransferase genes involved in betaine synthesis. *Archives of Microbiology* **186**:495–506. DOI: https://doi.org/10.1007/s00203-006-0167-8, PMID: 17019606

Marnett LJ. 2000. Oxyradicals and DNA damage. *Carcinogenesis* **21**:361–370. DOI: https://doi.org/10.1093/carcin/21.3.361, PMID: 10688856

Martin WF. 2017. Too much eukaryote LGT. *BioEssays* **39**:1700115. DOI: https://doi.org/10.1002/bies.201700115

Martin WF. 2018. Eukaryote lateral gene transfer is lamarckian. *Nature Ecology & Evolution* **2**:754. DOI: https://doi.org/10.1038/s41559-018-0521-7, PMID: 29535449

Martin PR, Mulks MH. 1992. Sequence analysis and complementation studies of the argJ gene encoding ornithine acetyltransferase from neisseria gonorrhoeae. *Journal of Bacteriology* **174**:2694–2701. DOI: https://doi.org/10.1128/jb.174.8.2694-2701.1992, PMID: 1339419

Mashek DG, Li LO, Coleman RA. 2007. Long-chain acyl-CoA synthetases and fatty acid channeling. *Future Lipidology* **2**:465–476. DOI: https://doi.org/10.2217/17460875.2.4.465, PMID: 20354580

Matsuzaki M, Misumi O, Shin-I T, Maruyama S, Takahara M, Miyagishima SY, Mori T, Nishida K, Yagisawa F, Nishida K, Yoshida Y, Nishimura Y, Nakao S, Kobayashi T, Momoyama Y, Higashiyama T, Minoda A, Sano M, Nomoto H, Oishi K, et al. 2004. Genome sequence of the ultrasmall unicellular red alga cyanidioschyzon merolae 10D. *Nature* **428**:653–657. DOI: https://doi.org/10.1038/nature02398, PMID: 15071595

McCoy JG, Bailey LJ, Ng YH, Bingman CA, Wrobel R, Weber AP, Fox BG, Phillips GN. 2009. Discovery of sarcosine dimethylglycine methyltransferase from *galdieria sulphuraria*. *Proteins: Structure, Function, and Bioinformatics* **74**:368–377. DOI: https://doi.org/10.1002/prot.22147, PMID: 18623062

Meharg AA, MacNair MR. 1992. Suppression of the High Affinity Phosphate Uptake System: A Mechanism of Arsenate Tolerance in *Holcus lanatus* L. . *Journal of Experimental Botany* **43**:519–524. DOI: https://doi.org/10.1093/jxb/43.4.519

Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, Witman GB, Terry A, Salamov A, Fritz-Laylin LK, Maréchal-Drouard L, Marshall WF, Qu LH, Nelson DR, Sanderfoot AA, Spalding MH, Kapitonov VV, Ren Q, Ferris P, Lindquist E, Shapiro H, et al. 2007. The chlamydomonas genome reveals the evolution of key animal and plant functions. *Science* **318**:245–250. DOI: https://doi.org/10.1126/science.1143609, PMID: 17932292

Moreira D, LÃ³pez-Archilla A-I, Amils R, MarÃn I. 1994. Characterization of two new thermoacidophilic microalgae: Genome organization and comparison with *Galdieria sulphuraria* . *FEMS Microbiology Letters* **122**:109–114. DOI: https://doi.org/10.1111/j.1574-6968.1994.tb07152.x

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. 2007. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Research* **35**:W182–W185. DOI: https://doi.org/10.1093/nar/gkm321, PMID: 17526522

Nautiyal A, Rani PS, Sharples GJ, Muniyappa K. 2016. Mycobacterium tuberculosis RuvX is a holliday junction resolvase formed by dimerisation of the monomeric YqgF nuclease domain. *Molecular Microbiology* **100**:656–674. DOI: https://doi.org/10.1111/mmi.13338, PMID: 26817626

Nelson-Sathi S, Dagan T, Landan G, Janssen A, Steel M, McInerney JO, Deppenmeier U, Martin WF. 2012. Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of haloarchaea. *PNAS* **109**:20537–20542. DOI: https://doi.org/10.1073/pnas.1209119109, PMID: 23184964

Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* **32**:268–274. DOI: https://doi.org/10.1093/molbev/msu300, PMID: 25371430

Nordberg H, Cantor M, Dusheyko S, Hua S, Poliakov A, Shabalov I, Smirnova T, Grigoriev IV, Dubchak I. 2014. The genome portal of the department of energy joint genome institute: 2014 updates. *Nucleic Acids Research* **42**:D26–D31. DOI: https://doi.org/10.1093/nar/gkt1069, PMID: 24225321

Nozaki H, Takano H, Misumi O, Terasawa K, Matsuzaki M, Maruyama S, Nishida K, Yagisawa F, Yoshida Y, Fujiwara T, Takio S, Tamura K, Chung SJ, Nakamura S, Kuroiwa H, Tanaka K, Sato N, Kuroiwa T. 2007. A 100%-complete sequence reveals unusually simple genomic features in the hot-spring red alga cyanidioschyzon merolae. *BMC Biology* **5**:28. DOI: https://doi.org/10.1186/1741-7007-5-28, PMID: 17623057

Nyyssola A, Kerovuo J, Kaukinen P, von Weymarn N, Reinikainen T. 2000. Extreme halophiles synthesize betaine from glycine by methylation. *Journal of Biological Chemistry* **275**:22196–22201. DOI: https://doi.org/10.1074/jbc.M910111199, PMID: 10896953

O'Brien EA, Koski LB, Zhang Y, Yang L, Wang E, Gray MW, Burger G, Lang BF. 2007. TBestDB: a taxonomically broad database of expressed sequence tags (ESTs). *Nucleic Acids Research* **35**:D445–D451. DOI: https://doi.org/10.1093/nar/gkl770, PMID: 17202165

Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**:299–304. DOI: https://doi.org/10.1038/35012500, PMID: 10830951

Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. 1999. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* **27**:29–34. DOI: https://doi.org/10.1093/nar/27.1.29, PMID: 9847135

Okuyama M, Mori H, Chiba S, Kimura A. 2004. Overexpression and characterization of two unknown proteins, YicI and YihQ, originated from Escherichia coli. *Protein Expression and Purification* **37**:170–179. DOI: https://doi.org/10.1016/j.pep.2004.05.008, PMID: 15294295

Paulose B, Chhikara S, Coomey J, Jung HI, Vatamaniuk O, Dhankher OP. 2013. A γ-glutamyl cyclotransferase protects arabidopsis plants from heavy metal toxicity by recycling glutamate to maintain glutathione homeostasis. *The Plant Cell* **25**:4580–4595. DOI: https://doi.org/10.1105/tpc.113.111815, PMID: 24214398

Phaniendra A, Jestadi DB, Periyasamy L. 2015. Free radicals: properties, sources, targets, and their implication in various diseases. *Indian Journal of Clinical Biochemistry* **30**:11–26. DOI: https://doi.org/10.1007/s12291-014-0446-0, PMID: 25646037

Philippe H, Douady CJ. 2003. Horizontal gene transfer and phylogenetics. *Current Opinion in Microbiology* **6**: 498–505. DOI: https://doi.org/10.1016/j.mib.2003.09.008, PMID: 14572543

Pinto GT. 1975. Roberto, *nuove stazioni italiane di "Cyanidium caldarium". Delpinoa : Nuova Serie Del Bullettino dell'Orto Botanico Della Università Di Napoli* **15**:125–139.

Price DC, Chan CX, Yoon HS, Yang EC, Qiu H, Weber AP, Schwacke R, Gross J, Blouin NA, Lane C, Reyes-Prieto A, Durnford DG, Neilson JA, Lang BF, Burger G, Steiner JM, Löffelhardt W, Meuser JE, Posewitz MC, Ball S, et al. 2012. Cyanophora paradoxa genome elucidates origin of photosynthesis in algae and plants. *Science* **335**:843–847. DOI: https://doi.org/10.1126/science.1213561, PMID: 22344442

Psylinakis E, Boneca IG, Mavromatis K, Deli A, Hayhurst E, Foster SJ, Vårum KM, Bouriotis V. 2005. Peptidoglycan N-acetylglucosamine deacetylases from *Bacillus cereus*, highly conserved proteins in *Bacillus anthracis*. *The Journal of Biological Chemistry* **280**:30856–30863. DOI: https://doi.org/10.1074/jbc.M407426200, PMID: 15961396

Qiu H, Price DC, Weber AP, Reeb V, Yang EC, Lee JM, Kim SY, Yoon HS, Bhattacharya D. 2013. Adaptation through horizontal gene transfer in the cryptoendolithic red alga galdieria phlegrea. *Current Biology* **23**:R865–R866. DOI: https://doi.org/10.1016/j.cub.2013.08.046, PMID: 24112977

Qiu H, Price DC, Yang EC, Yoon HS, Bhattacharya D. 2015. Evidence of ancient genome reduction in red algae (Rhodophyta). *Journal of Phycology* **51**:624–636. DOI: https://doi.org/10.1111/jpy.12294, PMID: 26986787

Qiu H, Yoon HS, Bhattacharya D. 2016. Red algal phylogenomics provides a robust framework for inferring evolution of key metabolic pathways. *PLOS Currents* **8**. DOI: https://doi.org/10.1371/currents.tol.7b037376e6d84a1be34af756a4d90846, PMID: 28018750

Qiu H, Rossoni AW, Weber APM, Yoon HS, Bhattacharya D. 2018. Unexpected conservation of the RNA splicing apparatus in the highly streamlined genome of galdieria sulphuraria. *BMC Evolutionary Biology* **18**:41. DOI: https://doi.org/10.1186/s12862-018-1161-x, PMID: 29606099

Rademacher N, Kern R, Fujiwara T, Mettler-Altmann T, Miyagishima SY, Hagemann M, Eisenhut M, Weber AP. 2016. Photorespiratory glycolate oxidase is essential for the survival of the red alga *cyanidioschyzon merolae* under ambient CO_2 conditions. *Journal of Experimental Botany* **67**:3165–3175. DOI: https://doi.org/10.1093/jxb/erw118, PMID: 26994474

Rao KS, Albro M, Dwyer TM, Frerman FE. 2006. Kinetic mechanism of glutaryl-CoA dehydrogenase. *Biochemistry* **45**:15853–15861. DOI: https://doi.org/10.1021/bi0609016, PMID: 17176108

Raymond JA, Kim HJ. 2012. Possible role of horizontal gene transfer in the colonization of sea ice by algae. *PLOS ONE* **7**:e35968. DOI: https://doi.org/10.1371/journal.pone.0035968, PMID: 22567121

Reeb V, Bhattacharya D. 2010. The thermo-acidophilic cyanidiophyceae (Cyanidiales). In: Seckbach J, Chapman D. J (Eds). *Red Algae in the Genomic Age*. Springer. p. 409–426. DOI: https://doi.org/10.1007/978-90-481-3795-4_22

Reyes-Prieto A, Weber AP, Bhattacharya D. 2007. The origin and establishment of the plastid in algae and plants. *Annual Review of Genetics* **41**:147–168. DOI: https://doi.org/10.1146/annurev.genet.41.110306.130134, PMID: 17600460

Rhoads A, Au KF. 2015. PacBio sequencing and its applications. *Genomics, Proteomics & Bioinformatics* **13**:278–289. DOI: https://doi.org/10.1016/j.gpb.2015.08.002, PMID: 26542840

Richards TA, Monier A. 2016. A tale of two tardigrades. *PNAS* **113**:4892–4894. DOI: https://doi.org/10.1073/pnas.1603862113, PMID: 27084885

Rico-Díaz A, Vizoso Vázquez Ángel, Cerdán ME, Becerra M, Sanz-Aparicio J. 2014. Crystallization and preliminary X-ray diffraction data of β-galactosidase from *Aspergillus niger* . *Acta Crystallographica Section F Structural Biology Communications* **70**:1529–1531. DOI: https://doi.org/10.1107/S2053230X14019815

Robellet X, Flipphi M, Pégot S, Maccabe AP, Vélot C. 2008. AcpA, a member of the GPR1/FUN34/YaaH membrane protein family, is essential for acetate permease activity in the hyphal fungus *aspergillus nidulans*. *Biochemical Journal* **412**:485–493. DOI: https://doi.org/10.1042/BJ20080124, PMID: 18302536

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**:139–140. DOI: https://doi.org/10.1093/bioinformatics/btp616, PMID: 19910308

Rojas AL, Nagem RA, Neustroev KN, Arand M, Adamska M, Eneyskaya EV, Kulminskaya AA, Garratt RC, Golubev AM, Polikarpov I. 2004. *Crystal structures of beta-galactosidase from penicillium* sp. *and its complex with* galactose. *Journal of Molecular Biology* **343**:1281–1292. DOI: https://doi.org/10.1016/j.jmb.2004.09.012, PMID: 15491613

Rossoni AW. 2018. Cold acclimation of the thermoacidophilic red alga *Galdieria sulphuraria* - Changes in gene expression and involvement of horizontally acquired genes. *Plant and Cell Physiology*:pcy240. DOI: https://doi.org/10.1093/pcp/pcy240

Rouhier N, Lemaire SD, Jacquot JP. 2008. The role of glutathione in photosynthetic organisms: emerging functions for glutaredoxins and glutathionylation. *Annual Review of Plant Biology* **59**:143–166. DOI: https://doi.org/10.1146/annurev.arplant.59.032607.092811, PMID: 18444899

Sá-Pessoa J, Paiva S, Ribas D, Silva IJ, Viegas SC, Arraiano CM, Casal M. 2013. SATP (YaaH), a succinate-acetate transporter protein in Escherichia coli. *The Biochemical Journal* **454**:585–595. DOI: https://doi.org/10.1042/BJ20130412, PMID: 23844911

Salzberg SL. 2017. Horizontal gene transfer is not a hallmark of the human genome. *Genome Biology* **18**:85. DOI: https://doi.org/10.1186/s13059-017-1214-2, PMID: 28482857

Schönknecht G, Chen WH, Ternes CM, Barbier GG, Shrestha RP, Stanke M, Bräutigam A, Baker BJ, Banfield JF, Garavito RM, Carr K, Wilkerson C, Rensing SA, Gagneul D, Dickenson NE, Oesterhelt C, Lercher MJ, Weber

AP. 2013. Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote. *Science* **339**:1207–1210. DOI: https://doi.org/10.1126/science.1231707, PMID: 23471408

Schönknecht G, Weber AP, Lercher MJ. 2014. Horizontal gene acquisitions by eukaryotes as drivers of adaptive evolution. *BioEssays* **36**:9–20. DOI: https://doi.org/10.1002/bies.201300095, PMID: 24323918

Schrimsher JL, Schendel FJ, Stubbe J, Smith JM. 1986. Purification and characterization of aminoimidazole ribonucleotide synthetase from Escherichia coli. *Biochemistry* **25**:4366–4371. DOI: https://doi.org/10.1021/bi00363a028, PMID: 3530323

Seckbach J. 1972. On the fine structure of the acidophilic hot-spring alga Cyanidium caldarium: a taxonomic approach. *Microbios* **5**:133–142. PMID: 4206412

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**:3210–3212. DOI: https://doi.org/10.1093/bioinformatics/btv351, PMID: 26059717

Simonetti A, Marzi S, Jenner L, Myasnikov A, Romby P, Yusupova G, Klaholz BP, Yusupov M. 2009. A structural view of translation initiation in bacteria. *Cellular and Molecular Life Sciences* **66**:423–436. DOI: https://doi.org/10.1007/s00018-008-8416-4, PMID: 19011758

Stadtman ER, Levine RL. 2000. Protein oxidation. *Annals of the New York Academy of Sciences* **899**:191–208. DOI: https://doi.org/10.1111/j.1749-6632.2000.tb06187.x, PMID: 10863540

Stanke M, Morgenstern B. 2005. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Research* **33**:W465–W467. DOI: https://doi.org/10.1093/nar/gki458, PMID: 15980513

Stauffer RE, Thompson JM. 1984. Arsenic and antimony in geothermal waters of Yellowstone National Park, Wyoming, USA. *Geochimica Et Cosmochimica Acta* **48**:2547–2561. DOI: https://doi.org/10.1016/0016-7037(84)90305-3

Takeda K, Akimoto C, Kawamukai M. 2001. Effects of the Escherichia coli sfsA gene on mal genes expression and a DNA binding activity of SfsA. *Bioscience, Biotechnology, and Biochemistry* **65**:213–217. DOI: https://doi.org/10.1271/bbb.65.213, PMID: 11272834

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, Deboy RT, Davidsen TM, Mora M, Scarselli M, Margarit y Ros I, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, et al. 2005. Genome analysis of multiple pathogenic isolates of streptococcus agalactiae: implications for the microbial "pan-genome". *PNAS* **102**:13950–13955. DOI: https://doi.org/10.1073/pnas.0506758102, PMID: 16172379

Toplin JA, Norris TB, Lehr CR, McDermott TR, Castenholz RW. 2008. Biogeographic and phylogenetic diversity of thermoacidophilic cyanidiales in Yellowstone National Park, Japan, and New Zealand. *Applied and Environmental Microbiology* **74**:2822–2833. DOI: https://doi.org/10.1128/AEM.02741-07, PMID: 18344337

UniProt Consortium T . 2018. UniProt: the universal protein knowledgebase. *Nucleic Acids Research* **46**:2699. DOI: https://doi.org/10.1093/nar/gky092, PMID: 29425356

van Doesburg W, van Eekert MH, Middeldorp PJ, Balk M, Schraa G, Stams AJ. 2005. Reductive dechlorination of beta-hexachlorocyclohexane (beta-HCH) by a dehalobacter species in Coculture with a sedimentibacter sp. *FEMS Microbiology Ecology* **54**:87–95. DOI: https://doi.org/10.1016/j.femsec.2005.03.003, PMID: 16329975

van Lieshout JFT, Verhees CH, Ettema TJG, van der Sar S, Imamura H, Matsuzawa H, van der Oost J, de Vos WM. 2003. Identification and Molecular Characterization of a Novel Type of α-galactosidase from *Pyrococcus furiosus* . *Biocatalysis and Biotransformation* **21**:243–252. DOI: https://doi.org/10.1080/10242420310001614342

Vernikos G, Medini D, Riley DR, Tettelin H. 2015. Ten years of pan-genome analyses. *Current Opinion in Microbiology* **23**:148–154. DOI: https://doi.org/10.1016/j.mib.2014.11.016, PMID: 25483351

Waditee R, Tanaka Y, Aoki K, Hibino T, Jikuya H, Takano J, Takabe T, Takabe T. 2003. Isolation and functional characterization of N-methyltransferases that catalyze betaine synthesis from glycine in a halotolerant photosynthetic organism aphanothece halophytica. *Journal of Biological Chemistry* **278**:4932–4942. DOI: https://doi.org/10.1074/jbc.M210970200, PMID: 12466265

Weber AP. 2007. A genomics approach to understanding the biology of thermo-acidophilic red algae. In: Seckbach J (Ed). *Algae and Cyanobacteria in Extreme Environments*. Springer. p. 503–518. DOI: https://doi.org/10.1007/978-1-4020-6112-7_27

Yang EC, Boo SM, Bhattacharya D, Saunders GW, Knoll AH, Fredericq S, Graf L, Yoon HS. 2016. Divergence time estimates and the evolution of major lineages in the florideophyte red algae. *Scientific Reports* **6**:21361. DOI: https://doi.org/10.1038/srep21361, PMID: 26892537

Ylä-Herttuala S. 1999. Oxidized LDL and Atherogenesisa.In: *Annals of the New York Academy of Sciences* **874**: 134–137. DOI: https://doi.org/10.1111/j.1749-6632.1999.tb09231.x, PMID: 10415527

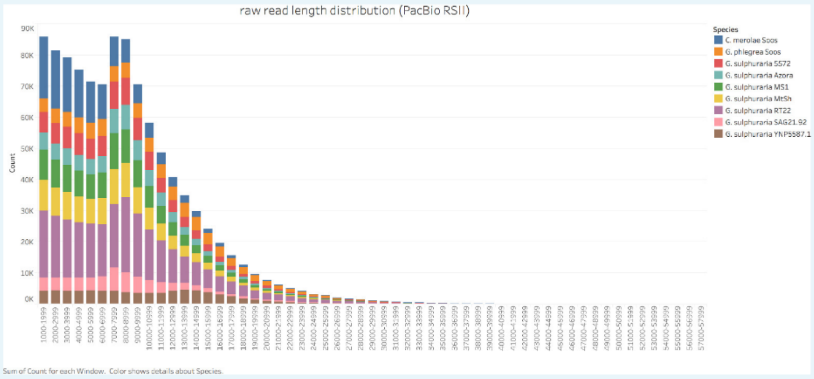
Yoon HS, Hackett JD, Ciniglia C, Pinto G, Bhattacharya D. 2004. A molecular timeline for the origin of photosynthetic eukaryotes. *Molecular Biology and Evolution* **21**:809–818. DOI: https://doi.org/10.1093/molbev/msh075, PMID: 14963099

Yu L, Petros AM, Schnuchel A, Zhong P, Severin JM, Walter K, Holzman TF, Fesik SW. 1997. Solution structure of an rRNA methyltransferase (ErmAM) that confers macrolide-lincosamide-streptogramin antibiotic resistance. *Nature Structural Biology* **4**:483–489. DOI: https://doi.org/10.1038/nsb0697-483, PMID: 9187657

Zhao FJ, McGrath SP, Meharg AA. 2010. Arsenic as a food chain contaminant: mechanisms of plant uptake and metabolism and mitigation strategies. *Annual Review of Plant Biology* **61**:535–559. DOI: https://doi.org/10.1146/annurev-arplant-042809-112152, PMID: 20192735

Appendix 1

DOI: https://doi.org/10.7554/eLife.45017.108


raw read length distribution (PacBio RSII)

Sum of Count for each Window. Color shows details about Species.

Appendix 1—figure 1. Raw read length distribution of the sequenced Cyanidiales strains. The strains were sequenced in 2016/2017 using PacBio's RS2 sequencing technology and P6-C4 chemistry (the only exception being *C. merolae Soos*, which was sequenced as pilot study using P4-C2 chemistry in 2014). Seven strains, namely *G. sulphuraria* 5572, *G. sulphuraria* 002, *G. sulphuraria* SAG21.92, *G. sulphuraria* Azora, *G. sulphuraria* MtSh, *G. sulphuraria* RT22 and *G. sulphuraria* MS1 were sequenced at the University of Maryland Institute for Genome Sciences (Baltimore, USA). The remaining three strains, *G. sulphuraria YNP5578.1*, *G. phlegrea* Soos and *C. merolae* Soos, were sequenced at the Max-Planck-Institut für Pflanzenzüchtungsforschung (Cologne, Germany).

DOI: https://doi.org/10.7554/eLife.45017.109

Appendix 1—table 1. Sequencing and Assembly stats.

The strains were sequenced using PacBio's RS2 sequencing technology and P6-C4 chemistry (the only exception being *C. merolae Soos*, which was sequenced using P4-C2 chemistry). For genome assembly, canu version 1.5 was used, followed by polishing three times using the Quiver algorithm. Genes were predicted with MAKER v3 beta(***Doolittle, 1999***; ***Doolittle, 1999***). The performance of genome assemblies (not shown here) and gene prediction was assessed using BUSCO v.3. Raw Reads: Number of raw PacBio RSII reads. Raw Reads N50: 50% of the raw sequence is contained in reads with sizes greater than the N50 value. Raw Reads GC: GC content of the raw reads in percent. Raw Reads (bp): Total number of sequenced basepairs (nucleotides) per species. Raw Coverage (bp): Genomic coverage by raw reads. This figure was computed once the assembly was finished. Unitigging (bp): Total number of basepairs that survived read correction and trimming. This amount of sequence is what the assembler considered when constructing the genome. Unitigging Coverage: Genomic coverage by corrected and trimmed reads. Genome Size (bp): Size of the polished genome. Genome GC: GC content of the polished genome. Contigs: Number of contigs. Contig N50: 50% of the final genomic sequence is contained in contigs sizes greater than the N50 value. Genes: Number of genes predicted by Maker v3 beta. BUSCO (C): Percentage of complete gene models. BUSCO (C + F): Percentage of complete and fragmented gene models. Fragmented gene models are also somewhat present. BUSCO (D): Percentage of duplicated gene models. BUSCO (M): Percentage of missing gene models.

Species	Raw reads	Raw reads N50	Raw reads GC	Raw reads (bp)	Raw reads coverage	Unitigging (bp)	Unitigging coverage	Genome size (bp)	Genome GC	Contigs	Contig N50	Genes	Busco (C)	Busco (C + F)	Busco (D)	Busco (M)
G. sulphuraria RT22	163764	12023	35.83%	1424372481	91.20	1108677098	70.99	15617852	37.43%	118	172878	6982	92.8%	94.5%	6.3%	5.5%
G. sulphuraria 002	131978	10109	37.90%	946093501	67.05	805608410	57.09	14110219	39.16%	107	189293	5912	87.5%	92.5%	5.0%	7.5%
G. sulphuraria 5572	101472	10449	36.45%	802203307	56.19	664626554	46.55	14277368	37.99%	108	229711	6472	91.5%	93.5%	5.0%	6.5%
G. sulphuraria MS1	128294	9991	36.18%	934546621	62.77	777587876	52.23	14887946	37.62%	129	172087	7441	90.8%	94.1%	4.0%	5.9%
G. sulphuraria MtSh	158936	13617	39.19%	1523875693	101.95	1235394614	82.65	14947614	40.04%	101	186619	6160	87.4%	91.7%	6.9%	8.3%
G. sulphuraria Azora	82544	10244	37.09%	651280930	46.31	551720524	39.23	14063793	40.10%	127	162248	6305	88.4%	92.0%	2.3%	8.0%

Appendix 1—table 1 continued on next page

Appendix 1—table 1 continued

Species	Raw reads	Raw reads N50	Raw reads GC	Raw reads (bp)	Raw reads coverage	Unitigging (bp)	Unitigging coverage	Genome size (bp)	Genome GC	Contigs	Contig N50	Genes	Busco (C)	Busco (C + F)	Busco (D)	Busco (M)
G. sulphuraria SAG21.92	71480	10341	36.67%	564874149	39.47	413793659	28.91	14312824	37.92%	135	158217	5956	83.8%	88.4%	3.6%	11.6%
G. sulphuraria YNP5587.1	77421	13842	36.69%	769606723	53.38	613905250	42.58	14416547	40.05%	115	170797	6118	91.8%	93.5%	5.0%	6.5%
G. phlegrea Soos	92263	14365	36.01%	966702049	65.00	619580741	41.66	14872696	37.52%	108	201071	6125	92.1%	93.8%	7.9%	6.2%
C. merolae Soos	154461	7924	52.92%	**848542698**	68.82	570542830	46.27	12329961	54.33%	35	567466	4406	85.2%	89.5%	2.0%	10.5%
G. sulphuraria 074W*								13712004	36.89%	433	172322	7177	83.8%	87.4%	2.3%	10.3%
C. merolae 10D*								16728945	54.81%	22	859119	5044	90.4%	93.4%	1.3%	6.6%
G. phlegrea DBV009*								11413183	37.86%	9311	1993	7836	68.3%	88.1%	3.6%	11.9%

DOI: https://doi.org/10.7554/eLife.45017.110

Appendix 2

DOI: https://doi.org/10.7554/eLife.45017.108

Archaeal ATPases and 'old' HGT

We compared the HGT results of this study to previous published claims of HGT in *G. sulphuraria* 074W (75 separate acquisitions followed by gene family expansion, 335 transcripts in total) (**Schönknecht et al., 2013**) and *G. phlegrea* DBV009 (13 genes from 11 acquisitions unique to this strain, excluding those shared with *G. sulphuraria* 074W and other red algae) (**Qiu et al., 2013**). Each HGT candidate was queried against our database, mapped to the existing OGs and phylogenetic trees were built for each sequence (where possible). The HGT candidates of *G. sulphuraria* 074W mapped into 100 different OGs, thus increasing the number of separate origins from 75 to 100 (more separate origins = less gene family expansion). 211 out of the 335 HGT candidates in *G. sulphuraria* 074W are 'archaeal STAND ATPases'. They clustered into OG0000000, OG0000003 and OG0000001 which are not classified as HGT. Thus, HGT origin for those gene families can be excluded. The remaining 124 *G. sulphuraria* 074W HGT candidates are spread across 98 OGs. Of those, 20 OGs overlap with our HGT findings, whereas 78 are OGs that do not have HGT origins (one was classified as EGT). All 13 HGT candidates in *G. phlegrea* DBV009 were found and their HGT origin could be confirmed. Some do not make the cut due to individual acquisitions by *G. phlegrea* DBV009 alone. However, considering the operon structures of the acquisition it seems plausible in this case.

In order to exclude the possibility that our database was 'missing' crucial non-eukaryotic species we queried all protein sequences against our own database and NCBI's uncurated nr database, including predicted models and environmental samples and implementing various search strategies. 219 out of the 335 HGT candidates in *G. sulphuraria* 074W did not report any hits outside the species itself (including the 211 'archaeal ATPases') and no functional evidence could be found besides the one obtained through manual curation of sequence alignments as reported by the author (**Schönknecht et al., 2013**).

As seen in the case of the human and the Tardigrade genome, the overestimation of HGT in eukaryotic genomes, followed by later re-correction, is not a new phenomenon (**Boothby et al., 2015**; **Crisp et al., 2015**; **Koutsovoulos et al., 2016**; **Salzberg, 2017**). There are several reasons that may have led to the drastic overestimation of HGT candidates in the case of *G. sulphuraria* 074W (100 OGs derived from HGT, instead of 58 OGs). Although published in 2013, the HGT analysis was performed in early 2007. By then, the RefSeq database contained 4.7 million accessions compared to 163.9 million accessions in May 2018. The low resolution regarding eukaryotic species may have led to many singletons, here defined as *Galdieria* being the only eukaryotic species in otherwise bacterial clusters, leading to the mislabelling of HGT. Further, the many small contigs derived from short read sequencing technologies of the last decade, combined with older assembly software [138] are known potential pitfalls (**Danchin, 2016**) for missassembly that may lead to the inclusion of bacterial contigs into the reference genome as a consequence of prior culture contamination. Lastly, this analysis occurred a decade prior to the tardigrade and human case that led to raised awareness and standards regarding HGT annotation as many claims of HGT were later refuted by further analyses. From a biological view the HGT origin of the Archaeal ATPases is disputable as a re-sequencing of the Genome using MinION technology (Rossoni, data unpublished) shows they always occur immediately adjacent to every single telomere, therefore adding another layer of complexity. The 'archaeal ATPase' was not only integrated into the genome, but also put under influence a non-random duplication mechanism responsible for spreading copies in a targeted manner to the subtelomeric region of each single contig (no exception!). Examples of similar cases may be found in the Variant Surface Glycoproteins (VSGs) of the Trypanosoma [139] and the Candidates for Secreted Effector Proteins (CSEPs) in the powdery mildew fungus Blumeria graminis [140]. As those genes are vital for the infection of the host, they are subjects of very strong natural selection and profit from high evolutionary rates achieved at the subtelomeric regions. But the high evolutionary rates also made it impossible to correctly embed the aforementioned gene families in a phylogenetic tree. As such, it is not to be excluded that a similar case occurred regarding *Galdieria sulphuraria*'s 'archaeal ATPases', although a permissive search might indicate an archaeal origin of single protein domains. Also, as only a patchy subset of the ATPases reacts to temperature fluctuations, it cannot be determined that temperature is the driving factor.

Appendix 3

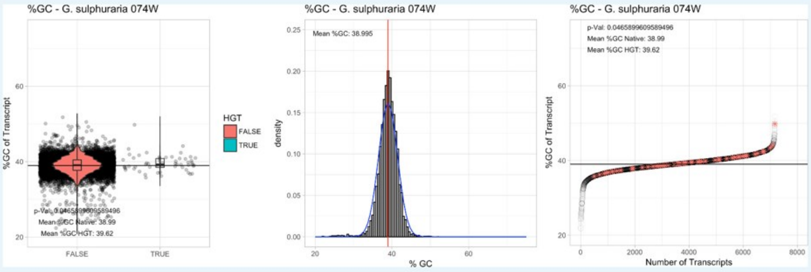
DOI: https://doi.org/10.7554/eLife.45017.108

%GC

Appendix 3—table 1. %GC analysis of the Cyanidiales transcriptomes. %GC content of HGT genes was compared to the %GC content of native genes using students test. Legend: HGT Genes: number of HGT gene candidates found in species. Avg. %GC Native: average %GC of native transcripts. Avg. %GC HGT: average %GC of HGT candidates. P-Val (T-test): significance value (p-value) of student's test. Delta: difference in %GC between average %GC of native genes and the average %GC of HGT candidates.

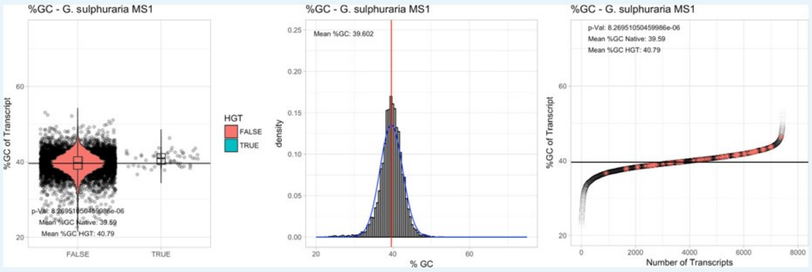
	HGT genes	Avg. %GC Native	Avg. %GC HGT	p-Val (T-test)	Delta
Galdieria_ sulphuraria_074W	55	38.99	39.62	0.046	0.63
Galdieria_ sulphuraria_MS1	58	39.59	40.79	0	1.2
Galdieria_ sulphuraria_RT22	54	39.54	40.85	0	1.31
Galdieria_ sulphuraria_SAG21	47	40.04	41.47	0	1.43
Galdieria_ sulphuraria_MtSh	47	41.33	42.48	0	1.15
Galdieria_ sulphuraria_Azora	58	41.34	42.57	0	1.23
Galdieria_ sulphuraria _YNP55871	46	41.33	42.14	0.006	0.81
Galdieria_ sulphuraria_5572	53	39.68	40.5	0.002	0.82
Galdieria_ sulphuraria_002	52	40.76	41.35	0.016	0.59
Galdieria_ phlegrea_DBV08	54	39.97	40.58	0.016	0.61
Galdieria_ phlegrea_Soos	44	39.57	40.73	0	1.16
Cyanidioschyzon_ merolae_10D	33	56.57	56.57	0.996	0
Cyanidioschyzon_ merolae_Soos	34	54.84	54.26	0.479	−0.58

DOI: https://doi.org/10.7554/eLife.45017.113



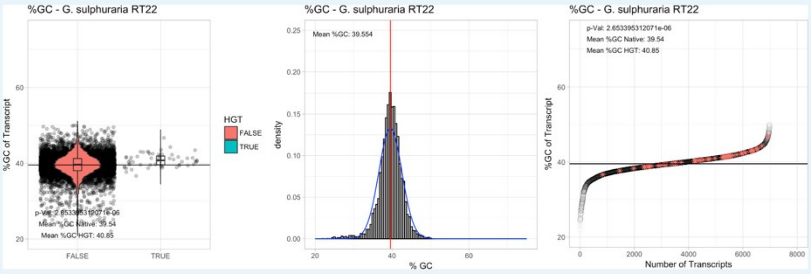
Appendix 3—figure 1. %GC – *Galdieria sulphuraria 074W:* (Left) Violin plot showing the %GC distribution across native transcripts and HGT candidates. (Mid) Cumulative %GC distribution of transcripts. Red line shows the average, blue line a normal distribution based on the average value. (Right) Ranking all transcripts based upon their %GC content. Red '*' demarks HGT candidates. As the %GC content was normally distributed, students test was applied for the determination of significant differences between the native gene and the HGT candidate subset.

DOI: https://doi.org/10.7554/eLife.45017.114



Appendix 3—figure 2. %GC – *Galdieria sulphuraria MS1:* (Left) Violin plot showing the %GC distribution across native transcripts and HGT candidates. (Mid) Cumulative %GC distribution of transcripts. Red line shows the average, blue line a normal distribution based on the average value. (Right) Ranking all transcripts based upon their %GC content. Red '*' demarks HGT candidates. As the %GC content was normally distributed, students test was applied for the determination of significant differences between the native gene and the HGT candidate subset.

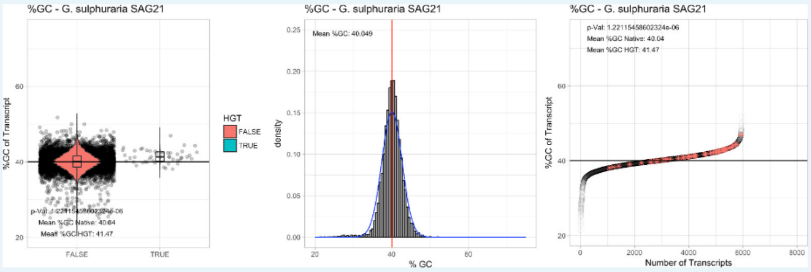
DOI: https://doi.org/10.7554/eLife.45017.115



Appendix 3—figure 3. %GC – *Galdieria sulphuraria RT22:* (Left) Violin plot showing the %GC distribution across native transcripts and HGT candidates. (Mid) Cumulative %GC distribution of transcripts. Red line shows the average, blue line a normal distribution based on the

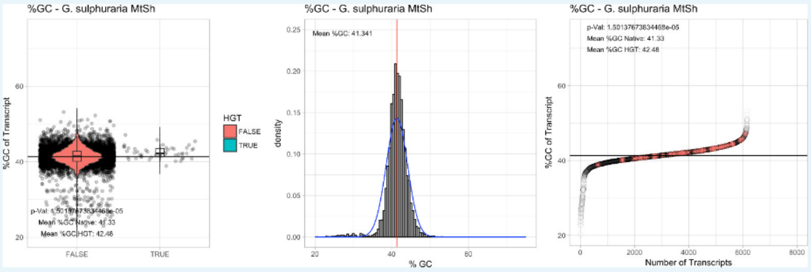
average value. (Right) Ranking all transcripts based upon their %GC content. Red '*' demarks HGT candidates. As the %GC content was normally distributed, students test was applied for the determination of significant differences between the native gene and the HGT candidate subset.

DOI: https://doi.org/10.7554/eLife.45017.116



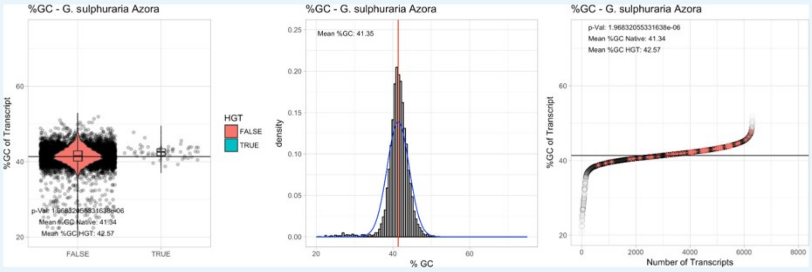
Appendix 3—figure 4. %GC – *Galdieria sulphuraria SAG21:* (Left) Violin plot showing the % GC distribution across native transcripts and HGT candidates. (Mid) Cumulative %GC distribution of transcripts. Red line shows the average, blue line a normal distribution based on the average value. (Right) Ranking all transcripts based upon their %GC content. Red '*' demarks HGT candidates. As the %GC content was normally distributed, students test was applied for the determination of significant differences between the native gene and the HGT candidate subset.

DOI: https://doi.org/10.7554/eLife.45017.117



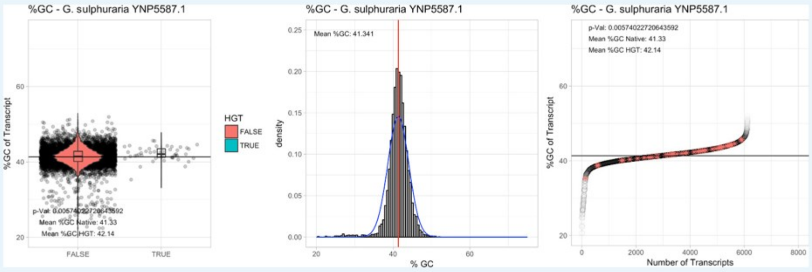
Appendix 3—figure 5. %GC – *Galdieria sulphuraria Mount Shasta (MtSh):* (Left) Violin plot showing the %GC distribution across native transcripts and HGT candidates. (Mid) Cumulative %GC distribution of transcripts. Red line shows the average, blue line a normal distribution based on the average value. (Right) Ranking all transcripts based upon their %GC content. Red '*' demarks HGT candidates. As the %GC content was normally distributed, students test was applied for the determination of significant differences between the native gene and the HGT candidate subset.

DOI: https://doi.org/10.7554/eLife.45017.118



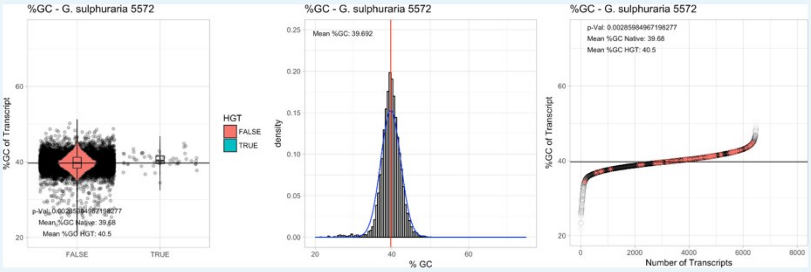
Appendix 3—figure 6. *Galdieria sulphuraria Azora:* (Left) Violin plot showing the %GC distribution across native transcripts and HGT candidates. (Mid) Cumulative %GC distribution of transcripts. Red line shows the average, blue line a normal distribution based on the average value. (Right) Ranking all transcripts based upon their %GC content. Red '*' demarks HGT candidates. As the %GC content was normally distributed, students test was applied for the determination of significant differences between the native gene and the HGT candidate subset.

DOI: https://doi.org/10.7554/eLife.45017.119



Appendix 3—figure 7. %GC – *Galdieria sulphuraria Mount Shasta YNP5578.1:* (Left) Violin plot showing the %GC distribution across native transcripts and HGT candidates. (Mid) Cumulative %GC distribution of transcripts. Red line shows the average, blue line a normal distribution based on the average value. (Right) Ranking all transcripts based upon their % GC content. Red '*' demarks HGT candidates. As the %GC content was normally distributed, students test was applied for the determination of significant differences between the native gene and the HGT candidate subset.

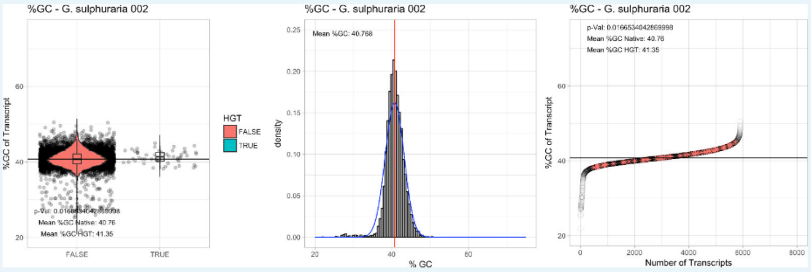
DOI: https://doi.org/10.7554/eLife.45017.120



Appendix 3—figure 8. %GC – *Galdieria sulphuraria 5572:* (Left) Violin plot showing the %GC distribution across native transcripts and HGT candidates. (Mid) Cumulative %GC distribution of transcripts. Red line shows the average, blue line a normal distribution based on the

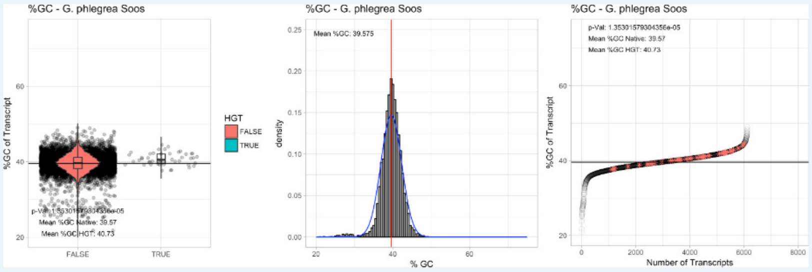
average value. (Right) Ranking all transcripts based upon their %GC content. Red '*' demarks HGT candidates. As the %GC content was normally distributed, students test was applied for the determination of significant differences between the native gene and the HGT candidate subset.

DOI: https://doi.org/10.7554/eLife.45017.121



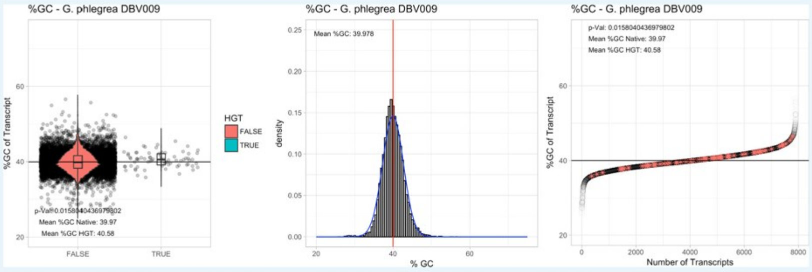
Appendix 3—figure 9. %GC – *Galdieria sulphuraria 002:* (Left) Violin plot showing the %GC distribution across native transcripts and HGT candidates. (Mid) Cumulative %GC distribution of transcripts. Red line shows the average, blue line a normal distribution based on the average value. (Right) Ranking all transcripts based upon their %GC content. Red '*' demarks HGT candidates. As the %GC content was normally distributed, students test was applied for the determination of significant differences between the native gene and the HGT candidate subset.

DOI: https://doi.org/10.7554/eLife.45017.122



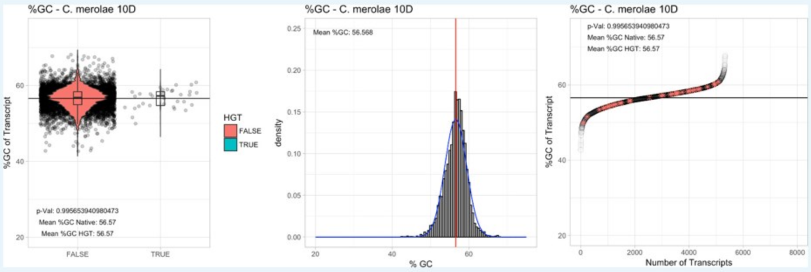
Appendix 3—figure 10. %GC – *Galdieria phlegrea Soos:* (Left) Violin plot showing the %GC distribution across native transcripts and HGT candidates. (Mid) Cumulative %GC distribution of transcripts. Red line shows the average, blue line a normal distribution based on the average value. (Right) Ranking all transcripts based upon their %GC content. Red '*' demarks HGT candidates. As the %GC content was normally distributed, students test was applied for the determination of significant differences between the native gene and the HGT candidate subset.

DOI: https://doi.org/10.7554/eLife.45017.123



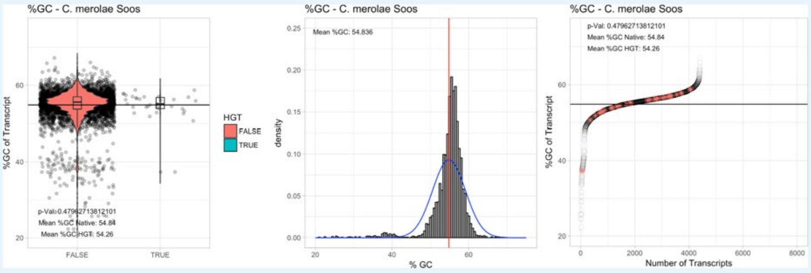
Appendix 3—figure 11. %GC – *Galdieria phlegrea DBV009:* (Left) Violin plot showing the % GC distribution across native transcripts and HGT candidates. (Mid) Cumulative %GC distribution of transcripts. Red line shows the average, blue line a normal distribution based on the average value. (Right) Ranking all transcripts based upon their %GC content. Red '*' demarks HGT candidates. As the %GC content was normally distributed, students test was applied for the determination of significant differences between the native gene and the HGT candidate subset.

DOI: https://doi.org/10.7554/eLife.45017.124



Appendix 3—figure 12. %GC – *Cyanidioschyzon merolae Soos:* (Left) Violin plot showing the %GC distribution across native transcripts and HGT candidates. (Mid) Cumulative %GC distribution of transcripts. Red line shows the average, blue line a normal distribution based on the average value. (Right) Ranking all transcripts based upon their %GC content. Red '*' demarks HGT candidates. As the %GC content was normally distributed, students test was applied for the determination of significant differences between the native gene and the HGT candidate subset.

DOI: https://doi.org/10.7554/eLife.45017.125



Appendix 3—figure 13. %GC – *Cyanidioschyzon merolae 10D:* (Left) Violin plot showing the %GC distribution across native transcripts and HGT candidates. (Mid) Cumulative %GC distribution of transcripts. Red line shows the average, blue line a normal distribution based

on the average value. (Right) Ranking all transcripts based upon their %GC content. Red '*' demarks HGT candidates. As the %GC content was normally distributed, students test was applied for the determination of significant differences between the native gene and the HGT candidate subset.

DOI: https://doi.org/10.7554/eLife.45017.126

Appendix 4

DOI: https://doi.org/10.7554/eLife.45017.108

Appendix 4—table 1. Single exon genes vs multiexonic. The ratio of single exon genes vs multiexonic genes was compared between HGT candidates and native Cyanidiales genes (Fisher enrichment test). Legend: HGT Genes: number of HGT gene candidates found in species. Single Exon HGT: number of single exon genes in HGT candidates. Multi Exon HGT: number of multiexonic genes in HGT candidates. Single Exon Native: number of single exon genes in native Cyanidiales genes. Multi Exon Native: number of multiexonic genes in native Cyanidiales genes. HGT SM Ratio percentage of single exon genes within the HGT candidate genes. Native SM Ratio percentage of single exon genes within the native genes. Delta: difference in percent between the percentage of single exon genes between the native genes and HGT candidates. Fisher p-val: p-value of fisher enrichment test.

	HGT genes	Single exon (HGT)	Multi exon (HGT)	Single exon (Native)	Multi exon (Native)	Fisher's p	Single exon % (HGT)	Single exon % (Native)	Multi exon % (HGT)	Multi exon % (Native)
Galdieria_ sulphuraria_ 074W	55	29	26	1879	5240	4.05E-05	52.7%	26.4%	47.3%	73.6%
Galdieria_ sulphuraria_ MS1	58	22	36	1224	6159	0.0001098	37.9%	16.6%	62.1%	83.4%
Galdieria_ sulphuraria_ RT22	54	26	28	1756	5172	0.0004079	48.1%	25.3%	51.9%	74.7%
Galdieria_ sulphuraria_ SAG21	47	8	39	901	5008	0.6852	17.0%	15.2%	83.0%	84.8%
Galdieria_ sulphuraria_ MtSh	47	17	30	1239	4874	0.01054	36.2%	20.3%	63.8%	79.7%
Galdieria_ sulphuraria_ Azora	58	14	39	966	5286	0.03558	24.1%	15.5%	75.9%	84.5%
Galdieria_ sulphuraria_ YNP55871	46	21	25	1548	4524	0.00341	45.7%	25.5%	54.3%	74.5%
Galdieria_ sulphuraria_ 5572	53	29	24	1389	5030	1.75E-07	54.7%	21.6%	45.3%	78.4%
Galdieria_ sulphuraria_ 002	52	26	26	140	4720	8.75E-07	50.0%	2.9%	50.0%	97.1%
Galdieria_ phlegrea_ DBV009	54	na	na	na	na	na	na	na	na	na
Galdieria_ phlegrea_ Soos	44	25	22	1369	4709	5.17E-06	56.8%	22.5%	43.2%	77.5%
Cyanidio schyzon_ merolae_ 10D	33	33	0	4744	26	1	100.0%	99.5%	0.0%	0.5%

Appendix 4—table 1 continued on next page

Appendix 4—table 1 continued

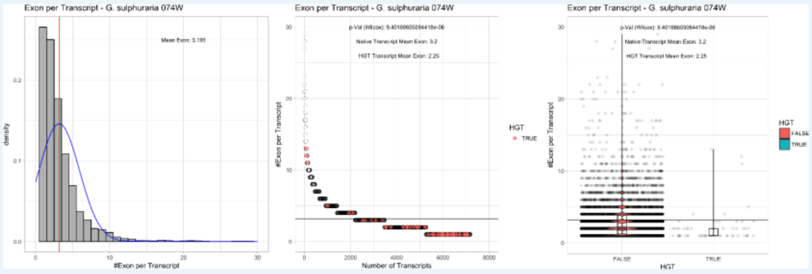
	HGT genes	Single exon (HGT)	Multi exon (HGT)	Single exon (Native)	Multi exon (Native)	Fisher's p	Single exon % (HGT)	Single exon % (Native)	Multi exon % (HGT)	Multi exon % (Native)
Cyanidioschyzon_merolae_Soos	34	33	1	3960	412	0.367	97.1%	90.6%	2.9%	9.4%

DOI: https://doi.org/10.7554/eLife.45017.128

Appendix 4—table 2. Exon/Gene ratio. The ratio of exons per gene was compared between HGT candidates and native Cyanidiales genes (Wilcox ranked test). Legend: HGT Genes: number of HGT gene candidates found in species. E/G All: average number of exons per gene across the whole transcriptome. E/G Native: average number of exons per gene across in native genes. E/G HGT: average number of exons per gene in HGT gene candidates. p-Val (Wilcox) SM Ratio p-value of non-parametric Wilcox test for significant differences. Delta: difference in average number of exons per gene the native genes and HGT candidates.

	HGT genes	Mean exon per transcript (HGT)	Mean exon per transcript (Native)	Wilcox (p)	Delta
Galdieria_sulphuraria_074W	55	2.25	3.2	9.40E-06	0.95
Galdieria_sulphuraria_MS1	58	2.5	3.88	1.41E-05	1.38
Galdieria_sulphuraria_RT22	54	2.63	3.95	3.42E-06	1.32
Galdieria_sulphuraria_SAG21	47	4.02	5.03	0.0004	1.01
Galdieria_sulphuraria_MtSh	47	3.15	4.32	0.0011	1.17
Galdieria_sulphuraria_Azora	58	2.68	4.03	9.92E-05	1.35
Galdieria_sulphuraria_YNP55871	46	2.61	3.65	2.30E-04	1.04
Galdieria_sulphuraria_5572	53	2.15	3.53	2.25E-07	1.38
Galdieria_sulphuraria_002	52	2.37	3.73	2.65E-06	1.36
Galdieria_phlegrea_DBV009	54	na	na	na	na
Galdieria_phlegrea_Soos	44	2.19	3.33	1.19E-05	1.14
Cyanidioschyzon_merolae_10D	33	1	1.01	1.00E + 00	0.01
Cyanidioschyzon_merolae_Soos	34	1.06	1.1	2.10E-01	0.04

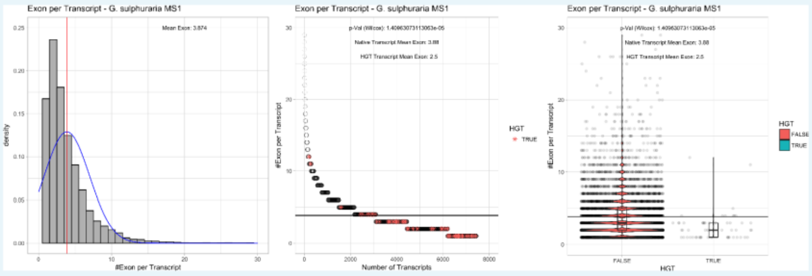
DOI: https://doi.org/10.7554/eLife.45017.129



Appendix 4—figure 1. Exon/Intron – *Galdieria sulphuraria 074W:* (Left) Mid) Cumulative %GC distribution of transcripts. Red line shows the average, blue line a normal distribution based on the average value. The data is categorical (genes have either one, two, three etc. exons) and does not follow a normal distribution. (Mid) Ranking all transcripts based upon their

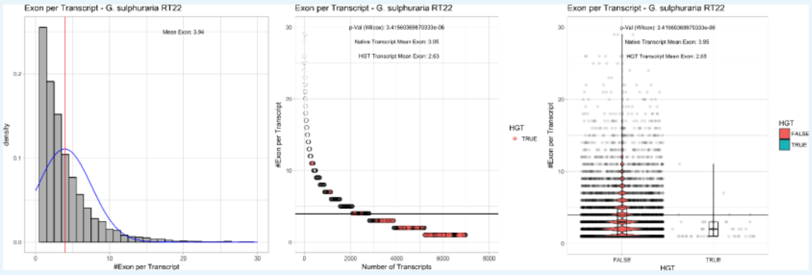
number of exons. Red '*' demarks HGT candidates. As the number of exons was not normally distributed, transcripts were ranked by number of exons. In order to resolve the high number of tied ranks (e.g. many transcripts have two exons) a bootstrap was implied by which the rank of transcripts sharing the same number of exons was randomly assigned 1000 times. Wilcoxon-Mann-Whitney-Test applied for the determination of significant rank differences between the native gene and the HGT candidate subset. (Right) Violin plot showing the number of exons per transcript distribution across native transcripts and HGT candidates.

DOI: https://doi.org/10.7554/eLife.45017.130



Appendix 4—figure 2. Exon/Intron – *Galdieria sulphuraria MS1:* (Left) Mid) Cumulative %GC distribution of transcripts. Red line shows the average, blue line a normal distribution based on the average value. The data is categorical (genes have either one, two, three etc. exons) and does not follow a normal distribution. (Mid) Ranking all transcripts based upon their number of exons. Red '*' demarks HGT candidates. As the number of exons was not normally distributed, transcripts were ranked by number of exons. In order to resolve the high number of tied ranks (e.g. many transcripts have two exons) a bootstrap was implied by which the rank of transcripts sharing the same number of exons was randomly assigned 1000 times. Wilcoxon-Mann-Whitney-Test applied for the determination of significant rank differences between the native gene and the HGT candidate subset. (Right) Violin plot showing the number of exons per transcript distribution across native transcripts and HGT candidates.

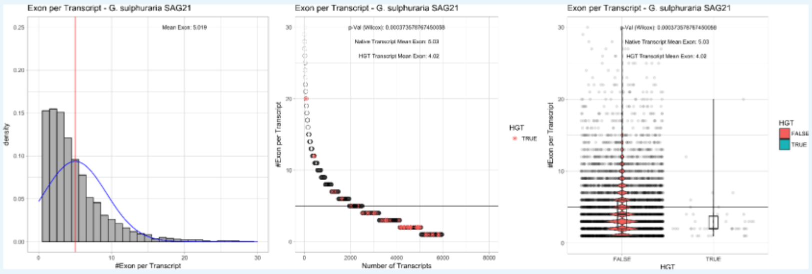
DOI: https://doi.org/10.7554/eLife.45017.131



Appendix 4—figure 3. Exon/Intron – *Galdieria sulphuraria RT22:* (Left) Mid) Cumulative %GC distribution of transcripts. Red line shows the average, blue line a normal distribution based on the average value. The data is categorical (genes have either one, two, three etc. exons) and does not follow a normal distribution. (Mid) Ranking all transcripts based upon their number of exons. Red '*' demarks HGT candidates. As the number of exons was not normally distributed, transcripts were ranked by number of exons. In order to resolve the high number of tied ranks (e.g. many transcripts have two exons) a bootstrap was implied by which the rank of transcripts sharing the same number of exons was randomly assigned 1000

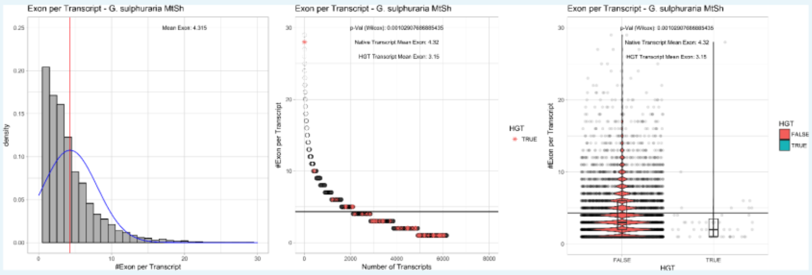
times. Wilcoxon-Mann-Whitney-Test applied for the determination of significant rank differences between the native gene and the HGT candidate subset. (Right) Violin plot showing the number of exons per transcript distribution across native transcripts and HGT candidates.

DOI: https://doi.org/10.7554/eLife.45017.132



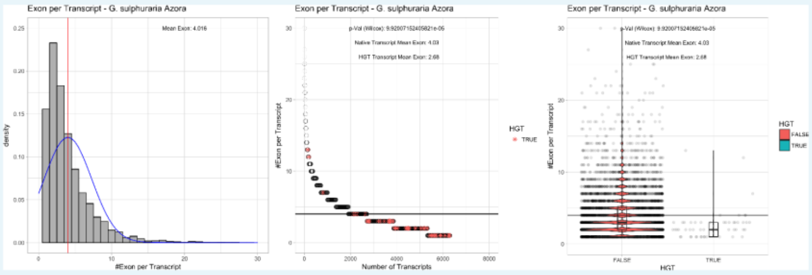
Appendix 4—figure 4. Exon/Intron – *Galdieria sulphuraria SAG21:* (Left) Mid) Cumulative % GC distribution of transcripts. Red line shows the average, blue line a normal distribution based on the average value. The data is categorical (genes have either one, two, three etc. exons) and does not follow a normal distribution. (Mid) Ranking all transcripts based upon their number of exons. Red '*' demarks HGT candidates. As the number of exons was not normally distributed, transcripts were ranked by number of exons. In order to resolve the high number of tied ranks (e.g. many transcripts have two exons) a bootstrap was implied by which the rank of transcripts sharing the same number of exons was randomly assigned 1000 times. Wilcoxon-Mann-Whitney-Test applied for the determination of significant rank differences between the native gene and the HGT candidate subset. (Right) Violin plot showing the number of exons per transcript distribution across native transcripts and HGT candidates.

DOI: https://doi.org/10.7554/eLife.45017.133



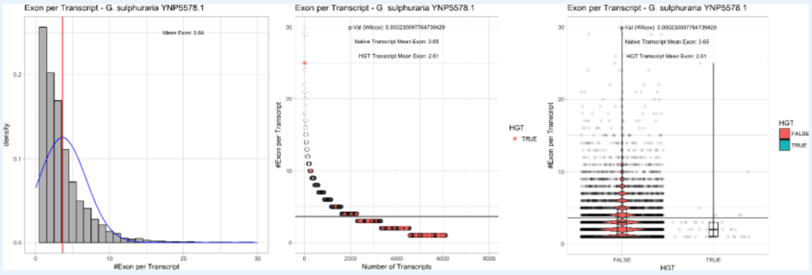
Appendix 4—figure 5. Exon/Intron – *Galdieria sulphuraria MtSh:* (Left) Mid) Cumulative %GC distribution of transcripts. Red line shows the average, blue line a normal distribution based on the average value. The data is categorical (genes have either one, two, three etc. exons) and does not follow a normal distribution. (Mid) Ranking all transcripts based upon their number of exons. Red '*' demarks HGT candidates. As the number of exons was not normally distributed, transcripts were ranked by number of exons. In order to resolve the high number of tied ranks (e.g. many transcripts have two exons) a bootstrap was implied by which the rank of transcripts sharing the same number of exons was randomly assigned 1000 times. Wilcoxon-Mann-Whitney-Test applied for the determination of significant rank differences between the native gene and the HGT candidate subset. (Right) Violin plot showing the number of exons per transcript distribution across native transcripts and HGT candidates.

DOI: https://doi.org/10.7554/eLife.45017.134



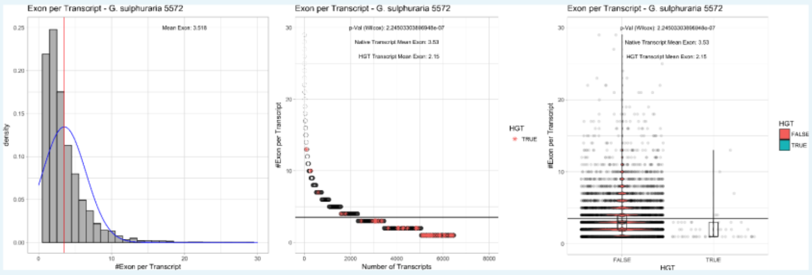
Appendix 4—figure 6. Exon/Intron – *Galdieria sulphuraria Azora:* (Left) Mid) Cumulative %GC distribution of transcripts. Red line shows the average, blue line a normal distribution based on the average value. The data is categorical (genes have either one, two, three etc. exons) and does not follow a normal distribution. (Mid) Ranking all transcripts based upon their number of exons. Red '*' demarks HGT candidates. As the number of exons was not normally distributed, transcripts were ranked by number of exons. In order to resolve the high number of tied ranks (e.g. many transcripts have two exons) a bootstrap was implied by which the rank of transcripts sharing the same number of exons was randomly assigned 1000 times. Wilcoxon-Mann-Whitney-Test applied for the determination of significant rank differences between the native gene and the HGT candidate subset. (Right) Violin plot showing the number of exons per transcript distribution across native transcripts and HGT candidates.

DOI: https://doi.org/10.7554/eLife.45017.135



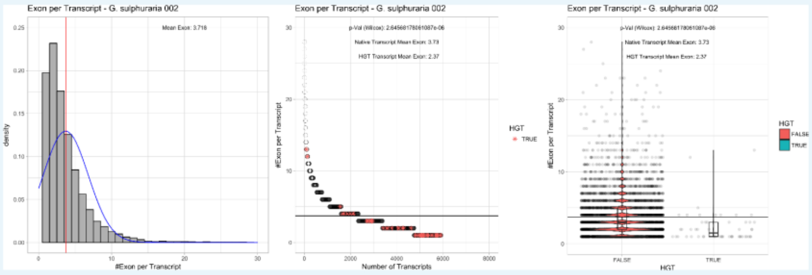
Appendix 4—figure 7. Exon/Intron – *Galdieria sulphuraria YNP5578.1:* (Left) Mid) Cumulative %GC distribution of transcripts. Red line shows the average, blue line a normal distribution based on the average value. The data is categorical (genes have either one, two, three etc. exons) and does not follow a normal distribution. (Mid) Ranking all transcripts based upon their number of exons. Red '*' demarks HGT candidates. As the number of exons was not normally distributed, transcripts were ranked by number of exons. In order to resolve the high number of tied ranks (e.g. many transcripts have two exons) a bootstrap was implied by which the rank of transcripts sharing the same number of exons was randomly assigned 1000 times. Wilcoxon-Mann-Whitney-Test applied for the determination of significant rank differences between the native gene and the HGT candidate subset. (Right) Violin plot showing the number of exons per transcript distribution across native transcripts and HGT candidates.

DOI: https://doi.org/10.7554/eLife.45017.136



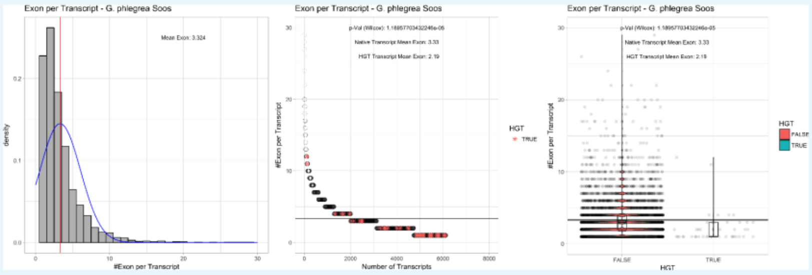
Appendix 4—figure 8. Exon/Intron – *Galdieria sulphuraria 5572:* (Left) Mid) Cumulative %GC distribution of transcripts. Red line shows the average, blue line a normal distribution based on the average value. The data is categorical (genes have either one, two, three etc. exons) and does not follow a normal distribution. (Mid) Ranking all transcripts based upon their number of exons. Red '*' demarks HGT candidates. As the number of exons was not normally distributed, transcripts were ranked by number of exons. In order to resolve the high number of tied ranks (e.g. many transcripts have two exons) a bootstrap was implied by which the rank of transcripts sharing the same number of exons was randomly assigned 1000 times. Wilcoxon-Mann-Whitney-Test applied for the determination of significant rank differences between the native gene and the HGT candidate subset. (Right) Violin plot showing the number of exons per transcript distribution across native transcripts and HGT candidates.

DOI: https://doi.org/10.7554/eLife.45017.137



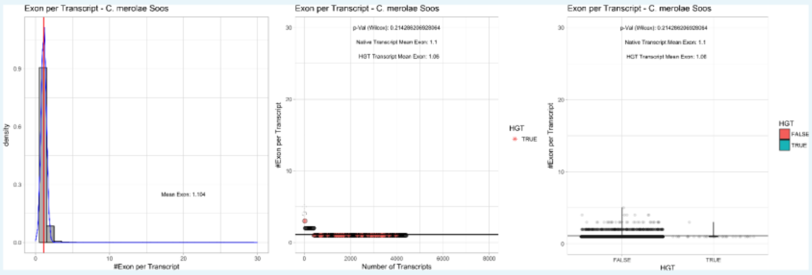
Appendix 4—figure 9. Exon/Intron – *Galdieria sulphuraria 002:* (Left) Mid) Cumulative %GC distribution of transcripts. Red line shows the average, blue line a normal distribution based on the average value. The data is categorical (genes have either one, two, three etc. exons) and does not follow a normal distribution. (Mid) Ranking all transcripts based upon their number of exons. Red '*' demarks HGT candidates. As the number of exons was not normally distributed, transcripts were ranked by number of exons. In order to resolve the high number of tied ranks (e.g. many transcripts have two exons) a bootstrap was implied by which the rank of transcripts sharing the same number of exons was randomly assigned 1000 times. Wilcoxon-Mann-Whitney-Test applied for the determination of significant rank differences between the native gene and the HGT candidate subset. (Right) Violin plot showing the number of exons per transcript distribution across native transcripts and HGT candidates.

DOI: https://doi.org/10.7554/eLife.45017.138



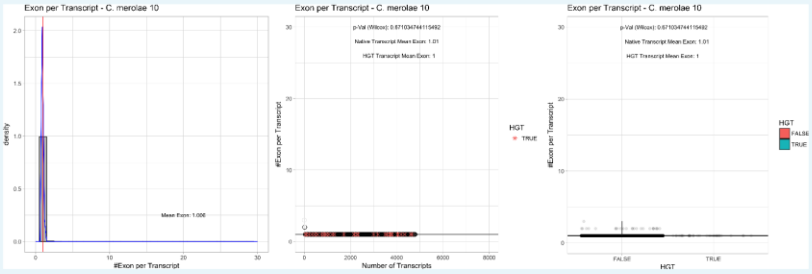
Appendix 4—figure 10. Exon/Intron – *Galdieria phlegrea Soos:* (Left) Mid) Cumulative %GC distribution of transcripts. Red line shows the average, blue line a normal distribution based on the average value. The data is categorical (genes have either one, two, three etc. exons) and does not follow a normal distribution. (Mid) Ranking all transcripts based upon their number of exons. Red '*' demarks HGT candidates. As the number of exons was not normally distributed, transcripts were ranked by number of exons. In order to resolve the high number of tied ranks (e.g. many transcripts have two exons) a bootstrap was implied by which the rank of transcripts sharing the same number of exons was randomly assigned 1000 times. Wilcoxon-Mann-Whitney-Test applied for the determination of significant rank differences between the native gene and the HGT candidate subset. (Right) Violin plot showing the number of exons per transcript distribution across native transcripts and HGT candidates.

DOI: https://doi.org/10.7554/eLife.45017.139



Appendix 4—figure 11. Exon/Intron – *Cyanidioschyzon merolae Soos:* (Left) Mid) Cumulative %GC distribution of transcripts. Red line shows the average, blue line a normal distribution based on the average value. The data is categorical (genes have either one, two, three etc. exons) and does not follow a normal distribution. (Mid) Ranking all transcripts based upon their number of exons. Red '*" demarks HGT candidates. As the number of exons was not normally distributed, transcripts were ranked by number of exons. In order to resolve the high number of tied ranks (e.g. many transcripts have two exons) a bootstrap was implied by which the rank of transcripts sharing the same number of exons was randomly assigned 1000 times. Wilcoxon-Mann-Whitney-Test applied for the determination of significant rank differences between the native gene and the HGT candidate subset. (Right) Violin plot showing the number of exons per transcript distribution across native transcripts and HGT candidates..

DOI: https://doi.org/10.7554/eLife.45017.140



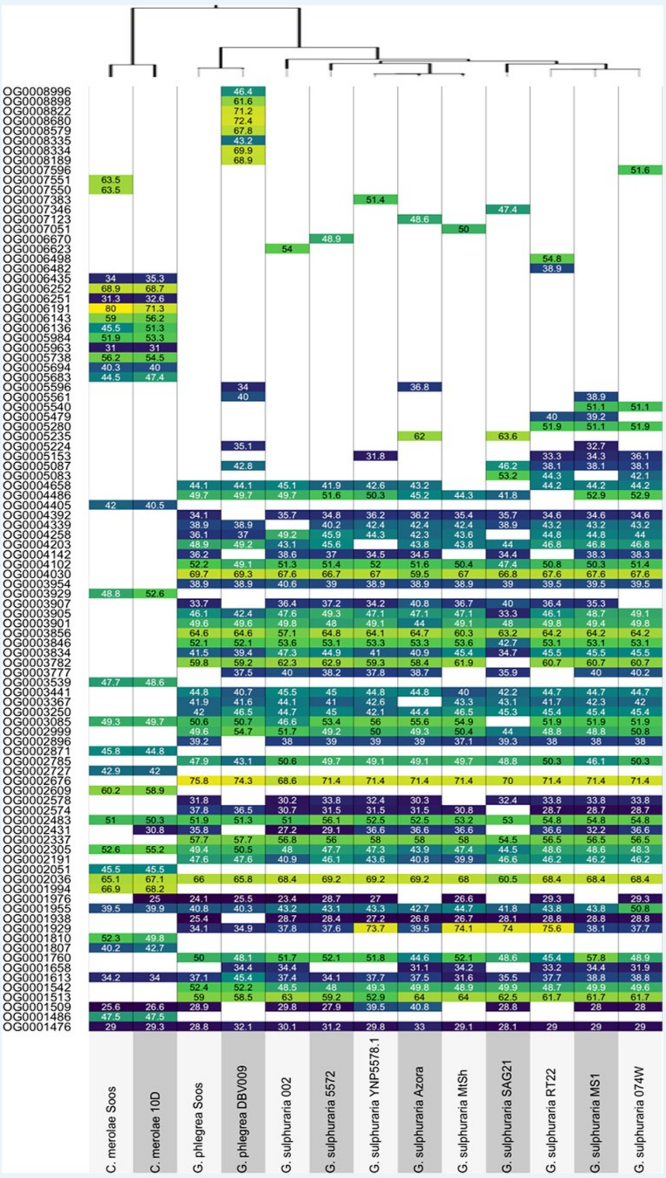
Appendix 4—figure 12. Exon/Intron – *Cyanidioschyzon merolae 074W:* (Left) Mid) Cumulative %GC distribution of transcripts. Red line shows the average, blue line a normal distribution based on the average value. The data is categorical (genes have either one, two, three etc. exons) and does not follow a normal distribution. (Mid) Ranking all transcripts based upon their number of exons. Red '*' demarks HGT candidates. As the number of exons was not normally distributed, transcripts were ranked by number of exons. In order to resolve the high number of tied ranks (e.g. many transcripts have two exons) a bootstrap was implied by which the rank of transcripts sharing the same number of exons was randomly assigned 1000 times. Wilcoxon-Mann-Whitney-Test applied for the determination of significant rank differences between the native gene and the HGT candidate subset. (Right) Violin plot showing the number of exons per transcript distribution across native transcripts and HGT candidates.

DOI: https://doi.org/10.7554/eLife.45017.141

Appendix 5

DOI: https://doi.org/10.7554/eLife.45017.108

Spliceosomal Introns and Exon/Gene



Appendix 5—figure 1. Best Blast Hit between each of the 13 Cyanidiales species and their most similar non-eukaryotic Ortholog in each OG-phylogeny. Values are given as average

percent protein identity between Cyanidiales and non-eukaryotic ortholog. White boxes represent missing Cyanidiales orthologs.

DOI: https://doi.org/10.7554/eLife.45017.143

Manuscript 2 commented by Elizabeth Pennisi (*Science*)

Plants and animals sometimes take genes from bacteria, study of algae suggests

PLANETARY SCIENCE

Windy season fails to revive fading Mars rover

Decision to end Opportunity mission could come in weeks

By **Paul Voosen**

The Opportunity mission is coming to its end. The 15-year-old Mars rover has sat silently for 6 months, and NASA's Jet Propulsion Laboratory (JPL) in Pasadena, California, is running out of tricks to revive it. Agency officials will soon decide whether to end the mission.

In June 2018, a planet-wide dust storm blotted out the sun over Opportunity for several months, strangling it of solar power and draining its batteries. Since then, JPL has sent the golf cart–size rover more than 600 commands to revive it. Engineers hoped seasonal winds, running high between November 2018 and the end of January, would clear the solar panels of dust. But the rover hasn't recovered. "We're running out of time," says John Callas, the mission project manager at JPL.

"The end of the windy season could spell the end of the rover," says Steven Squyres, the mission's principal investigator at Cornell University. "But if this is the end, I can't imagine a better way for it to happen … 15 years into a 90-day mission and taken out by one of the worst martian dust storms in many years."

The martian winter, which in 2011 ended the mission of Opportunity's twin rover, Spirit, is approaching. Sunlight is waning and temperatures are dropping. JPL is trying a few more long shots, such as commands that would tell Opportunity to switch to backup antennas, in case it is awake and trying to use a broken antenna. "After that, I don't know what to do next," Callas says. The plan was to have NASA headquarters weigh in on whether to continue the efforts after the windy season, he adds. Such a decision could come within weeks from agency science chief Thomas Zurbuchen.

In 2014, Opportunity made a shadow portrait in the late afternoon light.

PHOTO: NASA/JPL-CALTECH

Opportunity will leave a trail of superlatives. Although JPL only promised it would last 90 days on Mars, it ended up enduring at least 5000. It traversed a path 45 kilometers long, often driving backward because of an overheating steering control. Even after all that time, its cameras were still working beautifully, says Jim Bell, a planetary scientist at Arizona State University in Tempe who leads the rover's color camera team. Bell, for one, isn't giving up hope. The rover is perched on the rim of Endeavor crater, he notes, exposed to wind gusts that might still revive it. "No one has ever won a bet against it. I'm not about to start."

The mission explored whether Mars could have hosted life in the deep past. Soon after landing in Meridiani Planum in 2004, it revealed the first signs of past habitability when it drove across sulfate-rich sandstones. The stones likely formed as shallow muds in lagoons, says Raymond Arvidson, the rover's deputy principal investigator at Washington University in St. Louis, Missouri. "There was an ephemeral lake system, going dry, going wet. That's a huge discovery."

The rover later found more evidence for long periods of past habitability. Near crater rims, it spotted veins of gypsum, which forms as water evaporates. And, in 2013, it provided the first surface observations of 4-billion-year-old clays, a sign of truly abundant water. The finding, 9 years into its mission, validated observations from orbit and expanded the hunt for such clays, says Alberto Fairén, a planetary scientist at Cornell.

Few expected when they signed up for the mission that they'd still be working 15 years later. In the end, though, Bell adds, "Mars always wins." ■

EVOLUTION

Algae suggest eukaryotes get many gifts of bacteria DNA

Analysis revives debate on "horizontal gene transfer"

By **Elizabeth Pennisi**

Algae found in thermal springs and other extreme environments have heated up a long-standing debate: Do eukaryotes—organisms with a cell nucleus—sometimes get an evolutionary boost in the form of genes transferred from bacteria? The genomes of some red algae, single-celled eukaryotes, suggest the answer is yes. About 1% of their genes have foreign origins, and the borrowed genes may help the algae adapt to their hostile environment.

The new research, posted last week as a preprint on bioRxiv, has not persuaded the most vocal critic of the idea that eukaryotes regularly receive beneficial bacterial DNA. But other scientists have been won over. The group provides a "fairly nice, rock-solid case for horizontal gene transfer" into eukaryotes, says Andrew Roger, a protist genomicist at Dalhousie University in Halifax, Canada.

Many genome studies have shown that prokaryotes—bacteria and archaea—liberally swap genes among species, which influences their evolution. The initial sequencing of the human genome suggested our species, too, has picked up microbial genes. But further work demonstrated that such genes found in vertebrate genomes were often contaminants introduced during sequencing.

In 2015, after analyses of millions of protein sequences across many species, William Martin, a biologist at the University of Dusseldorf (UD) in Germany, and colleagues concluded in *Nature* that there is no significant ongoing transfer of prokaryotic genes into eukaryotes. Martin believes any such transfers only occurred episodically early in the evolution of eukaryotes, as they internalized the bacteria that eventually became organelles such mitochondria or chloroplasts. If bacterial genes were continually moving into eukaryotes and being put to use, Martin says, a pattern of such gene accumulation should be discernible within

III.2C Comment 1 of Manuscript 2

the eukaryotic family tree, but there is none.

Debashish Bhattacharya, an evolutionary genomicist at Rutgers University in New Brunswick, New Jersey, and UD plant biochemist Andreas Weber took a closer look at a possible case of bacteria-to-eukaryote gene transfer that Martin has challenged. The initial sequencing of genomes from two species of red algae called Cyanidiophyceae had indicated that up to 6% of their DNA had a prokaryotic origin. These so-called extremophiles, which live in acidic hot springs and even inside rock, can't afford to maintain superfluous DNA. They appear to contain only genes needed for survival. “When we find a bacterial gene, we know it has an important function or it wouldn't last” in the genome, Bhattacharya says.

He and Weber turned to a newer technology that deciphers long pieces of DNA. The 13 red algal genomes they studied contain 96 foreign genes, nearly all of them sandwiched between typical algal genes in the DNA sequenced, which makes it unlikely they were accidentally introduced in the lab. “At the very least, this argument that [putative transferred genes are] all contamination should finally be obsolete,” says Gerald Schoenknecht, a plant physiologist at Oklahoma State University in Stillwater.

The transferred genes seem to transport or detoxify heavy metals, or they help the algae extract nourishment from the environment or cope with high temperature and other stressful conditions. “By acquiring genes from extremophile prokaryotes, these red algae have adapted to more and more extreme environments,” Schoenknecht says.

Martin says the new evidence doesn't persuade him. “They go to great lengths to find exactly what I say they should find if [horizontal gene transfer to eukaryotes] is real, but they do not find it,” he asserts. Others argue that gene transfer to eukaryotes is so rare, and the pressure to get rid of any but the most important borrowed genes is so strong, that transferred genes might not accumulate over time as Martin expects.

Of course, Roger says, “What's happening in red algae might not be happening in animals like us.” Humans and all other multicellular eukaryotes, including plants, have specialized reproductive cells, such as sperm or eggs or their stem cells, and only bacterial genes picked up by those cells could be passed on.

Despite this obstacle, several insect researchers say they see evidence of such gene transfer. John McCutcheon, a biologist at Montana State University in Missoula who studies mealy bugs, is one. “I've moved beyond asking ‘if [the bacterial genes] are there,’ to how they work,” he says. The red algae, he adds, “is a very clear case.” ■

U.S. RESEARCH FUNDING

Shutdown ends, but not worry

Science agencies dig out amid fears of yet another closure

By **Jeffrey Mervis** *and* **David Malakoff**

The longest U.S. government shutdown in history is over. But federal research agencies that were shuttered for 35 days won't be returning to normal anytime soon, officials warn. And any relief could be fleeting: Another closure could come on 16 February if Congress and President Donald Trump can't agree on funding for Trump's proposed border wall.

“Scientists will need to be patient” as agencies dig out from an avalanche of incomplete paperwork, unanswered emails, and canceled meetings, says Sarah Nusser, vice president for research at Iowa State University in Ames. “You're not going to get all your questions answered immediately.”

The impasse, which began on 22 December 2018 and ended on 25 January, halted most operations at more than a half-dozen agencies that conduct or fund research, including NASA, the National Science Foundation (NSF), the Food and Drug Administration, and the departments of agriculture, the interior, and commerce. It left more than 800,000 federal workers, contractors, and grantees, including many postdoctoral researchers, without paychecks. It forced many academic researchers to cancel or delay planned projects and froze the grantmaking machinery at key agencies at one of the busiest times of year for reviewing proposals.

A new shutdown would be like “a concussion followed by a repeat blow ...”

Donald Weber,
U.S. Department of Agriculture

This week, agencies began to regain their footing after Congress and the White House agreed to a so-called continuing resolution (CR) that funds the agencies at existing budget levels for 3 weeks. The pause is supposed to give negotiators time to finalize a deal on the wall and pending spending bills that would increase the budgets of many of the reopened agencies.

At NSF, the restart means processing routine award transactions that were frozen, resuming conversations with scientists who have questions about current awards or upcoming competitions, and rescheduling more than 100 review panels—involving 2000 proposals—that were scrubbed during the shutdown. But, “We won't know all of the challenges we face” until staff have had a chance to settle in, Amanda Hallberg Greenwell, head of NSF's Office of Legislative and Public Affairs in Alexandria, Virginia, said as the agency prepared to reopen.

At NASA headquarters in Washington, D.C., returning employees were clearing spiderwebs and dusting workspaces as harried support staff helped restart computers and smartphones that had been turned off. “It took me nearly 4 hours just to reset my passwords and get software updates,” one NASA researcher, who is not authorized to speak to the press, told *Science*.

At the National Oceanic and Atmospheric Administration in Silver Spring, Maryland, part of the Department of Commerce, one priority was getting public websites that provide a wealth of earth science data back online. Another was rescheduling fisheries surveys and other research cruises that had been canceled.

But reopening the government may not be enough to save some research projects. “We are having to do some serious priority-setting. ... Research projects will have to undergo triage,” says Donald Weber, a biocontrol scientist at the U.S. Department of Agriculture's Beltsville Agricultural Research Center in Maryland. That could mean killing off field studies too delayed by the shutdown to pursue. And the possibility of another closure in just a few weeks has everyone on edge. “A repeat shutdown would be very damaging,” Weber says. “I liken it to a concussion followed by a repeat blow, which would render agency programs punch-drunk.”

University groups are urging Congress and the White House to end the agony for agencies by passing final 2019 spending bills—regardless of the outcome for the wall—and not another CR. “[T]he U.S. research enterprise,” said Peter McPherson, president of the Association of Public and Land-grant Universities in Washington, D.C., “does not operate anywhere close to full strength when agencies are only guaranteed to be open 3 weeks at a time.” ■

Published by AAAS

Manuscript 2 commented by Carolin M Kobras and Daniel Falush (*eLife*)

Gene Transfer: Adapting for life in the extreme

eLIFE
elifesciences.org

INSIGHT

GENE TRANSFER

Adapting for life in the extreme

Red algae have adapted to extreme environments by acquiring genes from bacteria and archaea.

CAROLIN M KOBRAS AND DANIEL FALUSH

Related research article Rossoni AW, Price DC, Seger M, Lyska D, Lammers P, Bhattacharya D, Weber APM. 2019. The genomes of polyextremophilic Cyanidiales contain 1% horizontally transferred genes with diverse adaptive functions. *eLife* **8**:e45017. DOI: 10.7554/eLife.45017

Most humans have nearly the same complement of genes, all of which have come from our primate ancestors (***Salzberg, 2017***). On the other hand, even closely related strains of the bacterium *Escherichia coli* can differ by hundreds of genes (***Touchon et al., 2009***) despite having a much smaller genome. These genes have been acquired via a process called horizontal gene transfer (HGT), which is an important driver of adaptation, as it allows bacteria and other prokaryotes to gain the genes they need in order to thrive in certain environments (***Koonin et al., 2001***). Moreover, this exchanging of genes has resulted in many genetic elements in prokaryotes becoming highly mobile, making it easier for DNA to be transferred to a diverse range of hosts.

HGT has also been observed in animals, plants and other eukaryotes (***Husnik and McCutcheon, 2018***), but its role in determining genome composition and facilitating adaptation in these species remains unclear (***Ku and Martin, 2016***). Now, in eLife, Andreas Weber and co-workers at Heinrich Heine University, Arizona State University and Rutgers University – including Alessandro Rossoni as first author – report evidence for HGT between prokaryotes and the red alga Cyanidiales (***Rossoni et al., 2019***). These are remarkable single-cell organisms that can perform photosynthesis at temperatures up to 56˚C, and can live in extreme environments such as hot springs and acid rivers (***Schönknecht et al., 2013***). Cyanidiales can also be used to investigate HGT over geological timescales because they share a common ancestor that dates back 800 million years to a time before animals had even evolved.

Based on an analysis of ten new and three previously reported Cyanidiales genomes, Rossoni et al. found that 1% of genes had been obtained via HGT. Moreover, many of these genes coded for proteins that were needed to survive in extreme environments (such as proteins involved in detoxifying heavy metals like arsenic or mercury, or removing free radicals; ***Figure 1***). Additionally, prokaryotes adapted to the same extreme environment as Cyanidiales were commonly identified as the source of these genes. It seems likely, therefore, that HGT influenced the evolution of Cyanidiales, especially because the criterion used to detect HGT was conservative and the study did not attempt to detect gene transfer from other eukaryotes.

Comparing the new Cyanidiales genes to genes found in present-day bacteria and archaea databases did not yield any recent examples of HGT. This absence of recent events is unsurprising, as Rossoni et al. estimated that Cyanidiales acquire just one gene via HGT every 14.6 million years – the same amount of time it took for humans to diverge from the orangutan. Such a low rate makes finding a fresh transfer in a small number of genomes unlikely. Instead, the majority of HGT candidate genes found by Rossoni et al. have acquired introns (non-protein coding

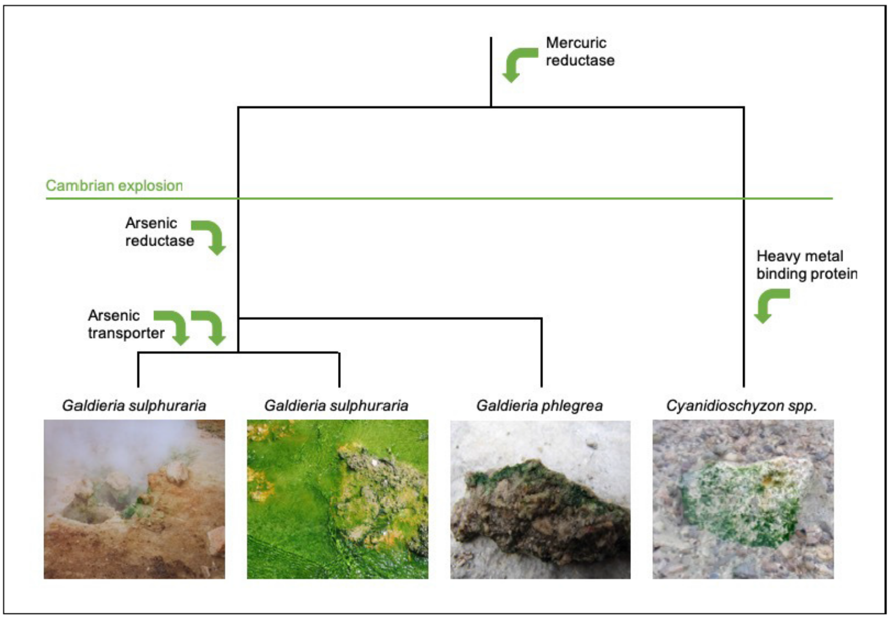


Figure 1. Horizontal gene transfer in the evolution of red algae. The evolutionary trajectory of the red algae Cyanidiales is shown from top to bottom. Rossini et al. investigated genetic changes that took place before and after the Cambrian explosion 541 million years ago, and found that Cyanidiales obtained 1% of their genes during this time by horizontal transfer. Many of these genes allowed Cyanidiales to adapt to extreme environments, such as genes related to the detoxification of heavy metals including mercury and arsenic (represented by green arrows). Some of the lineages of Cyanidiales that were sequenced by Rossoni et al. are shown in the bottom panels: two of these have the same taxonomic name despite having diverged from one another millions of years ago. Image credit: Andreas Weber (left panel), Debashish Bhattacharya (two middle panels), and Shin-ya Miyagishima (right panel).

segments of DNA), and then persisted over hundreds of millions of years.

Despite there being evidence to show HGT occurred, it still remains unclear how these transfers took place. The best-studied mechanisms by which eukaryotes acquire DNA from other organisms are sexual reproduction and by transferring DNA from symbionts (biological organisms that live cooperatively with other organisms). However, meiotic sex only occurs between closely related species, and therefore cannot explain how Cyanidiales appear to have gained DNA from such a diverse range of prokaryotes: moreover, the evolution of symbiotic transfer is uncommon in most taxonomic groups. Instead DNA was more likely obtained via viral infection or plasmids (circular molecules of double stranded DNA) being transferred between prokaryotes and eukaryotes (*Heinemann and Sprague, 1989*). Indeed, a recent study has shown that many eukaryotes, including red algae, can acquire plasmids carrying genes derived from plants, viruses and bacteria (*Lee et al., 2016*).

The work of Rossoni et al. suggests that, in terms of gene content evolution, Cyanidiales are more similar to humans than to *E. coli*, which is consistent with previous qualitive comparisons of HGT patterns in eukaryotes and prokaryotes (*Ku and Martin, 2016*). However, a number of mysteries still remain. For example, what are the most common modes of plasmid transmission in Cyanidiales? How do plasmids maintain themselves in populations? How often do they jump between species, and how far do they jump? To answer these questions we should first observe what is happening all around us today (*Popa et al., 2017*) and, if possible, study events that occur more frequently than once every 14.6 million years.

Carolin M Kobras is in the Milner Centre for Evolution, University of Bath, Bath, UK
https://orcid.org/0000-0003-4393-5829

Daniel Falush is in the Milner Centre for Evolution, University of Bath, Bath, UK
danielfalush@googlemail.com
https://orcid.org/0000-0002-2956-0795

Competing interests: The authors declare that no competing interests exist.

Published 15 July 2019

References

Heinemann JA, Sprague GF. 1989. Bacterial conjugative plasmids mobilize DNA transfer between bacteria and yeast. *Nature* **340**:205–209. DOI: https://doi.org/10.1038/340205a0, PMID: 2666856

Husnik F, McCutcheon JP. 2018. Functional horizontal gene transfer from bacteria to eukaryotes. *Nature Reviews Microbiology* **16**:67–79. DOI: https://doi.org/10.1038/nrmicro.2017.137, PMID: 29176581

Koonin EV, Makarova KS, Aravind L. 2001. Horizontal gene transfer in prokaryotes: quantification and classification. *Annual Review of Microbiology* **55**:709–742. DOI: https://doi.org/10.1146/annurev.micro.55.1.709, PMID: 11544372

Ku C, Martin WF. 2016. A natural barrier to lateral gene transfer from prokaryotes to eukaryotes revealed from genomes: the 70 % rule. *BMC Biology* **14**:89. DOI: https://doi.org/10.1186/s12915-016-0315-9, PMID: 27751184

Lee J, Kim KM, Yang EC, Miller KA, Boo SM, Bhattacharya D, Yoon HS. 2016. Reconstructing the complex evolutionary history of mobile plasmids in red algal genomes. *Scientific Reports* **6**:23744. DOI: https://doi.org/10.1038/srep23744, PMID: 27030297

Popa O, Landan G, Dagan T. 2017. Phylogenomic networks reveal limited phylogenetic range of lateral gene transfer by transduction. *The ISME Journal* **11**: 543–554. DOI: https://doi.org/10.1038/ismej.2016.116, PMID: 27648812

Rossoni AW, Price DC, Seger M, Lyska D, Lammers P, Bhattacharya D, Weber APM. 2019. The genomes of polyextremophilic Cyanidiales contain 1% horizontally transferred genes with diverse adaptive functions. *eLife* **8**:e45017. DOI: https://doi.org/10.7554/eLife.45017, PMID: 31149898

Salzberg SL. 2017. Horizontal gene transfer is not a hallmark of the human genome. *Genome Biology* **18**: 85. DOI: https://doi.org/10.1186/s13059-017-1214-2, PMID: 28482857

Schönknecht G, Chen WH, Ternes CM, Barbier GG, Shrestha RP, Stanke M, Bräutigam A, Baker BJ, Banfield JF, Garavito RM, Carr K, Wilkerson C, Rensing SA, Gagneul D, Dickenson NE, Oesterhelt C, Lercher MJ, Weber AP. 2013. Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote. *Science* **339**:1207–1210. DOI: https://doi.org/10.1126/science.1231707, PMID: 23471408

Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S, Bouchier C, Bouvet O, Calteau A, Chiapello H, Clermont O, Cruveiller S, Danchin A, Diard M, Dossat C, Karoui ME, Frapy E, Garry L, et al. 2009. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLOS Genetics* **5**:e1000344. DOI: https://doi.org/10.1371/journal.pgen.1000344, PMID: 19165319

Manuscript 3

Cold Acclimation of the Thermoacidophilic Red Alga *Galdieria sulphuraria*: Changes in Gene Expression and Involvement of Horizontally Acquired Genes

PCP
PLANT & CELL PHYSIOLOGY

Regular Paper

Cold Acclimation of the Thermoacidophilic Red Alga *Galdieria sulphuraria*: Changes in Gene Expression and Involvement of Horizontally Acquired Genes

Alessandro W. Rossoni[1,5], Gerald Schönknecht[2,5,*], Hyun Jeong Lee[3], Ryan L. Rupp[4], Samantha Flachbart[1], Tabea Mettler-Altmann[1], Andreas P.M. Weber[1] and Marion Eisenhut[1]

[1]Institute of Plant Biochemistry, Cluster of Excellence on Plant Sciences (CEPLAS), Heinrich Heine University, 40225 Düsseldorf, Germany
[2]Department of Plant Biology, Ecology & Evolution, Oklahoma State University, Stillwater, OK 74078, USA
[3]Graduate School of Semiconductor and Chemical Engineering, Chonbuk National University, Jeonju, South Korea
[4]Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI, 48824, USA
[5]These authors contributed equally to this work
*Corresponding author: E-mail, gschoen@okstate.edu.

(Received October 6, 2018; Accepted December 14, 2018)

Galdieria sulphuraria is a unicellular red alga that lives in hot, acidic, toxic metal-rich, volcanic environments, where few other organisms survive. Its genome harbors up to 5% of genes that were most likely acquired through horizontal gene transfer. These genes probably contributed to *G.sulphuraria*'s adaptation to its extreme habitats, resulting in today's polyextremophilic traits. Here, we applied RNA-sequencing to obtain insights into the acclimation of a thermophilic organism towards temperatures below its growth optimum and to study how horizontally acquired genes contribute to cold acclimation. A decrease in growth temperature from 42°C/46°C to 28°C resulted in an upregulation of ribosome biosynthesis, while excreted proteins, probably components of the cell wall, were downregulated. Photosynthesis was suppressed at cold temperatures, and transcript abundances indicated that C-metabolism switched from gluconeogenesis to glycogen degradation. Folate cycle and S-adenosylmethionine cycle (one-carbon metabolism) were transcriptionally upregulated, probably to drive the biosynthesis of betaine. All these cold-induced changes in gene expression were reversible upon return to optimal growth temperature. Numerous genes acquired by horizontal gene transfer displayed temperature-dependent expression changes, indicating that these genes contributed to adaptive evolution in *G.sulphuraria*.

Keywords: Cold stress • *Galdieria sulphuraria* • Horizontal gene transfer • RNA-Seq • Systems biology • Thermoacidophilic red alga.

Abbreviations: DEG, differentially expressed genes; HGT, horizontal gene transfer; RNA-seq, RNA-sequencing; SAM, S-adenosylmethionine; *T*-dependent, temperature-dependent.

Introduction

Galdieria sulphuraria is a polyextremophile unicellular red alga, thriving on—and in—soil and rocks in volcanic habitats across the world. It can endure extremely low pH (down to 0), high temperatures (up to 56°C), salt stress (up to 1.5 M NaCl), arsenic (up to a few g L^{-1}) and toxic heavy metals (e.g. 200 µg g^{-1} mercury) at concentrations lethal for most organisms (Doemel and Brock 1971, Reeb and Bhattacharya 2010). *Galdieria sulphuraria* can grow photoautotrophic, i.e. perform oxygenic photosynthesis, or can grow heterotrophic, utilizing a large variety of different metabolites as external carbon and energy source (Gross and Schnarrenberger 1995, Gross 1999). This metabolic flexibility in combination with its polyextremopilic traits has lately risen the interest in its biotechnological application. Pilot studies evaluated *G.sulphuraria*'s performance for urban wastewater treatment (Henkanatte-Gedera et al. 2017), in bioremediation (Fukuda et al. 2018), as food ingredient (Graziani et al. 2012), or for phycocyanin production (Sloth et al. 2006, Sloth et al. 2017).

Based on a high quality genome sequence it had been postulated that *G.sulphuraria*'s unusual polyextremophile life style was facilitated by horizontal gene transfer (HGT) from extremophile bacteria and archaea (Schönknecht et al. 2013). HGT is defined as transmission of genetic material between organisms other than by ('vertical') transmission from parents to their offspring (Koonin et al. 2001, Soucy et al. 2015). Whereas HGT in bacteria and archaea is widely recognized as important driver of adaptive evolution, the occurrence of HGT from non-eukaryotes to eukaryotes outside the context of pathogenicity and endosymbiosis remains disputed (Leger et al. 2018). HGT, followed by gene duplications probably gave rise to about 5% of protein-coding genes in the *G.sulphuraria* genome (Schönknecht et al. 2013). This estimate is based on both, phylogenetic incongruences between protein and organismal trees, and significant differences in genome signatures. The possibility of bacterial contamination could be excluded because most genes of bacterial or archaeal origin were located on large contigs where they were flanked by eukaryotic genes. Moreover, many genes of bacterial or archaeal origin had acquired introns, further excluding the possibility of a contamination. The analyses identifying HGT candidates in *G.sulphuraria* fulfilled most of the criteria recommended for reliable identification of HGT

Plant Cell Physiol. 0(0): 1–11 doi:10.1093/pcp/pcy240, Advance Access publication on 27 December 2018, available online at www.pcp.oxfordjournals.org

by Richards and Monier (2016). *Galdieria sulphuraria*'s genome was shaped by two phases of massive genome reduction and gene loss during the course of its evolution towards extremophily (Qiu et al. 2015). Although this reduced coding capacity may be partially compensated by alternative RNA splicing (Qiu et al. 2018), HGT has probably compensated some of the gene loss. So far, transcriptional evidence showing a functional activity of HGT candidates in *G.sulphuraria* has not been provided.

Here, we studied how a thermophilic organism, with an optimum growth temperature above 40°C, reacts to being exposed to 28°C. Besides data on growth rates (Doemel and Brock 1970), enzyme kinetics (McCoy et al. 2009) and a report that particular strains of *Galdieria* are able to grow in temperate environments (Gross et al. 2002), the systems biology of *G.sulphuraria* at temperatures below its growth optimum has remained unexplored. A temperature decrease was also expected to cause changes in expression levels of numerous genes, to analyze the involvement of HGT candidates in temperature acclimation. We aimed at (i) obtaining insight into how a thermophile organism reacts to cold stress and (ii) evaluating the impact of HGT candidates on the acclimation process to temperature decrease.

Results and Discussion

Which genes have the highest expression levels?

Before looking at temperature-dependent (*T*-dependent) changes in expression levels, it seems interesting to inspect, which genes do show the highest expression levels under 'normal' growth conditions, i.e. before temperature changes were applied. The 21 genes with highest expression levels, i.e. more than 40,000 Counts (**Table 1**), encode proteins that fall into three categories, (i) eight are hypothetical proteins, (ii) nine are part of the photosynthetic machinery, and (iii) the remaining four encode enzymes of the core carbon metabolism. There are hints that the four hypothetical proteins with the highest transcript levels might be components of the cell wall. Three of them contain an N-terminal secretory signal peptide (**Table 1**). Two of them are tyrosine-rich (Gasu_09270.1, 24.4% Y; Gasu_32120.1, 16.7% Y), and tyrosine-rich proteins are typical components of extracellular matrix in animals, play a key role in mussel adhesive proteins (Silverman and Roberto 2007), and have been detected in diatom cell walls (Buhmann et al. 2014). Searches for weak sequence similarities with PSI-BLAST (Altschul et al. 1997) and HHpred (Biegert et al. 2006) indicate distant relationships with a fasciclin domain-containing collagen-associated protein (Gasu_43630.1), with mucin (Gasu_09270.1 and Gasu_32120.1), or with fibrillin (Gasu_07490.1). This indicates that the four hypothetical proteins with the highest transcript levels are distantly related to proteins contributing to the extracellular matrix in animals. It has been reported that the cell wall of *Galdieria* contains 50% and more of protein (Bailey and Staehelin 1968). Very little is known about cell wall composition of unicellular thermoacidophilic red alga and the four highly expressed hypothetical proteins with weak similarity to extracellular matrix proteins seem to be good candidates for cell wall proteins. Analyses of 261 genes with the highest expression levels (Counts > 5000; sum up to 47% of all Counts) indicate enrichment of transcripts encoding proteins for photosynthesis (MAPMAN Bin 1, 8.7-fold, $P = 1.1 \times 10^{-13}$; GO: 0015979, 7.0-fold, 6.0×10^{-5}), translation (MAPMAN Bin 29, 1.7-fold, $P = 1.2 \times 10^{-4}$; GO: 0006412, 5.0-fold, $P = 5.0 \times 10^{-12}$) and glycolysis (MAPMAN Bin 4, 3.8-fold, $P = 0.018$; GO: 0006096, 5.1-fold, $P = 0.007$). Transcripts for proteins containing a putative signal sequence for excretion were 2.4-fold enriched ($P = 5.3 \times 10^{-4}$) among the top 261 transcripts. As in other photosynthetic eukaryotes (Külahoglu et al. 2014), in *G.sulphuraria* a large part of the transcriptional investment goes into photosynthesis, translation and primary carbon metabolism. In contrast to other photosynthetic eukaryotes, a considerable fraction of transcripts seems to encode putative cell wall proteins.

T-dependent changes in expression levels

To analyze transcriptional changes in response to temperature decrease, we performed a temperature shift experiment. *Galdieria sulphuraria* cells were cultivated at 42°C, which is the strain's standard growth temperature. Cells were shifted for 48 h to 28°C, shifted to 46°C for 48 h, and shifted to 28°C again for a final 48 h. Samples were taken in biological triplicates immediately before, 3 h and 12 h after each temperature change (see Supplementary Fig. S1). Growth rates remained constant during the course of the experiment, absorbance at 750 nm increased continuously from A_{750} = 0.65 to A_{750} = 0.85. This indicates that the applied temperature changes (42°C → 28°C → 46°C → 28°C) did probably not cause severe stress.

A principal component analysis of all expression values (Supplementary Fig. S2) showed large treatment effects compared to smaller scattering among biological replicates. Samples taken at higher (42°C/46°C) or at lower temperatures (28°C) are well separated when projected onto the plane described by the first two principal components. For 6545 out of 6623 annotated protein-coding genes transcripts were detected in more than 12 of 25 samples. For the clear majority of genes (6358 or 96%) transcripts were detected in all 25 samples. At a false discovery rate (FDR) < 1%, 3898 genes (59% of annotated genes) displayed a significant change in expression between the first time point ($t = 0$, $T = 42$°C) and at least one of the nine following time points. At most time points, the number of genes showing increased expression and the number showing decreased expression was roughly the same (Supplementary Fig. S3).

T-dependent expression changes of genes involved in protein biosynthesis

To obtain insight into transcriptional regulation of acclimation to cold stress in *G.sulphuraria*, *T*-dependent expression changes of annotated transcription factor genes were analyzed. Only one gene annotated as transcription factor displayed a rapid, 4.5-fold upregulation in response to temperature decrease, *Gasu_41350.1*, a Myb family transcription factor. The transcription factors ICE1 and CBF1 to CBF3, which control cold responses in higher plants, do not seem to have homologs in *G.sulphuraria*, or Rhodophyta in general. When we extended

Plant Cell Physiol. 0(0): 1–11 (2019) doi:10.1093/pcp/pcy240

Table 1 Genes with highest expression

Gene ID	Counts	Annotation	
Gasu_43630.1	322,152	Hypothetical protein (excreted?)	
Gasu_09270.1	174,832	Hypothetical protein	
Gasu_07490.1	155,379	Hypothetical protein (excreted?)	
Gasu_33650.1	146,828	Phycobilisome linker polypeptide	PS
Gasu_32120.1	91,788	Hypothetical protein (excreted?)	
Gasu_19410.1	85,684	RUBISCO activase	PS
Gasu_57970.1	68,148	Alcohol dehydrogenase	C-Metabolism
Gasu_23060.1	67,858	Hypothetical protein	
Gasu_34810.1	59,871	Hypothetical protein	
Gasu_57960.1	50,727	Glycerol dehydrogenase (HGT)	C-Metabolism
Gasu_15560.1	48,024	Hypothetical protein	
Gasu_18750.1	47,703	Hypothetical protein	
Gasu_56920.1	46,838	Ferredoxin component	PS
Gasu_42970.1	44,133	Photosystem I subunit O precursor	PS
Gasu_16200.1	43,404	Phycobilisome core linker protein	PS
Gasu_62010.1	42,970	Acetaldehyde dehydrogenase II (HGT)	C-Metabolism
Gasu_11610.1	42,097	Light-harvesting complex protein	PS
Gasu_37550.1	41,883	Light-harvesting complex protein	PS
Gasu_63700.1	41,687	Fructose-bisphosphate aldolase	C-Metabolism
Gasu_27040.1	40,873	Photosystem II PsbO protein	PS
Gasu_57520.1	40,637	Light-harvesting complex protein	PS

For 21 genes with the highest expression (>40,000 Counts; sum up to 19% of all Counts) gene ID, mean Counts and annotation are given. The ranking here is based on Counts and not on FPKM values, which have been used for all other analyses. Not being normalized by exonic length and library size, Counts provide an impression about how much the organism invests in making mRNA from a certain gene. 'HGT' indicates that this gene was probably acquired by horizontal gene transfer; 'excreted?' indicates that SignalP4 (Petersen et al. 2011) detected a secretory signal peptide; 'PS' and 'C-Metabolism' indicate genes involved in photosynthesis or core C-metabolism, respectively.

the search for homologs to all 27 transcription factor genes that are induced in *Arabidopsis* by cold in parallel to CBF1 to CBF3 (Park et al. 2015), homologous (10^{-30} < e-value < 10^{-20}) putative transcription factors were detected in *G.sulphuraria*. However, most of these transcription factor genes had decreased expression levels at lower temperatures, some more than 4-fold lower at 28°C (*Gasu_13060.1, Gasu_23730.1 and Gasu_41660.1*). There seems to be little conservation among the transcription factors that control acclimation to cold stress between green plants and Rhodophyta.

An EdgeR analysis comparing all samples at high (42°C/46°C) against all samples at low temperature (28°C) identified 1590 differentially expressed genes (DEG; FDR < 0.01), 753 with higher expression at high temperatures and 837 with higher expression at low temperature. Among DEG's differing between all warmer (42°C/46°C) and all colder temperatures (28°C), pathway analyses identified 'Ribosome' (Kyoto Encyclopedia of Genes and Genomes (KEGG) map 03010) as 4.9-fold enriched ($P = 6.2 \times 10^{-32}$) and 'Ribosome biogenesis in eukaryotes' (KEGG map 03008) as 3.5-fold enriched ($P = 2.2 \times 10^{-7}$). Hierarchical cluster analysis of the complete data set (Supplementary Fig. S4) resulted, among other clusters, in one cluster, where one third of genes (37 of 106) encode ribosomal proteins (Supplementary Fig. S5). Out of 147 ribosomal proteins annotated in the genome of *G.sulphuraria*, 37 (i.e. 25%) were in this cluster. Genes in this cluster showed increased expression after 3 h, at decreased temperature ($T = 28$°C), with expression levels staying elevated for the remaining 45 h cold period. Upon increasing temperature to $T = 46$°C, expression of these genes decreased within 3 h, and remained low (**Fig. 1**). Increase in transcript abundance, at colder temperatures (28°C), in this cluster was between 2- and 70-fold (median 6.5). Genes with the terms 'ribosome' or 'ribosomal' in their annotation were 16-fold enriched in this cluster ($P = 4.1 \times 10^{-38}$). Genes associated with the GO term 'structural constituent of ribosome' (GO: 0003735) were 24-fold enriched in this cluster ($P = 9.2 \times 10^{-38}$). Other genes, in this cluster and neighboring clusters, encode additional proteins required for protein biosynthesis. Different analyses indicated an increase in expression of genes encoding the translational machinery at colder temperatures (28°C).

This increase in expression of genes encoding the translational machinery was not unexpected. Protein biosynthesis is optimized to a certain temperature range. A reduction in growth temperature results in a reduction of protein biosynthesis rates, and to compensate for this, ribosome production is increased. Increased expression of ribosomal proteins upon temperature decrease has been observed in eubacteria (Jones and Inouye 1994) as well as eukaryotes (Tai et al. 2007). This is a clear indication that *G.sulphuraria* is thermophilic and not

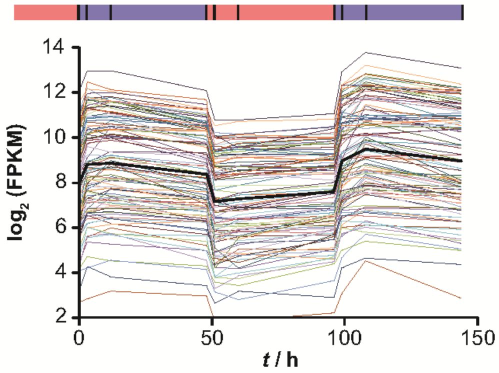


Fig. 1 Cold-induced upregulation of genes involved in protein biosynthesis. Time course of log$_2$(FPKM) values (biological replicates averaged) from 106 genes forming a cluster (compare Supplementary Fig. S5) that contains 37 genes encoding ribosomal proteins. The heavier, black trace represents the average of all 106 expression levels. The colored bar on top indicates the temperature protocol (light red for 42°C/46°C, light blue for 28°C) and time points when samples were taken (vertical black lines). Genes in this cluster display a rapid, lasting increase in expression at 28°C, and a rapid, lasting decrease in expression at 46°C.

thermotolerant. Protein biosynthesis in *G.sulphuraria* is adapted to high temperatures, and a reduction to 28°C is resulting in cold stress.

When exposed to growth temperatures below the optimal growth temperature, most cells express distinctive 'cold shock' proteins. Glycine-rich cold-inducible RNA-binding proteins are thought to function as mRNA chaperones and play an important role in cold tolerance in land plants (Kim et al. 2005), as well as animals (Nishiyama et al. 1997). Two genes encoding glycine-rich RNA-binding proteins in *G.sulphuraria* (*Gasu_13830.1 and Gasu_37390.1*) showed a 20- to 30-fold upregulation in cold. Interestingly, glycine-rich RNA-binding proteins seem to lack in *Cyanidioschyzon merolae* and most other Rhodophyta. In both, plants and animals it has been shown that some heat shock proteins, which cells express in response to elevated temperatures, are also essential for acclimation to cold. In the data set analyzed here, all genes annotated as heat shock proteins, or chaperones, or having similar annotations, either did not show significant *T*-dependent expression changes or had higher transcript levels at higher temperatures.

T-dependent changes in the expression of secreted proteins

For DEG's differing between all warmer (42°C/46°C) and all colder temperatures (28°C) enrichment analyses were performed. These analyses showed that genes encoding proteins containing a predicted secretory signal peptide were 2.0-fold enriched ($P = 2.8 \times 10^{-5}$) among genes with higher expression levels at high temperatures, and 1.9-fold depleted ($P = 5.3 \times 10^{-7}$) among genes with higher expression levels at lower temperature. This indicates that expression levels of putatively excreted proteins are downregulated at 28°C. To further analyze expression levels of genes encoding proteins with a secretory signal peptide, these genes were binned by Δ_{Max}log$_2$(FPKM) or P_{Min}(EdgeR), which served as proxy for magnitude or statistical significance of expression differences, respectively. This binning showed (Supplementary Fig. S6) that genes encoding proteins with a secretory signal peptide were increasingly enriched among genes displaying larger or more significant expression differences. The five genes with the highest *T*-dependent change in expression (*Gasu_54130.1, Gasu_54440.1, Gasu_10210.1, Gasu_06850.1 and Gasu_31500.1*; 4670- to 307-fold) all five encode a secretory signal and had their highest frequency at 46°C. *Galdieria sulphuraria* is predicted to have 260 proteins with an N-terminal secretory peptide. Many of those proteins are likely to be part of the cell wall. Some of the genes encoding these 260 proteins were among those with the highest expression levels at high temperatures (see above), and among those showing the largest expression decreases upon shift to lower temperatures. It seems the expression of excreted proteins, which are probably part of the cell wall decreases significantly at 28°C.

T-dependent expression changes of genes realizing C-metabolism

A decrease in temperature resulted in pronounced changes in frequency of transcripts realizing metabolic pathways in *G.sulphuraria*. After 12 h and 48 h at 28°C genes annotated as enzymes were 1.5-fold ($P = 0.002$) and 1.8-fold ($P = 7.4 \times 10^{-14}$) enriched among upregulated DEG's, respectively. Transcripts involved in amino acid biosynthesis showed little enrichment among upregulated DEG's after 3 h at 28°C. Yet after 12 h, pathway analysis (KEGG map 01230) and enrichment analysis (MAPMAN Bin 13) indicated a 2-fold enrichment ($P = 0.006$ and $P = 0.011$, respectively) of amino acid biosynthesis. Further pathway analyses showed that metabolism of a variety of amino acids seems upregulated after 12 h at 28°C (Supplementary Table S1). Upon temperature elevation to 46°C, amino acid biosynthesis seemed to be generally downregulated, again. The upregulation of amino acid biosynthesis at 28°C, as indicated by gene expression frequencies, could either allow for increased protein biosynthesis, especially ribosomal proteins (see above), or might drive the biosynthesis of betaine (see below), or both.

Measurements of photosynthetic oxygen evolution at different temperatures indicated cessation of photosynthesis at 28°C (**Fig. 2B**). Compared to this, the time course of expression levels of genes encoding the photosynthetic machinery (**Fig. 2A**) did not show pronounced, rapid *T*-dependent changes. This lack of correlation between expression of genes encoding the photosynthetic machinery and photosynthetic activity indicates posttranscriptional regulation of photosynthesis. The rapid temperature decrease to 28°C probably resulted in an inhibition of the light-dependent processes of photosynthesis.

When inspecting a hierarchical cluster analysis (Supplementary Fig. S4), we noticed that glycogen-degrading enzymes, such as 4-alpha-glucanotransferase (Gasu_34050.1), were upregulated upon

Plant Cell Physiol. 0(0): 1–11 (2019) doi:10.1093/pcp/pcy240

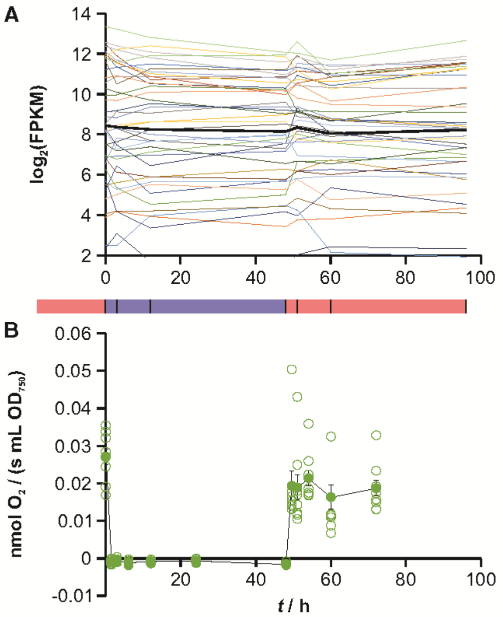


Fig. 2 *T*-dependent regulation of photosynthesis. (A) Expression profiles of genes involved in photosynthesis (MAPMAN Bin: 1) show no common, and mostly small changes in expression. The heavier, black trace represents the average of all 43 expression levels. (B) Photosynthetic O_2 evolution is inhibited after transfer to 28°C and quickly resumes after transfer to 46°C. For each time point (immediately before and 1.5, 3, 6, 12, 24 and 48 h after *T* change) eight measurements (two measurements each on four different samples; open symbols) were performed (closed symbols average ± SE). The colored bar in the middle indicates the temperature protocol (light red for 42°C/46°C, light blue for 28°C) and time points when samples for RNA isolation were taken (vertical black lines).

temperature decrease, while enzymes catalyzing glycogen biosynthesis and gluconeogenesis, such as glycogenin (Gasu_06400.1) and pyruvate, water dikinase (Gasu_42070.1), displayed opposite *T*-dependence. The genes encoding glycogenin and pyruvate, water dikinase were among those displaying the highest increase in expression 12 h after temperature increase from 28°C to 46°C, 29- and 19-fold, respectively. To investigate the *T*-dependence of central carbon metabolism, all genes encoding enzymes of the central carbon metabolism that displayed significant changes in transcript frequencies 12 h after temperature change were extracted (Supplementary Table S2). Most of the extracted genes encode enzymes catalyzing largely irreversible reactions, which are known to promote either catabolic (glycogen degradation and glycolysis) or anabolic (gluconeogenesis and sucrose or glycogen synthesis) carbon metabolism. Pronounced *T*-dependent transcriptional control of central carbon metabolism was observed at two points, (i) glycogen metabolism and (ii) the transition between glycolysis and mitochondrial pyruvate oxidation (Supplementary Table S2). After temperature decrease, enzymes catalyzing glycogen degradation, such as glycogen phosphorylase, had increased transcript frequencies. After temperature increase enzymes catalyzing glycogen biosynthesis, such as glycogenin, had increased transcript frequencies. After temperature decrease, enzymes feeding phosphoenolpyruvate (PEP) into mitochondrial pyruvate oxidation, such as pyruvate kinase and PEP carboxylase, had increased transcript frequencies, while pyruvate dehydrogenase kinase, which limits mitochondrial pyruvate oxidation, and pyruvate, water dikinase, which feeds pyruvate into gluconeogenesis, had decreased transcript frequencies. After transferring cells back to higher temperatures, transcript frequencies displayed opposite changes, probably limiting mitochondrial pyruvate oxidation and feeding pyruvate into gluconeogenesis, instead (Supplementary Table S2). These results indicated a predominance of glycogen degradation and glycolysis at low temperatures, and a predominance of gluconeogenesis followed by glycogen synthesis, at higher temperatures.

When inspecting a hierarchical cluster analysis (Supplementary Fig. S4), we also noticed that transcripts for enzymes catalyzing fatty acid and membrane lipid biosynthesis (anabolism) increased upon temperature decrease (42°C → 28°C), while transcripts for enzymes catalyzing membrane lipid and fatty acid degradation (catabolism) decreased in frequency. Temperature increase (28°C → 46°C) reversed these transcript frequency changes. To investigate the *T*-dependence of lipid metabolism, genes encoding enzymes of lipid metabolism and displaying significant changes in transcript frequencies 12 h after temperature change were extracted (Supplementary Table S3). Increased fatty acid biosynthesis at lower temperatures seemed to drive membrane lipid biosynthesis and not storage lipid biosynthesis. A 2.5-fold increase in transcript frequency of triacylglycerol lipase (Gasu_17090.1; EC 3.1.1.3), at 28°C, seemed to indicate an upregulation of storage lipid degradation. In higher plants, adaptation to lower temperatures is related to a larger fraction of unsaturated fatty acids, and an increasing fraction of unsaturated fatty acids at lower temperatures has been show for *Galdieria* as well (Nagashima 1994). Pathway analysis indicated a 4.9-fold enrichment (P = 0.0094) of transcripts required for biosynthesis of unsaturated fatty acids (KEGG map 01040) among DEG's upregulated within 12 h after temperature decrease. Four genes annotated as desaturases showed significantly increased transcript frequencies at 28°C (Supplementary Table S3). In addition to lipid desaturation, in *G.sulphuraria*, a decrease in temperature seemed to result in increased fatty acid and membrane lipid biosynthesis, possibly increasing the area of internal membrane systems. Upon temperature increase from 28°C to 46°C, transcripts for fatty acid and membrane lipid biosynthesis were downregulated within 3 h, and fatty acid degradation seemed upregulated (Supplementary Table S3).

Genes encoding all three steps of the folate cycle and the key enzyme of the S-adenosylmethionine (SAM) cycle were located together in a cluster of 89 genes (Supplementary Fig. S7). Transcripts from genes in this cluster displayed an increase in expression at 28°C during the first 12 h, and a rapid, lasting decrease in expression at 46°C (**Fig. 3**). Genes encoding enzymes for different metabolic pathways were enriched in this cluster of 89 genes. Pathway analysis indicated a 6.0-fold

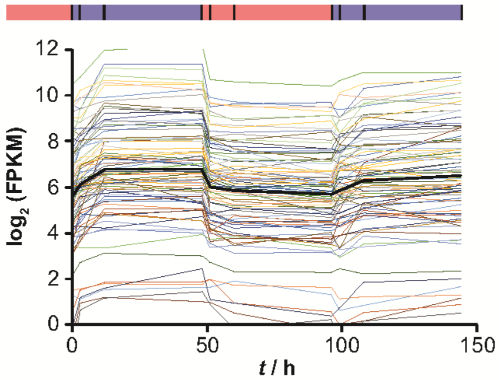


Fig. 3 Cold-induced upregulation of genes involved in different metabolic pathways. Time course of $\log_2$(FPKM) values (biological replicates averaged) from 89 genes forming a cluster (compare Supplementary Figs. S4, S7) that contains 34 enzymes catalyzing one-carbon metabolism, polyamine biosynthesis, fatty acid biosynthesis and amino acid biosynthesis. The heavier, black trace represents the average of all 89 expression levels. The colored bar on top indicates the temperature protocol (light red for 42°C and 46°C, light blue for 28°C) and time points when samples were taken (vertical black lines).

enrichment ($P = 5.9 \times 10^{-4}$) of 'Biosynthesis of amino acids' (KEGG map 01230; compare above), a 7.4-fold enrichment ($P = 0.04$) of 'Glycerophospholipid metabolism' (KEGG map 00564), and a 13.5-fold enrichment ($P = 0.04$) of 'One carbon pool by folate' (KEGG map 00670). The genes encoding the three steps of the folate cycle (**Fig. 4**), displayed 2.7- to 9.6-fold higher transcript frequency at 28°C (12 h) compared to 42°C. Two of three paralogs of SAM synthetase, the key enzyme of the SAM cycle, displayed 3.4- and 3.8-fold higher transcript frequencies at 28°C. All enzymes catalyzing folate and SAM cycle (**Fig. 4**) had elevated transcript frequencies at lower temperatures. Other genes from the same cluster indicate an upregulation of metabolic and transport processes connected to SAM and folate cycle (Supplementary Figs. S7, S8). Transcript levels for adenosine kinase and for an ATP:ADP antiporter were 4.6-fold and 2.9-fold elevated at 28°C (12 h), respectively. Adenosine kinase phosphorylates adenosine into AMP, which is phosphorylated to ADP by adenylate kinase, showing increased transcription levels at 28°C, as well (Supplementary Figs. S7, S8). At lower temperatures, regeneration of ADP from adenosine and transmembrane ATP: ADP exchange seemed to be upregulated. Three transcripts for SAM transporters, also from the same cluster, with 2.1 to 3.5-fold higher transcript frequencies at 28°C indicated that not only SAM biosynthesis but also transmembrane SAM transport was upregulated. Biosynthesis of serine from glyceraldehyde-3-phosphate (GAP) to provide one-carbon units for folate and SAM cycle, seemed upregulated at 28°C (Supplementary Figs. S7, S8). Glycine decarboxylase, which feeds methylene-tetrahydrofolate into the folate cycle was 1.7-fold upregulated at 28°C. In contrast, transcript frequency for formate-tetrahydrofolate ligase, which would drain the folate cycle, was reduced to 0.3-fold at 28°C (Supplementary Figs. S7, S8). These changes in transcript levels all point to an upregulation of SAM biosynthesis. But, what is this SAM used for?

The transcript with the largest increase in frequency, 13-fold (28°C, 12 h), in this cluster (Supplementary Fig. S7) encodes sarcosine dimethylglycine methyltransferase (SDMT, Gasu_06500.1). SDMT uses SAM to methylate sarcosine (methylglycine) into dimethylglycine and dimethylglycine into betaine (**Fig. 4**). Among 122 genes in the *G.sulphuraria* genome encoding proteins that are annotated as SAM-dependent or as interacting with SAM, no other displayed comparably large *T*-dependent transcription changes. Three paralogous SDMT genes in *G.sulphuraria* likely originated from a single horizontal gene transfer event (Schönknecht et al. 2013), followed by gene duplications. Transcripts for the other two SDMTs (Gasu_07580.1 and Gasu_07590.1) had 1.4- and 3.9-fold increased frequencies at 28°C (12 h) compared to 42°C. SDMT homologs have not be detected in any other eukaryotic organism, and enzymatic activity, i.e. SAM-dependent methylation of sarcosine and dimethylglycine, has been demonstrated experimentally for one SDMT (Gasu_07580.1) from *G.sulphuraria* (McCoy et al. 2009). Based on the transcript changes summarized in **Figs. 3**, **4**, we postulate that at lower temperatures folate cycle and SAM cycle are upregulated to provide methyl residues for betaine biosynthesis by SDMT, and triose phosphates are metabolized into serine to drive the folate cycle and provide glycine for betaine biosynthesis. It has been shown experimentally that *G.sulphuraria* does accumulate betaine under salt stress (Schönknecht et al. 2013; Fig. S15). Four transcripts in the cluster of 89 genes (Supplementary Fig. 7) indicate that polyamine biosynthesis might be upregulated at 28°C, which would also consume SAM. In the same cluster is Gasu_37790.1, encoding glutamine synthetase (EC 6.3.1.2), the enzyme that assimilates ammonia into amino acids. Elevated ammonia assimilation probably provides the organic N required for amino acid biosynthesis (see above).

It seems that increased SAM biosynthesis at lower temperatures drives betaine biosynthesis (**Fig. 4**), increasing the demand for amino acids. Betaines and polyamines have been implicated in tolerating cold stress and other stresses (Sakamoto and Murata 2002). Obviously, cold stress is relative! For an organism such as *G.sulphuraria*, which has its growth optimum at 45°C (Doemel and Brock 1971), a growth temperature of 28°C is likely to cause cold stress. Our results indicate that *G.sulphuraria* acclimates to this cold stress by synthesis of betaines, comparable to the biochemical acclimation of some non-thermophilic organisms at lower temperatures.

Summarizing, the following *T*-dependent changes in metabolic fluxes are indicated by the observed expression changes: after transfer to 28°C, storage polysaccharides in the form of glycogen (Shimonaga et al. 2008) are broken down, and the resulting hexose phosphates are metabolized in the glycolytic pathway to produce PEP for the citric acid cycle (Supplementary Table S2). Intermediates of glycolytic pathway and citric acid cycle provide carbon skeletons for increased amino acid biosynthesis (Supplementary Table S1). Amino acids are used for the increased synthesis of ribosomal proteins

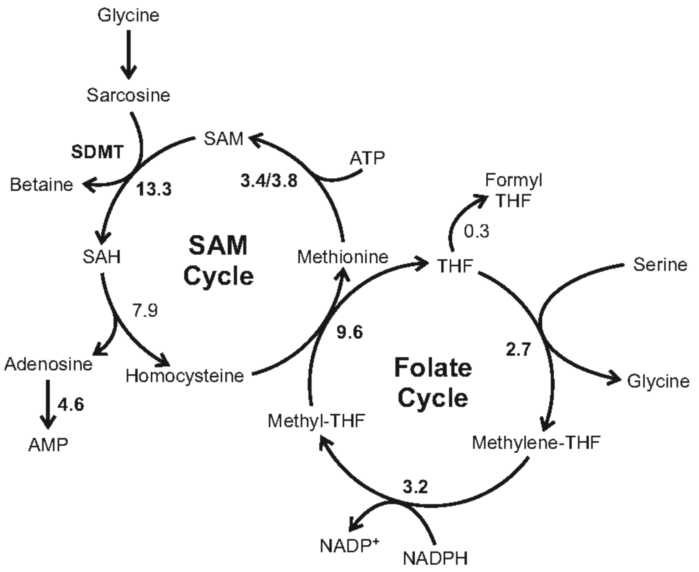


Fig. 4 One-carbon metabolism is upregulated at 28°C. Genes encoding enzymes catalyzing one-carbon metabolism, i.e. S-adenosylmethionine (SAM) cycle and folate cycle, have increased expression levels at 28°C. For enzyme names, EC numbers and gene names see Supplementary Fig. S8. Numbers give fold increase in transcript frequency at 28°C (12 h), compared to 42°C, with bold face indicating transcripts that cluster together. SAH, S-adenosylhomocysteine; SAM, S-adenosylmethionine; SDMT, sarcosine dimethylglycine methyltransferase; THF, tetrahydrofolate.

(**Fig. 1**) and to drive C1 metabolism, folate cycle and SAM cycle (**Fig. 4**). SAM and amino acids feed the metabolic reactions producing betaine, which has been shown to facilitate acclimation to stress in *G.sulphuraria*. After increasing the temperature to 46°C, pyruvate and triose phosphates from photosynthesis seem to enter gluconeogenesis producing hexose phosphates, which are metabolized into glycogen (Supplementary Table S2).

Genes acquired via horizontal gene transfer display *T*-dependent changes in expression

FPKM (fragments per kilobase of exon per million reads mapped) values displayed a log-Normal distribution, as expected (Supplementary Fig. S9A), and the distribution of FPKM values and the total number of reads were comparable for all 25 samples. A comparison of FPKM values from HGT candidates with the entire transcriptome indicated an almost 4-fold lower expression level of HGT candidates (Supplementary Fig. S9A). Yet a further analysis indicated that this low expression level reflects low expression of Archaeal ATPases (217 Archaeal ATPases out of 337 HGT candidates), while the remaining HGT candidates (120) displayed expression levels as expected for genes encoding transporters or enzymes (Supplementary Fig. S9B)—the two functional groups dominating HGT candidates (Schönknecht et al. 2013).

Most clusters showing pronounced *T*-dependent changes in gene expression contained one or more HGT candidates (marked in orange in Supplementary tables and figures). Some HGT candidates, such as SDMT (**Fig. 4**; see above) displayed relatively large *T*-dependent expression changes. We therefore tested whether HGT candidates were enriched among DEG's. This seemed to be the case. Genes were binned by $\Delta_{\mathrm{Max}}\log_2$(FPKM) or P_{Min}(EdgeR), which served as proxy for magnitude or statistical significance of expression differences, respectively (**Fig. 5**). There was an increase of the percentage of HGT candidates towards larger or more statistically significant gene expression changes, which did revert at the largest $\Delta_{\mathrm{Max}}\log_2$(FPKM) values and smallest *P*-values. HGT candidates were enriched close to 2-fold among genes displaying relatively large (4–8-fold) or statistically more significant *T*-dependent changes in gene expression. This enrichment (**Fig. 5**) indicates that HGT candidates may play a role in cold acclimation in *G.sulphuraria*. A good example are the three SDMT's, which had been interpreted to improve salt tolerance in *G.sulphuraria* (Schönknecht et al. 2013), and which seem to play a role in cold acclimation (**Figs. 3, 4**).

The two largest protein families in *G.sulphuraria* did probably originate from a single horizontal gene transfer followed by multiple rounds of gene duplications (Schönknecht et al. 2013). These proteins show similarity to Archaeal ATPases, which belong to the large family of STAND ATPases, soluble ATPases, which seem to be involved in stabilization or modification of other macromolecules (Leipe et al. 2004). Due to higher gene copy numbers in (hyper)thermophilic bacteria and archaea it has been speculated that Archaeal ATPases might be involved in adaptation to high temperatures (Schönknecht et al. 2013). Our RNA-Seq data indicate that genes encoding these putative

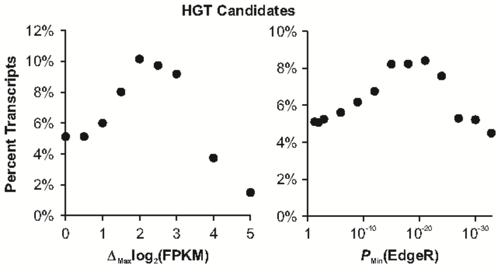


Fig. 5 HGT candidates are enriched among genes displaying larger or more significant expression differences. Genes were binned by either $\Delta_{Max}\log_2$(FPKM) (left) or P_{Min}(EdgeR) values (right), which were taken as a proxy how large in magnitude or how statistically significant expression differences are. For each bin the percentage of HGT candidates was calculated. There is a close to 2-fold increase in the percentage of HGT candidates towards larger or statistically better-supported differences in expression levels.

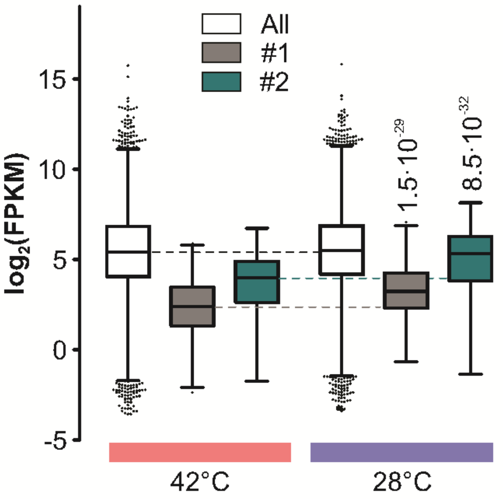


Fig. 6 Genes encoding Archaeal ATPases are expressed, and expression increases at lower temperatures. Box plots summarize $\log_2$(FPKM) values for all genes (All, white), and for genes encoding Archaeal ATPases of family #1 (#1, grey), or family #2 (#2, dark green) at 42°C (left, light red bar) and at 3 h 28°C (right, light blue bar, compare Supplementary Fig. S10 for complete time courses). Whiskers give 1 to 99 percentile intervals. Horizontal dashes lines connect median values for each of the three groups of genes (All, #1 and #2), and values above to right boxes give *P*-values (paired *t*-test) comparing $\log_2$(FPKM) values at 42°C and 28°C for Archaeal ATPases of family #1 or #2.

Archaeal ATPases in *G.sulphuraria* are expressed. The expression levels of most Archaeal ATPase genes were relatively low, on average about 4-fold lower than the median expression level of $\log_2$(FPKM) = 5.4 for all transcripts (Supplementary Fig. S9B). Unexpectedly, expression levels for Archaeal ATPases did increase at lower temperatures (**Fig. 6**). While the expression changes were only about 2-fold (on average 1.8-fold and 2.5-fold for family #1 and #2, respectively), they were very consistent and statistically significant ($P = 1.5 \times 10^{-29}$ and $P = 8.5 \times 10^{-32}$ for family #1 and #2, respectively). Comparing the time courses for family #1 and #2 of the Archaeal ATPases, family #1 showed more of a continuous decrease or increase in expression during the cold or warm periods, respectively, which in family #2 was superimposed by rapid increase or decrease in expression during the first 3 h of cold or warm periods, respectively (Supplementary Fig. S10). In a hierarchical cluster analysis of all transcripts (Supplementary Fig. S4), most family #1 members (89 of 131, plus 20 family #2 members) grouped together in one cluster of 494 genes, and many family #2 members (47 out of 96, plus 1 family #1 member) grouped together in another cluster of 143 genes. Enrichment analyses were performed for both clusters to obtain insight, which functional categories dominated these clusters. For the larger cluster, containing most family #1 members, several MAPMAN categories, such as 'protein', 'protein.degradation.subtilases', 'amino acid metabolism', or 'DNA binding' were depleted, while the protein family of peptidyl-prolyl *cis-trans* isomerase B was enriched (P = 3.1%). Peptidyl-prolyl *cis-trans* isomerases catalyze the *cis-trans* isomerization of proline imidic peptide bonds and regulate protein folding. It is tempting to speculate that Archaeal ATPases might be involved in regulating or stabilizing protein structure, as other STAND ATPases do.

Genes that most likely originated by horizontal gene transfer display above average *T*-dependent expression changes (**Fig. 5**), indicating a possible function in cold acclimation. Enzymes, such as SDMT (**Fig. 4**), probably catalyze the biosynthesis of metabolites promoting cold acclimation. Archaeal ATPases may stabilize proteins, or other macromolecules under cold. In summary, it seems that HGT candidates might have acquired new or additional functions that contribute to cold acclimation. For an Arctic *Chlamydomonas* species is was recently shown that improved freezing tolerance was acquired by HGT of a heat shock protein from a psychrophilic bacterium (Liu et al. 2018). These findings indicate that adaptation of (unicellular) eukaryotic organisms to environmental stress has probably been facilitated by HGT.

Materials and Methods

Cultivation of *G.sulphuraria*

Three replicate cultures of *G.sulphuraria* strain 074 W were pre-grown at photoautotrophic conditions in 2x Allen Medium (Allen 1959) at 42°C and constant illumination of 90 µmol photons m^{-2} s^{-1}, for two weeks. The cultures were shifted to 28°C for 48 h, to 46°C for 48 h, and to back 28°C for another 48 h. Immediately before each temperature change, and 3 h and 12 h after each temperature change samples were taken (Supplementary Fig. S1). Growth parameters were recorded by determining the optical density at λ = 750 nm (OD_{750}).

RNA isolation and sequencing

For RNA isolation, 10 mL samples were centrifuged for 5 min at 3,500 rpm and 4°C. RNA was isolated from cell pellets using the acid guanidinium thiocyanate-phenol-chloroform extraction method (Chomczynski and Sacchi 2006). To remove DNA, the isolates were treated with RNase-free DNase I (New England Biolabs, Ipswich, USA) for 30 min.

2 × 100 bp paired end RNA-Seq libraries were prepared following the 'Illumina TruSeq RNA Sample Prep v2 LS Protocol' (Illumina, San Diego, USA). All reagents were scaled by 2/3 to the volume proposed in the protocol. RNA integrity and the quality of all sequencing libraries was assessed using a 2100 Bioanalyzer (Agilent Technologies, Santa Clara, USA). Libraries were equimolarly pooled (2 nM) and sequenced with the Illumina HiSeq2000 platform at the Genomics and Transcriptomics Laboratory of the Biologisch-Medizinisches Forschungszentrum in Düsseldorf, Germany. A total of 25 samples was analyzed, biological duplicates for time points *t*0, *t*4, *t*5, *t*6 and *t*9, and triplicates for time points *t*1, *t*2, *t*3, *t*4, *t*7 and *t*8 (Supplementary Fig. S1).

Data analysis

The 2 × 100 bp paired end reads were mapped to the reference genome sequence of *G.sulphuraria* 074 W, available for download at NCBI (GCF_000341285.1). Read mapping and gene expression estimates were performed using RSEM (Li and Dewey 2011) running with the 'very sensitive' option. Differential gene expression was calculated with EdgeR (Robinson et al. 2010) implementing the quasi-likelihood F-test (QLF-test) functions in order to address the uncertainty in estimating the dispersion for each gene (Lun et al. 2016) (Supplementary Table S4). A significance threshold of 0.01 was applied after the *P*-value was adjusted with FDR via Benjamini–Hochberg correction (Benjamini and Hochberg 1995).

Specific sets of DEG were calculated comparing (i) all samples taken at warmer temperatures (42°C or 46°C) against all samples taken at 28°C, as well as (ii) each time point in relation to first time point (e.g. *t*0 vs. *t*1, *t*0 vs. *t*2, *t*0 vs. *t*3, etc.; Supplementary Table S4) and (iii) all consecutive time points along the temperature shift timeline (e.g. *t*0 vs. *t*1, *t*1 vs. *t*2, *t*2 vs. *t*3, etc.). As proxy for an overall expression change of each gene, we calculated $\Delta_{Max}\log_2$(FPKM) values as the difference between the two time points with the highest and the lowest average $\log_2$(FPKM) value. As a proxy for which genes show the 'most significant' expression difference, we extracted the lowest *P*-value for each gene from EdgeR analyses, P_{Min}(EdgeR), comparing each sample to each other sample (for all 25 samples). Since $\Delta_{Max}\log_2$(FPKM) and P_{Min}(EdgeR) values are sensitive for outliers, these values were not used to rank genes, but were only used to analyze whether certain categories of genes (such as HGT candidates) were enriched among sets of genes with increasing magnitude or significance of expression difference.

Gene set enrichment analyses were performed with Gene Ontology terms (The Gene Ontology Consortium 2015), MapMan categories (Thimm et al. 2004) based on the Mercator Automated Sequence Annotation Pipeline (http://www.plabipd.de/portal/mercator-sequence-annotation), and *G.sulphuraria* protein families (Schönknecht et al. 2013). Metabolic pathway analysis was performed using KOBAS 2 (http://kobas.cbi.pku.edu.cn/) (Xie et al. 2011). *P*-values for gene set enrichment analyses and metabolic pathway analyses were calculated with Fisher's exact test. *P*-values from multiple comparisons were corrected by the Benjamini–Hochberg procedure.

Hierarchical cluster analysis was performed on both the entire transcriptome (Supplementary Fig. S4) and enriched gene sets with MeV 4.9.0 (Howe et al. 2011). SignalP4 (Petersen et al. 2011) was used to detected N-terminal signal peptides, indicating excretion of the gene product. TargetP (Emanuelsson et al. 2007) was used to estimate the compartmental localization of gene products.

Data deposit

The complete RNA-seq data set is provided in Supplementary Table S4. The RNA-seq data reported in this article have been submitted to the National Center for Biotechnology Information Gene Expression Omnibus with accession number GSE118890.

Photosynthetic oxygen evolution

Photosynthetic oxygen generation was measured after 1.5 h, 3 h, 6 h, 12 h, 24 h and 48 h at 42°C and 28°C during the shift experiment using a Clark-type electrode. One milliliter samples were taken directly from the shaking batch cultures and injected into the measuring cuvette containing a stir bar to keep *G.sulphuraria* in suspension and provide homogeneous gas levels. In total, eight samples for each of the 10 time points were taken and analyzed. The measuring cuvette was connected to a heating device to keep sample temperature constant at 42°C/28°C during measurements.

Supplementary Data

Supplementary data are available at PCP online.

Funding

Deutsche Forschungsgemeinschaft in the framework of the Research Group FOR 1261–2 (WE2231/11–1 to A.P.M.W.); and the Excellence Cluster CEPLAS (EXC 1028 to A.P.M.W.); and by the Deutscher Akademischer Austauschdienst (German Academic Exchange Service, 57063099 to G.S. and H.J.L.). This material is based upon work supported by (while G.S. was serving at) the National Science Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Acknowledgments

We thank Dr. Andrea Bräutigam (Bielefeld, Germany) for comments on the manuscript.

Disclosures

The authors have no conflicts of interest to declare.

References

Allen, M.B. (1959) Studies with *Cyanidium caldarium*, an anomalously pigmented chlorophyte. *Archiv. Mikrobiol.* 32: 270–277.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389–3402.

Bailey, R.W. and Staehelin, L.A. (1968) The chemical composition of isolated cell walls of *Cyanidium caldarium*. *J. Gen. Microbiol.* 54: 269–276.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* 57: 289–300.

Biegert, A., Mayer, C., Remmert, M., Söding, J. and Lupas, A.N. (2006) The MPI Bioinformatics Toolkit for protein sequence analysis. *Nucleic Acids Res.* 34: W335–W339.

Buhmann, M.T., Poulsen, N., Klemm, J., Kennedy, M.R., Sherrill, C.D. and Kröger, N. (2014) A tyrosine-rich cell surface protein in the diatom *Amphora coffeaeformis* identified through transcriptome analysis and genetic transformation. *PLoS One* 9: e110369.

Chomczynski, P. and Sacchi, N. (2006) The single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction: twenty-something years on. *Nature Protocols*. 1: 581–585.

Consortium, T.G.O. (2015) Gene Ontology Consortium: going forward. *Nucleic Acids Res.* 43: D1049–D1056.

Doemel, W.N. and Brock, T.D. (1970) Upper temperature limit of *Cyanidium caldarium*. *Arch. Mikrobiol.* 72: 326–332.

Doemel, W.N. and Brock, T.D. (1971) The physiological ecology of *Cyanidium caldarium*. *J. Gen. Microbiol.* 67: 17–32.

Emanuelsson, O., Brunak, S., von Heijne, G. and Nielsen, H. (2007) Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protoc.* 2: 953–971.

Fukuda, S., Yamamoto, R., Iwamoto, K. and Minoda, A. (2018) Cellular accumulation of cesium in the unicellular red alga *Galdieria sulphuraria* under mixotrophic conditions. *J. Appl. Phycol.*

Graziani, G., Schiavo, S., Nicolai, M.A., Fogliano, V., Pollio, A. and Pinto, G. (2012) Microalgae as human food: chemical and nutritional characteristics of the thermo-acidophilic microalga *Galdieria sulphuraria*. *Food Funct* 4: 144–152.

Gross, W. (1999) Revision of comparative traits for the acido- and thermophilic red algae *Cyanidium* and *Galdieria*. *In* Enigmatic Microorganisms and Life in Extreme Environments. Edited by Seckbach, J. pp. 439–446. Kluwer, Dordrecht, The Netherlands.

Gross, W., Oesterhelt, C., Tischendorf, G. and Lederer, F. (2002) Characterization of a non-thermophilic strain of the red algal genus *Galdieria* isolated from Soos (Czech Republic). *Eur. J. Phycol.* 37: 477–482.

Gross, W. and Schnarrenberger, C. (1995) Heterotrophic growth of two strains of the acido-thermophilic red alga *Galdieria sulphuraria*. *Plant Cell Physiol.* 36: 633–638.

Henkanatte-Gedera, S.M., Selvaratnam, T., Karbakhshravari, M., Myint, M., Nirmalakhandan, N., Van Voorhies, W., et al. (2017) Removal of dissolved organic carbon and nutrients from urban wastewaters by *Galdieria sulphuraria*: laboratory to field scale demonstration. *Algal Res.* 24: 450–456.

Howe, E.A., Sinha, R., Schlauch, D. and Quackenbush, J. (2011) RNA-Seq analysis in MeV. *Bioinformatics* 27: 3209–3210.

Jones, P.G. and Inouye, M. (1994) The cold-shock response—a hot topic. *Mol. Microbiol.* 11: 811–818.

Kim, Y.-O., Kim, J.S. and Kang, H. (2005) Cold-inducible zinc finger-containing glycine-rich RNA-binding protein contributes to the enhancement of freezing tolerance in *Arabidopsis thaliana*. *Plant J.* 42: 890–900.

Koonin, E.V., Makarova, K.S. and Aravind, L. (2001) Horizontal gene transfer in prokaryotes: quantification and classification. *Annu. Rev. Microbiol.* 55: 709–742.

Külahoglu, C., Denton, A.K., Sommer, M., Maß, J., Schliesky, S., Wrobel, T.J., et al. (2014) Comparative transcriptome atlases reveal altered gene expression modules between two Cleomaceae C3 and C4 plant species. *Plant Cell* 26: 3243–3260.

Leger, M.M., Eme, L., Stairs, C.W. and Roger, A.J. (2018) Demystifying eukaryote lateral gene transfer (Response to Martin 2017 DOI: 10.1002/bies.201700115). *BioEssays* 40: 1700242.

Leipe, D.D., Koonin, E.V. and Aravind, L. (2004) STAND, a class of P-loop NTPases including animal and plant regulators of programmed cell death: multiple, complex domain architectures, unusual phyletic patterns, and evolution by horizontal gene transfer. *J. Mol. Biol.* 343: 1–28.

Li, B. and Dewey, C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12: 1–16.

Liu, C., Zhao, X. and Wang, X. (2018) Identification and characterization of the psychrophilic bacterium CidnaK gene in the Antarctic *Chlamydomoas* sp. ICE-L under freezing conditions. *J. Appl. Phycol.*

Lun, A.T.L., Chen, Y. and Smyth, G.K. (2016) It's DE-licious: a recipe for differential expression analyses of RNA-seq experiments using Quasi-Likelihood Methods in edgeR. *In* Statistical Genomics: Methods and Protocols. Edited by Mathé, E. and Davis, S. pp. 391–416. Springer New York, New York, NY.

McCoy, J.G., Bailey, L.J., Ng, Y.H., Bingman, C.A., Wrobel, R., Weber, A.P.M., et al. (2009) Discovery of sarcosine dimethylglycine methyltransferase from *Galdieria sulphuraria*. *Proteins Struct.*74: 368–377.

Nagashima, H. (1994) Natural products of the Cyanidiophyceae. *In* Evolutionary Pathways and Enigmatic Algae: Cyanidium Caldarium (Rhodophyta) and Related Cells. Edited by Seckbach, J. pp. 201–214. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Nishiyama, H., Itoh, K., Kaneko, Y., Kishishita, M., Yoshida, O. and Fujita, J. (1997) A glycine-rich RNA-binding protein mediating cold-inducible suppression of mammalian cell growth. *J. Cell Biol.* 137: 899–908.

Park, S., Lee, C.-M., Doherty, C.J., Gilmour, S.J., Kim, Y. and Thomashow, M.F. (2015) Regulation of the *Arabidopsis* CBF regulon by a complex low-temperature regulatory network. *Plant J.* 82: 193–207.

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

Qiu, H., Price, D.C., Yang, E.C., Yoon, H.S. and Bhattacharya, D. (2015) Evidence of ancient genome reduction in red algae (Rhodophyta). *J. Phycol.* 51: 624–636.

Qiu, H., Rossoni, A.W., Weber, A.P.M., Yoon, H.S. and Bhattacharya, D. (2018) Unexpected conservation of the RNA splicing apparatus in the highly streamlined genome of *Galdieria sulphuraria*. *BMC Evol. Biol.* 18: 41.

Reeb, V. and Bhattacharya, D. (2010) The thermo-acidophilic Cyanidiophyceae (Cyanidiales). In Red Algae in the Genomic Age. Edited by Seckbach, J. and Chapman, D.J. pp. 409–426. Springer Netherlands.

Richards, T.A. and Monier, A. (2016) A tale of two tardigrades. *Proc. Natl. Acad. Sci. USA*. 113: 4892–4894.

Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26: 139–140.

Sakamoto, A. and Murata, N. (2002) The role of glycine betaine in the protection of plants from stress: clues from transgenic plants. *Plant. Cell Environ.* 25: 163–171.

Schönknecht, G., Chen, W.-H., Ternes, C.M., Barbier, G.G., Shrestha, R.P., Stanke, M., et al. (2013) Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote. *Science* 339: 1207–1210.

Shimonaga, T., Konishi, M., Oyama, Y., Fujiwara, S., Satoh, A., Fujita, N., et al. (2008) Variation in storage a-glucans of the Porphyridiales (Rhodophyta). *Plant Cell Physiol.* 49: 103–116.

Silverman, H.G. and Roberto, F.F. (2007) Understanding marine mussel adhesion. *Mar. Biotechnol.* 9: 661–681.

Sloth, J.K., Jensen, H.C., Pleissner, D. and Eriksen, N.T. (2017) Growth and phycocyanin synthesis in the heterotrophic microalga *Galdieria sulphuraria* on substrates made of food waste from restaurants and bakeries. *Bioresour. Technol.* 238: 296–305.

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

Soucy, S.M., Huang, J. and Gogarten, J.P. (2015) Horizontal gene transfer: building the web of life. *Nat. Rev. Genet.* 16: 472–482.

Tai, S.L., Daran-Lapujade, P., Walsh, M.C., Pronk, J.T. and Daran, J.-M. (2007) Acclimation of *Saccharomyces cerevisiae* to low temperature: a chemostat-based transcriptome analysis. *Mol. Biol. Cell* 18: 5100–5112.

Thimm, O., Bläsing, O., Gibon, Y., Nagel, A., Meyer, S., Krüger, P., et al. (2004) MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J.* 37: 914–939.

Xie, C., Mao, X., Huang, J., Ding, Y., Wu, J., Dong, S., et al. (2011) KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Res.* 39: W316–W322.

Manuscript 4

Transcriptional response of the extremophile red alga *Cyanidioschyzon merolae* to changes in CO2 concentrations

Journal of Plant Physiology 217 (2017) 49–56

Contents lists available at ScienceDirect

Journal of Plant Physiology

journal homepage: www.elsevier.com/locate/jplph

Transcriptional response of the extremophile red alga *Cyanidioschyzon merolae* to changes in CO_2 concentrations[☆]

Nadine Rademacher[a,1], Thomas J. Wrobel[a,1], Alessandro W. Rossoni[a], Samantha Kurz[a], Andrea Bräutigam[b], Andreas P.M. Weber[a], Marion Eisenhut[a,*]

[a] *Institute of Plant Biochemistry, Cluster of Excellence on Plant Sciences (CEPLAS), Heinrich Heine University, Universitätsstraße 1, 40225 Düsseldorf, Germany*
[b] *Leibniz-Institut für Pflanzengenetik und Kulturpflanzenforschung (IPK), Corrensstraße 3, 06466 Stadt Seeland, OT Gatersleben, Germany*

ARTICLE INFO

Keywords:
Cyanidioschyzon merolae
RNA-seq
CO_2 limitation
Photorespiration

ABSTRACT

Cyanidioschyzon merolae (*C. merolae*) is an acidophilic red alga growing in a naturally low carbon dioxide (CO_2) environment. Although it uses a ribulose 1,5-bisphosphate carboxylase/oxygenase with high affinity for CO_2, the survival of *C. merolae* relies on functional photorespiratory metabolism. In this study, we quantified the transcriptomic response of *C. merolae* to changes in CO_2 conditions. We found distinct changes upon shifts between CO_2 conditions, such as a concerted up-regulation of photorespiratory genes and responses to carbon starvation. We used the transcriptome data set to explore a hypothetical CO_2 concentrating mechanism in *C. merolae*, based on the assumption that photorespiratory genes and possible candidate genes involved in a CO_2 concentrating mechanism are co-expressed. A putative bicarbonate transport protein and two α-carbonic anhydrases were identified, which showed enhanced transcript levels under reduced CO_2 conditions. Genes encoding enzymes of a PEPCK-type C_4 pathway were co-regulated with the photorespiratory gene cluster. We propose a model of a hypothetical low CO_2 compensation mechanism in *C. merolae* integrating these low CO_2-inducible components.

1. Introduction

Photosynthetic biomass production is initialized by the fixation of one molecule of CO_2 to the acceptor molecule ribulose 1,5-bisphosphate, catalyzed by the enzyme ribulose 1,5-bisphosphate carboxylase/oxygenase (Rubisco). The resulting two molecules of 3-phosphoglycerate (3-PGA) are fed into the Calvin-Benson-Bassham cycle (CCB) for reduction to carbohydrates and regeneration of the acceptor molecule. Rubisco catalyzes also an oxygenation reaction in which O_2 is added to the acceptor molecule, resulting in a proportion of Rubisco occupied with the non-productive side reaction yielding one molecule of 3-PGA and one molecule of 2-phosphoglycolate (2-PG). The latter inhibits multiple essential enzymes and must hence be efficiently detoxified. The photorespiratory pathway converts two molecules of 2-PG into one molecule of 3-PGA under consumption of energy and release of CO_2 and ammonia (reviewed in Bauwe et al., 2010; Hagemann et al., 2016). Two factors are critical for the rate of carboxylation versus oxygenation by Rubisco: The specificity of the enzyme for CO_2 and the ratio of $[CO_2]$ to $[O_2]$ at the site of Rubisco. Rubisco enzymes from organisms of various photosynthetic lineages have different evolutionary origins, and differ in substrate specificity and reaction velocity. Cyanobacteria and land plants use Rubisco Form 1A and 1B, which are of cyanobacterial origin, while non-green algae, such as red algae, contain Form 1C and 1D, which are of proteobacterial origin (Hauser et al., 2015). As a consequence of the reaction mechanism, higher specificity for CO_2 decreases Rubisco's velocity, while an increase in velocity is accompanied by a decline in specificity (Tcherkez et al., 2006). Cyanobacteria employ a fast Rubisco with low specificity for CO_2, while land plant Rubisco achieves a higher CO_2 specificity at the expense of velocity (Savir et al., 2010). Rubisco enzymes of thermophilic cyanidiophycean red algae, such as *Galdieria parta* and *Cyanidium caldarium*, exhibit the highest CO_2 specificity and thus lowest velocity, measured to date (Uemura et al., 1997).

In addition to the capacities of the enzyme, the $[CO_2]$ to $[O_2]$ ratio next to Rubisco strongly influences the carboxylation versus oxygenation rate. To overcome this constraint, many photosynthetic organisms evolved a CO_2 concentrating mechanism (CCM), which raises the CO_2 concentration in close vicinity to Rubisco and thereby enhances the carboxylation rate (Giordano et al., 2005; Raven et al., 2012). The occurrence of CCMs was demonstrated for cyanobacteria, most algae

Abbreviations: CA, carbonic anhydrase; CCM, CO_2 concentrating mechanism; HC, high CO_2; LC, low CO_2; RNA-seq, RNA-sequencing

[☆] This article is part of a special issue entitled: Light driven reactions in model algae published at the Journal of Plant Physiology 217C.

[*] Corresponding author: Postal address: Institute of Plant Biochemistry, Heinrich Heine University, Universitätsstraße 1, 40225 Düsseldorf, Germany.
E-mail address: m.eisenhut@hhu.de (M. Eisenhut).

[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.jplph.2017.06.014
Received 7 April 2017; Received in revised form 20 June 2017; Accepted 22 June 2017
Available online 04 July 2017

and aquatic plants, as well as for C_4 and Crassulacean acid metabolism plants (reviewed in Raven et al., 2008). The cyanobacterial CCM employs inorganic carbon (C_i) uptake mechanisms for cytoplasmic bicarbonate (HCO_3^-) accumulation and specific microcompartments, the carboxysomes, which encapsulate Rubisco and carbonic anhydrase (CA). The cytoplasmic HCO_3^- diffuses into the carboxysome and is converted into CO_2 by CA. Thus, the CO_2 concentration is increased by a factor of up to 1000 next to Rubisco (Badger and Price, 2003; Kaplan and Reinhold, 1999). However, even the action of a CCM cannot fully repress the oxygenase activity of Rubisco (Eisenhut et al., 2006, 2008; Nakamura et al., 2005; Zelitch et al., 2009). While among aquatic photosynthetic organisms the cyanobacterial and green algal CCM are well studied, it remains unclear whether thermophilic red algae also employ such a mechanism to improve photosynthetic efficiency (Giordano et al., 2005; Zenvirth et al., 1985).

Cyanidioschyzon merolae (*C. merolae*) is a model organism for cyanidiophycean red algae. The 16 Mbp genome of the single-cell organism is fully sequenced (Matsuzaki et al., 2004) and techniques for targeted gene knockout by homologous recombination as well as transient transformation for, *e.g.*, localization studies are available (Imamura et al., 2010; Minoda et al., 2004; Watanabe et al., 2011). *C. merolae* tolerates temperatures up to 57 °C and prefers acidic (pH < 2) growth medium (Seckbach, 1995). Under these conditions, CO_2 is the prevalent inorganic carbon (C_i) species in the aquatic environment. Although red algae use a Rubisco with high specificity for CO_2 over O_2, the reduced solubility of CO_2 at high temperatures forces *C. merolae* to perform a plant-like photorespiratory cycle (Rademacher et al., 2016).

In this work, we quantified the transcriptional response of *C. merolae* in response to changes in CO_2 concentrations by RNA-sequencing (RNA-seq) and applied the data set to predict a hypothetical CCM in *C. merolae* and to search for possible components.

2. Material and methods

2.1. *C. merolae* cultivation

C. merolae 10D wildtype (WT) cells were cultivated in 2x modified Allen's growth medium, pH 2 (Minoda et al., 2004), at 30 °C, bubbled with high CO_2 concentrations (5% CO_2 in air, HC) or low CO_2 concentrations (0.04% CO_2 in air, LC) at 90 µmol photons m^{-2} s^{-1} light in a Multi-Cultivator MC 1000-OD system (Photon Systems Instruments, Drasov, Czech Republic).

For the CO_2 shift experiment, three independent biological replicates of continuously HC grown *C. merolae* WT cells were cultivated for 24 h under HC conditions with an initial optical density at 750 nm (OD_{750}) of 0.7. After 24 h, cells were shifted by changing the CO_2 concentration in the aeration for 24 h to LC conditions and afterwards shifted back to HC conditions for a 24 h recovery phase. For RNA extraction, 5 mL samples (OD_{750} = 1.0) were taken immediately before the shift to LC (HC 0 h), 3 h after shift to LC (LC 3 h), 24 h after shift to LC (LC 24 h), and 24 h after a recovery phase at HC (HC 24 h). A sampling scheme is illustrated in Fig. 1A.

2.2. Gene expression analysis by RNA-seq

For RNA isolation, the 5 mL samples were centrifuged for 5 min at 4 °C (3000 rpm). RNA extraction from the cell pellet was performed using the EURx GeneMatrix Universal RNA Purification Kit (Roboklon, Berlin, Germany) following the manufacturer's protocol for RNA cell extraction. DNA was removed by treatment with RNase-free DNaseI (New England Biolabs, Ipswich, USA).

Libraries were prepared using the TruSeq RNA Sample Prep Kit v2 (Illumina, San Diego, USA). RNA integrity, sequencing library quality and fragment size were checked on a 2100 Bioanalyzer (Agilent). Average library size was 320 bp with equimolar pooling (2 nM). Demultiplexed Illumina reads were aligned with RSEM (parameters: −very-sensitive; −calc-pme; −calc-ci; −gibbs-burnin 500) with the default aligner bowtie2 (Li and Dewey, 2011) to the reference transcriptome of *C. merolae* 10D (ASM9120v1.30.gtf) (Nozaki et al., 2007), which was retrieved from the ENSEMBL database (Yates et al., 2016).

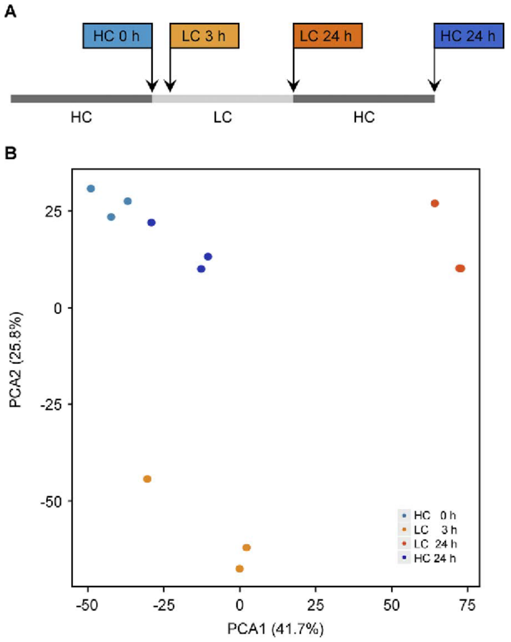


Fig. 1. Overview of the CO_2 shift experiment. A. Experimental set-up of the CO_2 shift experiment. *C. merolae* cells were cultivated under high CO_2 (5% CO_2 in air, HC) conditions, shifted for 24 h to low CO_2 (0.04% CO_2 in air, LC) conditions, and then shifted back to HC conditions. Samples for RNA-seq were taken immediately before the shift to LC (HC 0 h), 3 h (LC 3 h) and 24 h (LC 24 h) after the shift to LC, and 24 h after the recovery at HC (HC 24 h). The color code for the different samples applies for the complete study. B. Principle component analysis of RNA-seq data.

Differential gene expression was analyzed using the EdgeR package (McCarthy et al., 2012) in R. All sequenced conditions were analyzed in a pairwise manner with HC 0 h as reference point. A q-value of 0.01 was chosen as significance threshold for single gene differential expression after correction for multiple testing via the Bonferroni algorithm to limit the false positive rate to close to zero at the cost of a higher false negative rate (Krzywinski and Altman, 2014). For *K*-means clustering transcripts per million (TPM) values were scaled to their average. Sum of square errors were used to determine the suitable number of clusters at 10, *K*-means clustering with Euclidean distance carried out 10,000 times, and the clustering with the best SSE ratio used for further analysis. Principle component analysis was performed on scaled TPM values.

The MapMan-based functional categorization of all genes in the *C. merolae* genome was performed by comparing their protein sequence to *Arabidopsis* TAIR10 (http://www.arabidopsis.org/) using the standalone version of NCBI BLASTP (2.2.31+) with default settings. The MapMan categorization was transferred from TAIR10. Functional enrichment was performed on hierarchical, independent MapMan categories reduced to their first and second level using Fisher's exact tests. All *P*-values were corrected for multiple testing via the Benjamini Hochberg algorithm (Table S1) (Benjamini and Hochberg, 1995). GO terms for *C. merolae* proteins were retrieved from the Uniprot database and matched to gene identifiers using the ENSEMBL database (Yates et al., 2015). GO term enrichment was tested with the TopGO package (parameters; nodeSize = 10, ontology = "BP") (https://bioconductor.org/packages/release/bioc/html/topGO.html; Alexa and Rahnenfuhrer, 2016) in R. Statistically significant terms were calculated with classic Fisher's exact test and without weighing and corrected according to Benjamini-Yekutieli to account for dependency of GO terms

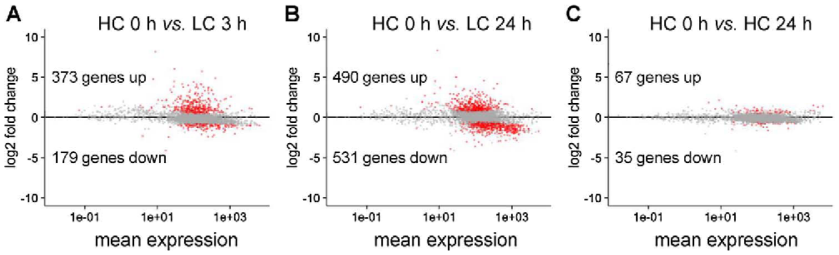


Fig. 2. Global transcriptional response of C. merolae toward changes in CO_2 concentrations. A. Short-term (3 h after shift from HC to LC conditions) effects of reduced CO_2 availability on gene expression. B. Long-term (24 h after shift from HC to LC conditions) effects of reduced CO_2 availability on gene expression. C. Recovery effects 24 h after re-shift from LC to HC conditions. Changes are given as log2-fold changes compared to HC 0 h. Significance was tested with EdgeR ($q < 0.01$; Robinson et al., 2010). Significantly changed genes are highlighted as red dots. Numbers of significantly up- and down-regulated genes are indicated.

(Benjamini and Yekutieli, 2001) (Table S2). Only significant enrichments ($q < 0.05$) are reported in the text. Transcription factors were annotated based on Pérez-Rodríguez et al. (2010). Heatmaps were created using the heatmap.2 package (https://cran.r-project.org/web/packages/gplots/gplots.pdf).

The complete RNA-seq data set is provided in Table S3. The read data have been submitted to the National Center for Biotechnology Information Gene Expression Omnibus under accession number GSE100372 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE100372).

2.3. *In silico analyses of carbonic anhydrases*

Amino acid sequences of CMT416C and CMI270C were retrieved from the *Cyanidioschyzon merolae* Genome Project data base (http://merolae.biol.s.u-tokyo.ac.jp) and the alignment performed with the ClustalW on the Phylogeny.fr platform (http://www.phylogeny.fr, Dereeper et al., 2008). The SOSUI platform (http://harrier.nagahama-i-bio.ac.jp/sosui/, Hirokawa et al., 1998) was employed to search for transmembrane regions. TargetP1.1 (http://www.cbs.dtu.dk/services/TargetP/, Emanuelsson et al., 2000) was used for prediction of hypothetical targeting peptides.

2.4. *Analysis of subcellular localization of carbonic anhydrases*

For subcellular localization of the carbonic anhydrases (CA) CMT416C and CMI270C in *Nicotiana benthamiana* (*N. benthamiana*) protoplasts and *C. merolae* cells, respectively, two constructs were generated for each protein, fusing CMT416C and CMI270C with the yellow fluorescent protein (YFP) either at the N- or C-terminus. The coding sequences were amplified by PCR using *C. merolae* genomic DNA as template and cloned into the pUBN-YFP or pUBC-YFP vector applying gateway technology (Grefen et al., 2010). In pUBN-YFP and pUBC-YFP expression of the fusion proteins is under the control of the *UBIQUTIN 10* promoter. Primer sequences are listed in Table S4. Transient transformation of *N. benthamiana* leaves was carried out using the *Agrobacterium tumefaciens* strain GV3101. Protoplast isolation and microscope analysis was performed 2 d after infiltration using a Zeiss LSM 510 Meta confocal −scanning laser microscope as described in Breuers et al., 2012. Transient transformation of *C. merolae* cells was performed as described by Ohnuma et al., 2008. Microscope analysis was perfomed 24 h after transformation with a Zeiss LSM 780 microscope.

3. Results

3.1. *Effects of reduced CO_2 concentrations on transcriptome*

To analyze transcriptional changes in response to altered CO_2 concentrations, we performed a CO_2 shift experiment. *C. merolae* cells were cultivated for 24 h under HC (5% CO_2 in air) conditions, then shifted for 24 h to LC conditions (0.04% CO_2 in air), and finally shifted back to HC conditions for another 24 h. Samples were taken in biological triplicates immediately before the shift to LC (HC 0 h), 3 h after the shift (LC 3 h), 24 h after the shift (LC 24 h), and 24 h after the recovery under HC conditions (HC 24 h). The experimental set-up is illustrated in Fig. 1A. RNA-seq analysis generated 726,326,434 Illumina paired-end reads with about 18.4 Million paired-end reads per sample on average. 92% of the RNA-reads mapped to the reference genome of *C. merolae* (Matsuzaki et al., 2004).

To evaluate reproducibility among biological replicates, a principle component analysis (PCA) was performed on the transcripts per million (TPM) values. The biological triplicates clearly clustered together (Fig. 1B), indicating lower variation between biological replicates compared to the treatment. Furthermore, the separation of the samples in the PCA indicated long-term and short-term LC effects as principle components 1 and 2, accounting for 42% and 26% of transcriptional variation, respectively (Fig. 1B).

Consistent with the PCA of all expression values, we found a larger number of genes significantly ($q < 0.01$, Bonferroni corrected) changed 24 h (long-term, 1021 genes in total, Fig. 2B, Table S3) after the shift from HC to LC conditions than 3 h (short-term, 552 genes in total, Fig. 2A, Table S3) after the shift. For 102 genes, the 24 h cultivation phase under HC conditions was not sufficient to fully recover the initial HC expression situation (Fig. 2C, Table S3) accounting for the difference between HC 0 h and HC 24 h in the PCA (Fig. 1B).

3.2. *Identification of CO_2-dependent gene expression patterns*

To search for CO_2-dependent gene expression patterns, we performed *K*-means clustering. As a result, 10 different clusters were generated (Fig. 3). Clusters 1, 2, 3, and 4 contained genes, which were characterized by short-term reduced transcript levels. Genes of clusters 5 and 6 showed a decline in transcript level during the 24 h LC treatment. Clusters 7 and 8 contained genes with rapid transcript accumulation after 3 h LC conditions, while in cluster 9 transcript levels constantly increased until 24 h after LC shift. Cluster 10 contained genes that were specifically induced in expression 24 h after LC shift.

To functionally characterize these clusters, we tested them for enrichment of gene ontology (GO) terms and MapMan categories (Thimm et al., 2004) expecting similar enrichments with these independent methods. Among the clusters, which contained short-term (3 h after LC shift) down-regulated genes, we found cluster 2 enriched for genes of the GO terms photosynthesis, light reactions, and tetrapyrrole biosynthesis (Table S2) and cluster 3 enriched with terms related to vesicle-mediated transport. Clusters 5 and 6 contained genes with reduced abundance in long-term CO_2 deprivation (24 h after LC shift). Cluster 5 showed an enrichment in the MapMan category DNA synthesis/chromatin structure (Table S1) and corresponding GO terms related to DNA replication (Table S2), while cluster 6 was enriched in genes belonging to GO terms and MapMan categories related to translation and protein biosynthesis (Table S2) and protein biosynthesis (Table S1), respectively. For cluster 7, which contained short-term LC-induced genes, we observed a significant enrichment of genes involved in the MapMan category photosynthesis: light reactions (Table S1). Among the clusters 8, 9, and 10, which contained constantly and long-term LC-induced genes (Fig. 3), cluster 8 was significantly enriched for genes connected to photorespiration (Table S1). No significant enrichments were detected for cluster 9 and 10.

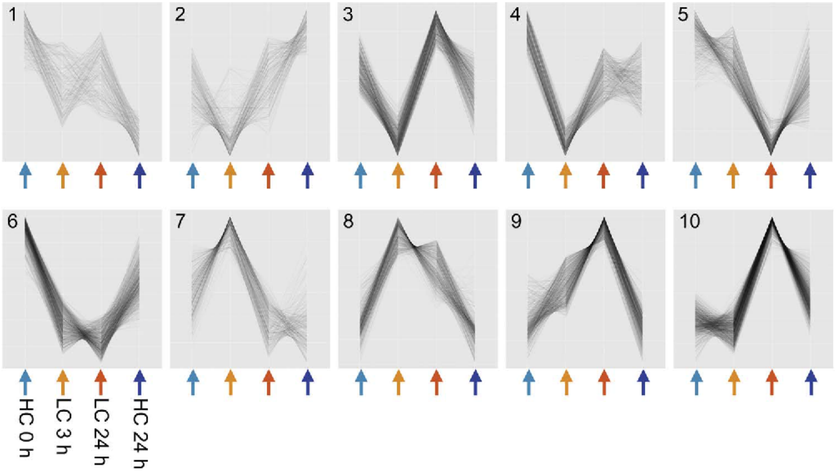


Fig. 3. Analysis of CO_2-dependent expression patterns by *K*-means clustering. *K*-means cluster are shown in *z*-scores. Presented are 10 different clusters (1–10), representing different expression patterns upon changes in CO_2 conditions. Color-coded arrows indicate different samples.

All genes encoding the photorespiratory enzymes (Rademacher et al., 2016) were strongly induced, ranging from 2.6-fold for hydroxypyruvate reductase (*CMQ289C*) to 53-fold for alanine:glyoxylate aminotransferase (*CMS429C*) 3 h after shift from HC to LC conditions. Though at a lower level compared to the LC 3 h value, these genes were still significantly up-regulated 24 h after the LC shift (Fig. 4, Table S5). Besides the photorespiratory genes, we also identified core elements of a hypothetical C_4 cycle, funneling pyruvate into a PEP oxaloacetate cycle (Fig. 4, Table S5), significantly up-regulated during the LC treatment. Cluster 8 contained the genes encoding phosphoenolpyruvate carboxylase (PPC, CME095C) and phosphoenolpyruvate carboxykinase (PEPCK, CMN285C). With a 22-fold induction, pyruvate phosphate dikinase (*PPDK, CMF012C*) was one of the 10 most up-regulated genes 3 h after the CO_2 shift. The transcript accumulation was even higher 24 h after LC shift. Accordingly, *PPDK* belonged to cluster 9. The pyruvate regenerating mechanism via malate dehydrogenase and NADP-malic enzyme exists in *C. merolae* as both genes were expressed at HC 0 h. Their expression however was significantly reduced in response to the LC shift (Fig. 4, Table S5).

3.3. Bicarbonate transporters and carbonic anhydrases

We furthermore employed the RNA-seq data set to search for possible components of a hypothetical CCM. We followed the rationale that candidate genes involved in an inducible CCM should be co-regulated with genes encoding proteins of the photorespiratory metabolism, as demonstrated for the green alga *Chlamydomonas reinhardtii* (*C. reinhardtii*) (Fang et al., 2012). Thus, we investigated clusters 8 and 9, which contained the majority of photorespiratory genes, for the occurrence of HCO_3^- transporters and CAs, which might participate in a hypothetical CCM. BlastP analyses identified two candidate proteins in *C. merolae* homologous to known HCO_3^- transport proteins in *C. reinhardtii* (Table S6). The protein CMN251C showed

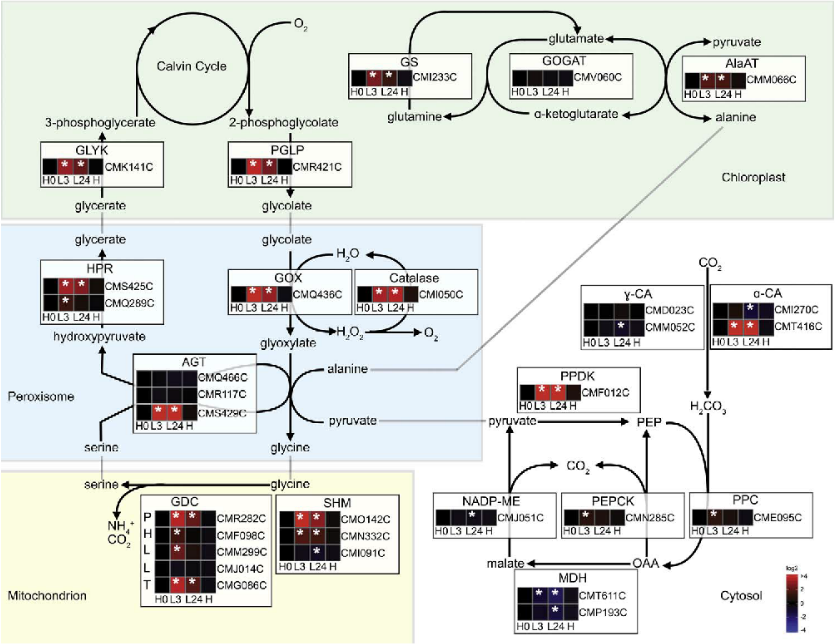


Fig. 4. The transcriptional response of genes involved in photorespiratory metabolism and HCO_3^- homeostasis to changes in CO_2 availability. The expression profiles of all genes involved in photorespiration and carbon sequestering from HCO_3^- are presented as log2-fold changes relative to their expression at sampling point H0, and are illustrated as heat map. The sampling points are indicated by H0 (HC 0 h), L3 (LC 3 h), L24 (LC 24 h), and H (HC 24 h). All significant expression changes, at $q < 0.05$, are indicated by white asterisks.

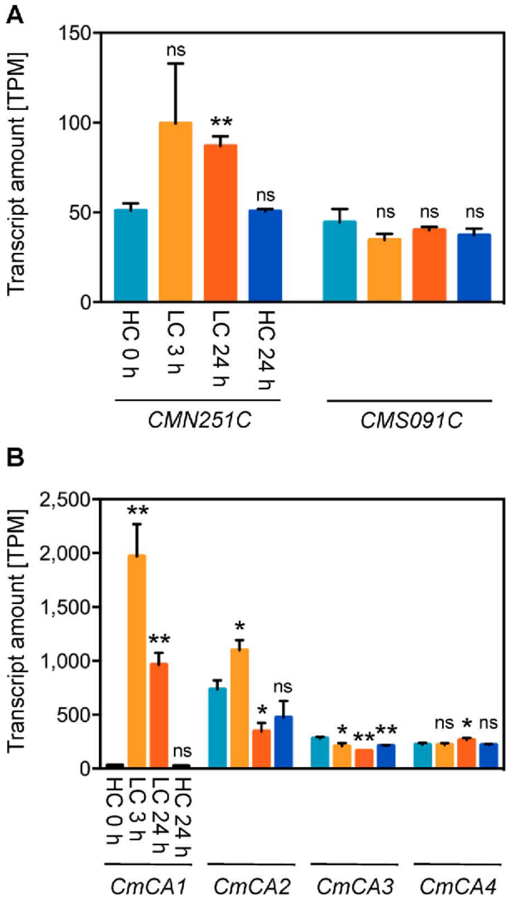


Fig. 5. Effect of changes in CO_2 concentrations on transcript abundances of genes encoding hypothetical HCO_3^- transporters and carbonic anhydrases. A: CO_2-dependent changes in transcript amounts of CMN251C and CMS091C, encoding hypothetical HCO_3^- transporters. B. CO_2-dependent changes in transcript amounts of CmCA1 (CMT416C), CmCA2 (CMI270C), CmCA3 (CMM052C), and CmCA4 (CMD023C). Transcript amounts are given in transcripts per million (TPM). Significant differences in relation to the HC 0 h value (paired *t*-test, two-tailed), are indicated as * ($P < 0.05$) and ** ($P < 0.01$), ns: not significant.

homology to the HIGH-LIGHT INDUCED GENE3 (HLA3) HCO_3^- transporter, which is induced under limiting CO_2 conditions in *C. reinhardtii* (Tirumani et al., 2014). The protein CMS091C is homologous to the CHLOROPLAST ENVELOPE PROTEINS1 and 2 (CCP1 and CCP2), which showed HCO_3^- transporter function in *C. reinhardtii* (Tirumani et al., 2014). Expression of the gene *CMN251C* followed the same transcription pattern as the photorespiratory genes with a 2-fold transcript accumulation 3 h and 24 h after the shift to LC conditions (Fig. 5A, Table S5, cluster 8). In contrast, expression of *CMS091C* did not cluster together with the photorespiratory genes. The transcription pattern was not affected by the CO_2 conditions (Fig. 5A, Table S5) and grouped in cluster 4.

BlastP analyses with known α-, β-, and γ-CAs of the model plant *Arabidopsis thaliana* (*A. thaliana*) (Fabre et al., 2007) revealed four proteins homologous to CAs in *C. merolae*: CmCA1 (CMT416C), CmCA2 (CMI270C), CmCA3 (CMM052C), and CmCA4 (CMD023C). CmCA1 and CmCA2 were homologous to α-CAs and CmCA3 and CmCA4 homologous to γ-CAs of *A. thaliana* (Table S7). Genes encoding α-CAs showed CO_2-dependent transcriptional dynamics (Fig. 5B, Table S5). Expression of *CmCA1* was most affected by the changing CO_2 conditions. While under HC conditions the gene was only weakly expressed, we observed a 67-fold accumulation of *CmCA1* transcript 3 h after the shift to LC conditions and a 33-fold increased value 24 h after the shift. The final 24 h HC treatment lead to full recovery to the initial HC transcript value (cluster 8). Transcript amounts of *CmCA2* increased 1.5-fold 3 h after the LC shift, went down to 50% of the initial HC value after 24 h LC treatment and fully recovered to the HC default value after 24 h HC conditions (cluster 7). In contrast, both γ-CAs, *CmCA3* and *CmCA4,* were quite constantly expressed in *C. merolae* under HC and LC conditions (Fig. 5B, Table S5). The LC shift induced a slight reduction (0.7-fold) in transcript abundances for *CmCA3* (cluster 6), while transcript amounts were slightly but significantly increased (1.2-fold) for *CmCA4* 24 h after shift to LC conditions (cluster 10).

3.4. Subcellular localization of CmCA1 and CmCA2

On the basis of the LC-inducible expression of *CmCA1* and *CmCA2*, we investigated the encoded proteins in more detail. An amino acid alignment showed an identity of 90% between both proteins. Interestingly, CmCA1 contained an additional N-terminal extension of 109 amino acids length (Fig. S1). Closer investigation of this extension indicated the occurrence of a transmembrane region (23 amino acids length) predicted by the SOSUI tool (Hirokawa et al., 1998) and a mitochondrial target peptide predicted by TargetP (Emanuelsson et al., 2000) (Fig. S1). To determine in which subcellular compartment the α-CAs of *C. merolae* reside, localization constructs were designed with an N- or C-terminal YFP-fusion to avoid masking of a hypothetical target peptide signal. In transiently transformed tobacco protoplasts YFP:CmCA1, YFP:CmCA2, and CmCA2:YFP were detected in the cytosol. For CmCA1:YFP the fluorescent signal was mostly detected around chloroplasts, indicating a possible attachment of CmCA1 to the chloroplast envelope (Fig. 6A). Transient expression of the YFP fusion proteins in *C. merolae* cells consistently revealed cytosolic localization of YFP:CmCA1, YFP:CmCA2, and CmCA2:YFP. In the case of CmCA1:YFP we observed YFP fluorescence at the interface between chloroplast and mitochondrion (Fig. 6B).

4. Discussion

The impact of reduced CO_2 availability on the transcriptome of *C. merolae* was large. In the short-term (3 h after shift to LC conditions), 11% of all genes were changed in transcript abundance (Fig. 2), which is well in the range of a comparable study with *C. reinhardtii* 4 h after a shift from HC to LC conditions (Fang et al., 2012) and markedly different from the response of land plants (Eisenhut et al., 2017; Queval et al., 2012). Long-term (24 h) LC treatment increased the number of significant changes (21% of all genes changed, Fig. 2) in line with the larger variation captured in the PCA in the direction of long-term change compared to short-term change (Fig. 1B).

The three sampling time-points captured three pivotal phases in the responses of *C. merolae* to changes in CO_2 availability: the initial counter response including up-regulation of photorespiration and a potential CCM, the long-term acclimation to reduced carbon in the metabolic system, and the resupply with carbon in the system. Intriguingly, the short-term high abundance response evident in the clusters 8 and 9 (Fig. 3) included not only the full set of genes encoding the enzymes of the photorespiratory cycle (derived from Rademacher et al., 2016), but also CAs and genes known to be involved in the C_4 cycle (Table S5). This coordinated up-regulation (Fig. 4, Table S5) strongly underscores the importance of the photorespiratory pathway in *C. merolae* under CO_2 concentrations present in its ecological niche, and points to a concerted transcriptional regulon of the photorespiratory genes. Among the potential transcriptional regulators changing in expression concomitantly in clusters 8 and 9, the strongest changes in amplitude were present in putative histone de-acetylases (*CMQ158C*, *CMG132C*, Table S5), suggesting a potential role of epigenetic processes, in a gene similar to sigma factor B (*CMR165C*, Table

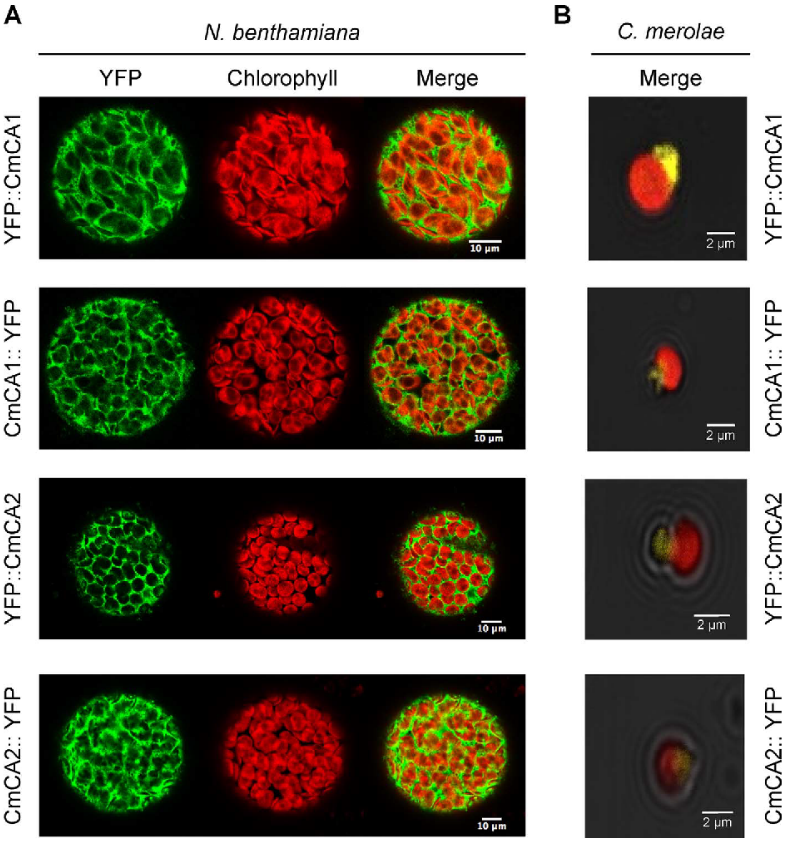


Fig. 6. Subcellular localization studies of carbonic anhydrases. CmCA1 (CMT416C) and CmCA2 (CMI270C) were N- and C-terminally fused with YFP, respectively. A. Fluorescent pictures of tobacco protoplast, transiently expressing YFP:CmCA1, CmCA1:YFP, YFP:CmCA2, and CmCA2:YFP fusion proteins, respectively. Pictures were recorded 48 h after infiltration of *N. benthamiana* leaves. B) Localization of CmCA1 and CmCA2 in *C. merolae*. Shown are merged pictures. The YFP signal is presented in green and yellow, respectively, chlorophyll autofluorescence in red. *C. merolae* cells were transformed 24 h before microscope analysis.

S5), which likely mediates chloroplast transcription, and in a regulator homologous to BRUTUS (*CMM141C*, Table S5), which controls the iron response transcription factor POPEYE in higher plants. Long-term carbon shortage at the LC 24 h time-point reduces growth (Rademacher et al., 2016) and consequently growth associated processes, such as DNA replication, the cell cycle, and protein biosynthesis were reduced in transcriptional abundance (Fig. 3: clusters 5 and 6, Table S1, Table S2). Reduction of transcriptional abundance for these processes has also been observed in carbon starved land plants (Brilhaus et al., 2016) and carbon starved *C. reinhardtii* cells (Brueggeman et al., 2012; Fang et al., 2012) pointing to a preserved ancient survival mechanism. During the response to LC, induced by low external CO_2 in algae or by drought induced stomatal closure in land plants, the responses are however different. In both cases, the species initially respond with a reduction in the transcriptional investment into photosynthesis (Fig. 3, cluster 2) (Brilhaus et al., 2016). The long-term responses however present a different pattern. The land plant maintains the low expression levels of photosynthetic genes, while the red alga recovers its transcriptional investment. *C. merolae* naturally grows under LC conditions and presumably recovers photosynthetic transcript abundance because its acclimation response re-enables efficient photosynthesis. The resupply of abundant CO_2 and presumably concomitant availability of fixed carbon mitigated the majority of changes induced by 24 h of LC status as is evident from the short distance between HC 0 h and HC 24 h points in the PCA (Fig. 1B), and from the overall pattern (Fig. 3). This mitigation of changes has also been observed in land plants upon resupply with water, which in turn caused resupply with carbon (Brilhaus et al., 2016). However, a low number of genes (Fig. 2C) did not return to the initial HC starting expression value after a 24 h HC treatment. Among these, we detected the gene encoding catalase (*CMI050C*), which participates in the photorespiratory metabolism. The ongoing significantly enhanced gene expression might indicate that this time period was not sufficient for *C. merolae* to fully return to the metabolic or energetic state of HC cells despite attaining a growth rate equal to before treatment (Rademacher et al., 2016).

Co-expression is a powerful tool to search for unknown components in functionally related processes. It was demonstrated for *C. reinhardtii* that genes involved in photorespiratory metabolism are co-regulated with genes involved in an inducible CCM (Fang et al., 2012) and a photorespiratory transporter was identified by co-expression analysis in higher plants (Bordych et al., 2013; Pick et al., 2013). Thus, we searched clusters 8 and 9, which contained photorespiratory genes, for candidates establishing a hypothetical CCM in *C. merolae*. We identified a hypothetical HCO_3^- transporter, two CAs, and enzymes constituting a PEPCK-type C_4 cycle within the LC-inducible clusters 8 and 9 (Table S5). We suggest the following hypothetical model (Fig. 7) for the concerted action of these proteins to function as CCM in *C. merolae*:

In its natural habitat with an acidic pH the major form of C_i is CO_2, which diffuses into the red algal cell. The cytoplasmic CAs (CmCA2 and CmCA3, Fig. 6) convert the CO_2 rapidly into HCO_3^- and thus intracellularly capture the C_i. Additionally, we identified the protein CMN251C, which shows homology to the LC-inducible HLA3 HCO_3^- transporter from *C. reinhardtii* (Table S6). The protein is predicted to reside in the plasma membrane. Thus, CMN251C is a promising candidate for C_i uptake under LC conditions in *C. merolae*. Photosynthesis in the red alga is most efficient at a low extracellular pH and decreases at a neutral pH (Zenvirth et al., 1985) calling into question whether HCO_3^- is the major C_i species taken up from the external medium by *C.*

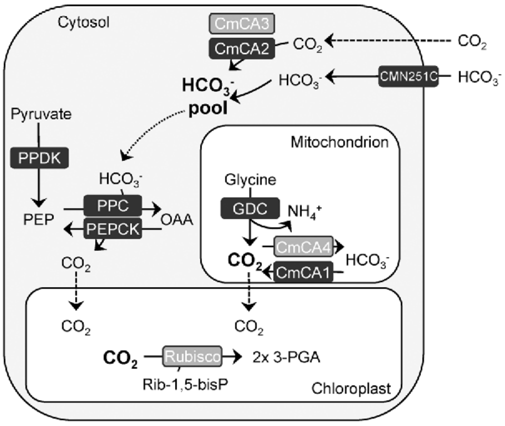


Fig. 7. Hypothetical model of a CCM *in C. merolae*. At low pH, C_i from the aquatic environment predominantly diffuses as CO_2 into the cell, where it is converted by cytoplasmic CAs, CmCA2 and CmCA3, into HCO_3^-. Traces of environmental HCO_3^- are imported by the plasma membrane protein CMN251, which is homologous to HLA3. From the cytoplasmic pool, HCO_3^- is drained by the action of PPC, generating OAA and pulling more environmental CO_2 into the cytoplasm. Decarboxylation of OAA by PEPCK allows stepwise release of CO_2, which diffuses into the chloroplast for fixation by Rubisco. Besides this PEPCK-type C_4 cycle, the concerted action of GDC and CmCA1 might function as a CO_2 pump, by accumulating high amounts of CO_2 in the mitochondrion, which is in close proximity to the chloroplast in the red algal cell. More details are presented in the text. Proteins, which are encoded by genes of the LC-inducible clusters 8 and 9 and presented in black boxes.

merolae. Possibly, CMN251C's role is the harvest of trace amounts of environmental HCO_3^- under C_i limited condition or serving as a CO_2 diffusion facilitator.

Genes encoding the components of a PEPCK-type C_4 pathway (Fig. 4) were also contained in the photorespiratory gene cluster 8. This might indicate a temporary HCO_3^- partitioning via PPC and PEPCK from the cytoplasmic HCO_3^- pool into organic acids, thereby increasing the concentration gradient and hence generating additional "pull" for CO_2 entry into the cell. By the action of PEPCK the CO_2 is released stepwise and allowed to diffuse into the chloroplast for fixation by Rubisco. A single cell C_4 metabolism as proposed in diatoms is a distinct possibility. For example, in LC-acclimated *Thalassiosira weissflogii* cells, it was shown that repression of PPC activity by the inhibitor 3,3-dichloro-2-dihydroxyphosphinoylmethyl-2-propenoate resulted in a 90% decrease in the rate of photosynthesis, which could be overcome by HC (Reinfelder et al., 2004). For the model diatom *Phaeodactylum tricornutum,* the primary function of C_4 metabolism might not be in CO_2 fixation but rather play a role in dissipating energy and maintaining pH homeostasis (Haimovich-Dayan et al., 2013). Alternatively, the transient upregulation of PPC and PEPCK under CO_2 limitation might have to do with overcoming short-term carbon limitation and gluconeogenesis from lipid stores and/or protein degradation.

The shift to LC conditions enhances the flux through the photorespiratory cycle in *C. merolae* (Rademacher et al., 2016). Accordingly, photorespiratory glycine cleavage in the mitochondrion by the GDC generates increased amounts of CO_2 and NH_3. The latter is likely protonated to NH_4^+ at the pH of the mitochondrial matrix leading to an alkalization (Eriksson et al., 1996). A function in maintaining the pH homeostasis was suggested for the LC-inducible mitochondrial CAs in *C. reinhardtii* (Eriksson et al., 1996). In accordance, we postulate that CmCA1, which was almost exclusively expressed under LC conditions (Fig. 5B), fulfills this function in *C. merolae*. Interestingly, we observed in transiently transformed *C. merolae* cells the YFP fluorescence signal of a CmCA1:YFP protein at the interface between chloroplast and mitochondrion (Fig. 6B). A localization to the mitochondrial membrane is in accordance with the *in silico* prediction of a transmembrane helix and a mitochondrial targeting peptide in the N-terminal extension specific for CmCA1 (Fig. S1). If CmCA1 was indeed attached to the mitochondrial membrane, its CO_2 production might allow high CO_2 accumulation in close proximity to the chloroplast. Thus, the mitochondrial CO_2 provision by the concerted action of GDC and CmCA1 might serve as a CO_2 pump for CO_2 enrichment around Rubisco as part of a CCM. Alternatively, CmCA1's role may be the intracellular trapping of the high volume of CO_2 released during photorespiratory glycine decarboxylation avoiding leakage of CO_2 from the cell.

Clusters 8 and 9 also contained numerous genes encoding proteins of unknown function (Table S3). These LC-responsive proteins are potential candidates for involvement in the LC acclimation in *C. merolae* either to detoxify Rubisco oxygenation products or to enrich CO_2 at the site of Rubisco.

Although the RNA-seq data only provide tantalizing suggestions about a CCM in the red alga and eventually cannot distinguish between a transport based CCM, a biochemical CCM, a mixed type as here suggested (Fig. 7), or an as of yet undiscovered mechanism, *C. merolae* has demonstrated higher apparent affinity to CO_2 under LC conditions (Rademacher et al., 2016; Zenvirth et al., 1985), raising the possibility that one or more of the mechanism described above are functional.

5. Conclusion

In conclusion, the red alga *C. merolae* shows a distinct transcriptional response to reduction in CO_2 availability. The CO_2 limitation makes the cells reduce transcripts of protein biosynthesis and DNA replication when growth is stalled. As one strategy to acclimate to LC conditions, abundance of photorespiratory transcripts is uniformly upregulated. Additionally, transcripts constituting a hypothetical PEPCK C_4 pathway, a C_i transporter, and the mitochondrial CA CmCA1 are higher in abundance. Whether this represents a CCM footprint or a physiological adaptation in pH maintenance or energy homeostasis remains to be genetically and physiologically tested.

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (WE 2231/8-2, WE2231/11-1). APMW acknowledges additional funding from EXC 1028. We thank the HHU BMFZ for technical support of the RNA-seq experiments.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jplph.2017.06.014.

References

topGO: Enrichment Analysis for Gene Ontology. R package version 2.26.0.

Badger, M.R., Price, G.D., 2003. CO_2 concentrating mechanisms in cyanobacteria: molecular components, their diversity and evolution. J. Exp. Bot. 54, 609–622.

Bauwe, H., Hagemann, M., Fernie, A.R., 2010. Photorespiration: players, partners and origin. Trends Plant Sci. J 15, 330–336.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B: Methodol. 57.

Benjamini, Y., Yekutieli, D., 2001. The control of the false discovery rate in multiple testing under dependency. Ann. Stat. 29, 1165–1188.

Bordych, C., Eisenhut, M., Pick, T.R., Kuelahoglu, C., Weber, A.P.M., 2013. Co-expression analysis as tool for the discovery of transport proteins in photorespiration. Plant Biol. 15, 686–693.

Breuers, F.K.H., Bräutigam, A., Geimer, S., Welzel, U., Stefano, G., Renna, L., Brandizzi, F., Weber, A.P.M., 2012. Dynamic remodeling of the plastid envelope membranes–a tool for chloroplast envelope *in vivo* localizations. Front. Plant Sci. 3, 1–10.

Brilhaus, D., Bräutigam, A., Mettler-Altmann, T., Winter, K., Weber, A.P., 2016. Reversible burst of transcriptional changes during induction of crassulacean acid metabolism in *Talinum triangulare*. Plant Physiol. 170, 102–122.

Brueggeman, A.J., Gangadharaiah, D.S., Cserhati, M.F., Casero, D., Weeks, D.P., Ladunga, I., 2012. Activation of the carbon concentrating mechanism by CO_2 deprivation coincides with massive transcriptional restructuring in *Chlamydomonas reinhardtii*. Plant Cell 24, 1860–1875.

Dereeper, A., Guignon, V., Blanc, G., Audic, S., Buffet, S., Chevenet, F., Dufayard, J.F., Guindon, S., Lefort, V., Lescot, M., Claverie, J.M., Gascuel, O., 2008. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res. 36, W465–9.

Eisenhut, M., Kahlon, S., Hasse, D., Ewald, R., Lieman-Hurwitz, J., Ogawa, T., Ruth, W., Bauwe, H., Kaplan, A., Hagemann, M., 2006. The plant-like C2 glycolate cycle and the bacterial-like glycerate pathway cooperate in phosphoglycolate metabolism in cyanobacteria. Plant Physiol. 142, 333–342.

Eisenhut, M., Ruth, W., Haimovich, M., Bauwe, H., Kaplan, A., Hagemann, M., 2008. The photorespiratory glycolate metabolism is essential for cyanobacteria and might have been conveyed endosymbiontically to plants. Proc. Natl. Acad. Sci. U. S. A. 105, 17199–17204.

Eisenhut, M., Bräutigam, A., Timm, S., Florian, A., Tohge, T., Fernie, A.R., Bauwe, H., Weber, A.P., 2017. Photorespiration is crucial for dynamic response of photosynthetic metabolism and stomatal movement to altered CO_2 availability. Mol. Plant. 10, 47–61.

Emanuelsson, O., Nielsen, H., Brunak, S., von Heijne, G., 2000. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J. Mol. Biol. 300, 1005–1016.

Eriksson, M., Karlsson, J., Ramazanov, Z., Gardeström, P., Samuelsson, G., 1996. Discovery of an algal mitochondrial carbonic anhydrase: molecular cloning and characterization of a low-CO_2-induced polypeptide in *Chlamydomonas reinhardtii*. Proc. Natl. Acad. Sci. U. S. A. 93, 12031–12034.

Fabre, N., Reiter, I.M., Becuwe-Linka, N., Genty, B., Rumeau, D., 2007. Characterization and expression analysis of genes encoding alpha and beta carbonic anhydrases in Arabidopsis. Plant Cell Environ. 30, 617–629.

Fang, W., Si, Y., Douglass, S., Casero, D., Merchant, S.S., Pellegrini, M., Ladunga, I., Liu, P., Spalding, M.H., 2012. Transcriptome-wide changes in *Chlamydomonas reinhardtii* gene expression regulated by carbon dioxide and the CO_2-concentrating mechanism regulator CIA5/CCM1. Plant Cell 24, 1876–1893.

Giordano, M., Beardall, J., Raven, J.A., 2005. CO_2 concentrating mechanism in algae: mechanisms, environmental modulation, and evolution. Annu. Rev. Plant Biol. 56, 99–131.

Grefen, C., Donald, N., Hashimoto, K., Kudla, J., Schumacher, K., Blatt, M.R., 2010. A ubiquitin-10 promoter-based vector set for fluorescent protein tagging facilitates temporal stability and native protein distribution in transient and stable expression studies. Plant J. 64, 355–365.

Hagemann, M., Kern, R., Maurino, V.G., Hanson, D.T., Weber, A.P.M., Sage, R.F., Bauwe, H., 2016. Evolution of photorespiration from cyanobacteria to land plants, considering protein phylogenies and acquisition of carbon concentrating mechanisms. J. Exp. Bot. 67, 2963–2976.

Haimovich-Dayan, M., Garfinkel, N., Ewe, D., Marcus, Y., Gruber, A., Wagner, H., Kroth, P.G., Kaplan, A., 2013. The role of C4 metabolism in the marine diatom *Phaeodactylum tricornutum*. New Phytol. 197, 177–185.

Hauser, T., Popilka, L., Hartl, F.U., Hayer-Hartl, M., 2015. Role of auxiliary proteins in Rubisco biogenesis and function. Nat. Plants 1, 15065.

Hirokawa, T., Boonćchieng, S., Mitaku, S., 1998. SOSUI: classification and secondary structure prediction system for membrane proteins. Bioinformatics 14, 378–379.

Imamura, S., Terashita, M., Ohnuma, M., Maruyama, S., Minoda, A., Weber, A.P., Inouye, T., Sekine, Y., Fujita, Y., Omata, T., Tanaka, K., 2010. Nitrate assimilatory genes and their transcriptional regulation in a unicellular red alga *Cyanidioschyzon merolae*: genetic evidence for nitrite reduction by a sulfite reductase-like enzyme. Plant Cell Physiol. 51, 707–717.

Kaplan, A., Reinhold, L., 1999. CO_2 concentrating mechanisms in photosynthetic microorganisms. Annu. Rev. Plant. Physiol. Plant. Mol. Biol. 50, 539–570.

Krzywinski, M., Altman, N., 2014. Points of significance. Comparing samples—part I. Nat. Methods 11, 215–216.

Li, B., Dewey, C.N., 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinf. 12, 1–16.

Matsuzaki, M., Misumi, O., Shin-I, T., et al., 2004. Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. Nature 428, 653–657.

McCarthy, D.J., Chen, Y., Smyth, G.K., 2012. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. 40, 4288–4297.

Minoda, A., Sakagami, R., Yagisawa, F., Kuroiwa, T., Tanaka, K., 2004. Improvement of culture conditions and evidence for nuclear transformation by homologous recombination in a red alga, Cyanidioschyzon merolae 10D. Plant Cell Physiol. 45, 667–671.

Nakamura, Y., Kanakagiri, S., Van, K., He, W., Spalding, M.H., 2005. Disruption of the glycolate dehydrogenase gene in the high-CO_2-requiring mutant HCR89 of *Chlamydomonas reinhardtii*. Can. J. Bot. 83, 820–833.

Nozaki, H., Takano, H., Misumi, O., et al., 2007. A 100%-complete sequence reveals unusually simple genomic features in the hot-spring red alga *Cyanidioschyzon merolae*. BMC Biol. 5, 28.

Ohnuma, M., Yokoyama, T., Inouye, T., Sekine, Y., Tanaka, K., 2008. Polyethylene glycol (PEG)-mediated transient gene expression in a red alga, *Cyanidioschyzon merolae* 10D. Plant Cell Physiol. 49, 117–120.

Pérez-Rodríguez, P., Riaño-Pachón, D.M., Corrêa, L.G., Rensing, S.A., Kersten, B., Mueller-Roeber, B., 2010. PlnTFDB: updated content and new features of the plant transcription factor database. Nucleic Acids Res. 38, D822–7.

Pick, T.R., Bräutigam, A., Schulz, M.A., Obata, T., Fernie, A.R., Weber, A.P., 2013. PLGG1, a plastidic glycolate glycerate transporter, is required for photorespiration and defines a unique class of metabolite transporters. Proc. Natl. Acad. Sci. U. S. A. 110, 3185–3190.

Queval, G., Neukermans, J., Vanderauwera, S., Van Breusegem, F., Noctor, G., 2012. Day length is a key regulator of transcriptomic responses to both CO_2 and H_2O_2 in Arabidopsis. Plant Cell Environ. 35, 374–387.

Rademacher, N., Kern, R., Fujiwara, T., Mettler-Altmann, T., Miyagishima, S.Y., Hagemann, M., Eisenhut, M., Weber, A.P., 2016. Photorespiratory glycolate oxidase is essential for the survival of the red alga *Cyanidioschyzon merolae* under ambient CO_2 conditions. J. Exp. Bot. 67, 3165–3175.

Raven, J.A., Cockell, C.S., De La Rocha, C.L., 2008. The evolution of inorganic carbon concentrating mechanisms in photosynthesis. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 363, 2641–2650.

Raven, J.A., Giordano, M., Beardall, J., Maberly, S.C., 2012. Algal evolution in relation to atmospheric CO_2: carboxylases, carbon-concentrating mechanisms and carbon oxidation cycles. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 367, 493–507.

Reinfelder, J.R., Milligan, A.J., Morel, F.M., 2004. The role of the C4 pathway in carbon accumulation and fixation in a marine diatom. Plant Physiol. 135, 2106–2111.

Savir, Y., Noor, E., Milo, R., Tlusty, T., 2010. Cross-species analysis traces adaptation of Rubisco toward optimality in a low-dimensional landscape. Proc. Natl. Acad. Sci. U. S. A. 107, 3475–3480.

Seckbach, J., 1995. The first eukaryotic cells – acid hot-spring algae. J. Biol. Phys. 20, 335–345.

Tcherkez, G.G., Farquhar, G.D., Andrews, T.J., 2006. Despite slow catalysis and confused substrate specificity: all ribulose bisphosphate carboxylases may be nearly perfectly optimized. Proc. Natl. Acad. Sci. U. S. A. 103, 7246–7251.

Thimm, O., Bläsing, O., Gibon, Y., Nagel, A., Meyer, S., Krüger, P., Selbig, J., Müller, L.A., Rhee, S.Y., Stitt, M., 2004. MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J. 37, 914–939.

Tirumani, S., Kokkanti, M., Chaudhari, V., Shukla, M., Rao, B.J., 2014. Regulation of CCM genes in *Chlamydomonas reinhardtii* during conditions of light-dark cycles in synchronous cultures. Plant Mol. Biol. 85, 277–286.

Uemura, K., Anwaruzzaman, Miyachi S., Yokota, A., 1997. Ribulose-1,5-bisphosphate carboxylase/oxygenase from thermophilic red algae with a strong specificity for CO_2 fixation. Biochem. Biophys. Res. Comm. 233, 568–571.

Watanabe, S., Ohnuma, M., Sato, J., Yoshikawa, H., Tanaka, K., 2011. Utility of a GFP reporter system in the red alga *Cyanidioschyzon merolae*. J. Gen. Appl. Microbiol. 57, 69–72.

Yates, A., Akanni, W., Amode, M.R., et al., 2016. Ensembl 2016. Nucleic Acids Res. 44, D710–6.

Zelitch, I., Schultes, N.P., Peterson, R.B., Brown, P., Brutnell, T.P., 2009. High glycolate oxidase activity is required for survival of maize in normal air. Plant Physiol. 149, 195–204.

Zenvirth, D., Volokita, M., Kaplan, A., 1985. Photosynthesis and inorganic carbon accumulation in the acidophilic alga *Cyanidioschyzon merolae*. Plant Physiol. 77, 237–239.

Manuscript 5

Systems Biology of Cold Adaptation in the Polyextremophilic Red Alga *Galdieria sulphuraria*

frontiers in Microbiology

ORIGINAL RESEARCH
published: 01 May 2019
doi: 10.3389/fmicb.2019.00927

Systems Biology of Cold Adaptation in the Polyextremophilic Red Alga *Galdieria sulphuraria*

*Alessandro W. Rossoni and Andreas P. M. Weber**

Cluster of Excellence on Plant Sciences (CEPLAS), Institute of Plant Biochemistry, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

Rapid fluctuation of environmental conditions can impose severe stress upon living organisms. Surviving such episodes of stress requires a rapid acclimation response, e.g., by transcriptional and post-transcriptional mechanisms. Persistent change of the environmental context, however, requires longer-term adaptation at the genetic level. Fast-growing unicellular aquatic eukaryotes enable analysis of adaptive responses at the genetic level in a laboratory setting. In this study, we applied continuous cold stress (28°C) to the thermoacidophile red alga *G. sulphuraria*, which is 14°C below its optimal growth temperature of 42°C. Cold stress was applied for more than 100 generations to identify components that are critical for conferring thermal adaptation. After cold exposure for more than 100 generations, the cold-adapted samples grew ~30% faster than the starting population. Whole-genome sequencing revealed 757 variants located on 429 genes (6.1% of the transcriptome) encoding molecular functions involved in cell cycle regulation, gene regulation, signaling, morphogenesis, microtubule nucleation, and transmembrane transport. CpG islands located in the intergenic region accumulated a significant number of variants, which is likely a sign of epigenetic remodeling. We present 20 candidate genes and three putative *cis*-regulatory elements with various functions most affected by temperature. Our work shows that natural selection toward temperature tolerance is a complex systems biology problem that involves gradual reprogramming of an intricate gene network and deeply nested regulators.

Keywords: microevolution, Cyanidiales, extremophile, temperature adaptation, cold stress, red algae

OPEN ACCESS

Edited by:
Baolei Jia,
Chung-Ang University, South Korea

Reviewed by:
Julian Whitelegge,
University of California, Los Angeles, United States
Thinesh Selvaratnam,
Lamar University, United States

****Correspondence:***
Andreas P. M. Weber
andreas.weber@uni-duesseldorf.de

Specialty section:
This article was submitted to Extreme Microbiology, a section of the journal Frontiers in Microbiology

Received: *02 March 2019*
Accepted: *12 April 2019*
Published: *01 May 2019*

Citation:
Rossoni AW and Weber APM (2019) Systems Biology of Cold Adaptation in the Polyextremophilic Red Alga Galdieria sulphuraria. Front. Microbiol. 10:927. doi: 10.3389/fmicb.2019.00927

INTRODUCTION

Small changes in average global temperature significantly affect the species composition of ecosystems. Indeed, 252 Ma years ago up to ~95% of marine species and ~70% of terrestrial vertebrates ceased to exist (Benton, 2008; Sahney and Benton, 2008). This event, known as the Permian–Triassic extinction, was triggered by a sharp increase in worldwide temperature (+8°C) and CO_2 concentrations (+2000 ppm) during a period spanning 48,000–60,000 years (McElwain and Punyasena, 2007; Shen et al., 2011; Burgess et al., 2014). In comparison, atmospheric CO_2 has increased by ~100 ppm and the global mean surface temperature by ~1°C since the sinking of the Titanic in 1912, a little more than 100 years ago. Anthropogenic climate change and its consequences have become a major evolutionary selective force (Palumbi, 2001). Higher temperatures and CO_2 concentrations result in increased seawater acidity, increased UV radiation, and changes in oceanwide water circulation and upwelling patterns. These rapid changes represent dramatically accelerating shifts in the demography and number of species, leading to loss of habitats and biodiversity (Hendry and Kinnison, 1999; Stockwell et al., 2003). A global wave of mass extinction appears inevitable (Kolbert, 2014). In this context, it is relevant to assess the effects of temperature change on genome evolution. Aquatic unicellular eukaryotes are particularly well-suited to addressing this question due to their short generation time and straightforward temperature control of their growth environment.

Microorganisms rapidly acclimate and subsequently adapt to environmental change (López-Rodas et al., 2009; Huertas et al., 2011; Romero-Lopez et al., 2012; Osundeko et al., 2014; Foflonker et al., 2018). These adaptations are driven by natural selection and involve quantitative changes in allele frequencies and phenotype within a short period of time, a phenomenon known as microevolution. The *Galdieria* lineage comprise a monophyletic clade of polyextremophilic, unicellular red algae (Rhodophyta) that thrive in acidic and thermal habitats worldwide (e.g., volcanoes, geysers, acid mining sites, acid rivers, urban wastewaters, and geothermal plants) where they represent up to 90% of the total biomass, competing with specialized Bacteria and Archaea (Seckbach, 1972; Castenholz and McDermott, 2010). Accordingly, members of the *Galdieria* lineage can cope with extremely low pH values, temperatures above 50°C, and high salt and toxic heavy metal ion concentrations (Doemel and Brock, 1971; Castenholz and McDermott, 2010; Reeb and Bhattacharya, 2010; Hsieh et al., 2018). Some members of this lineage also occur in more temperate environments (Gross et al., 2002; Ciniglia et al., 2004; Qiu et al., 2013; Barcytė et al., 2018; Iovinella et al., 2018).

Our work systematically analyzed the impact of prolonged exposure to suboptimal (28°C) and optimal (42°C) growth temperatures on the systems biology of *Galdieria sulphuraria* for a period spanning more than 100 generations. We chose *G. sulphuraria* as the model organism for this experiment due to its highly streamlined haploid genome (14 Mb, 6800 genes) that evolved out of two phases of strong selection for genome miniaturization (Qiu et al., 2015). In this genomic context, we expected maximal physiological effects of novel mutations, thus possibly reducing the fraction of random neutral mutations. Furthermore, we expected a smaller degree of phenotypical plasticity and hence a more rapid manifestation of adaptation at the genetic level.

MATERIALS AND METHODS

Experimental Design and Sampling

A starting culture of *G. sulphuraria* strain RT22 adapted to growth at 42°C was split into two batches, which were grown separately at 42°C (control condition) and 28°C (temperature stress) for a period spanning 8 months. Bacteria were cultured on agar plates under non-photosynthetic conditions, with glucose (50 mM) as the sole carbon source. To select for fast-growing populations, the five largest colonies of each generation were picked. The samples were propagated across generations by iteratively picking the five biggest colonies from each plate and transferring them to a new plate. The picked colonies were diluted in 1 ml Allen Medium containing 25 mmol glucose. The OD_{750} of the cell suspensions was measured at each re-plating step using a spectrophotometer. Approximately 1,000 cells were streaked on new plates to start the new generation. The remaining cell material was stored at −80°C until DNA extraction. This process was reiterated whenever new colonies with a diameter of 3–5 mm became visible. During the 240 days of this experiment, a total of 181 generations of *G. sulphuraria* RT22 were obtained for the culture grown at 42°C, whereas 102 generations were obtained for *G. sulphuraria* RT22 grown at 28°C.

DNA Extraction and Sequencing

DNA from each sample was extracted using the Genomic-tip 20/G column (QIAGEN, Hilden, Germany), following the steps of the yeast DNA extraction protocol provided by the manufacturer. DNA size and quality were assessed via gel electrophoresis and Nanodrop spectrophotometry (Thermo Fisher Scientific, Waltham, MA, United States). TruSeq DNA PCR-Free libraries (insert size = 350 bp) were generated. The samples were quantified using the KAPA library quantification kit, quality controlled using a 2100 Bioanalyzer (Agilent, Santa Clara, CA, United States), and sequenced on an Illumina (San Diego, CA, United States) HiSeq 3000 in paired-end mode (1 × 150 bp) at the Genomics and Transcriptomics Laboratory of the Biologisch-Medizinisches Forschungszentrum in Düsseldorf, Germany. The raw sequence reads are retrievable from the NCBI's Small Read Archive (SRA) database (Project ID: PRJNA513153).

Read Mapping and Variant Calling

Single nucleotide polymorphisms (SNPs) and insertions/deletions (InDels) were called separately on the dataset using the GATK software version 3.6-0-g89b7209 (McKenna et al., 2010). The analysis was performed according to GATKs best practices protocols (DePristo et al., 2011; Van der Auwera et al., 2013). The untrimmed raw DNA-Seq reads of each sample were mapped onto the genome of *G. sulphuraria* RT22 (NCBI, SAMN10666930) using the BWA aligner (Li and Durbin, 2009) with the –M option activated to mark shorter split hits as secondary. Duplicates were marked using Picard tools[1]. A set of known variants was bootstrapped for *G. sulphuraria* RT22 to build the covariation model and estimate empirical base qualities (base quality score recalibration). The bootstrapping process was iterated three times until convergence was reached (no substantial changes in the effect of recalibration between iterations were observed, indicating that the produced set of known sites adequately masked the true variation in the data). Finally, the recalibration model was built upon the final samples to capture the maximum number of variable sites. Variants

[1] http://broadinstitute.github.io/picard

were called using the haplotype caller in discovery mode with -ploidy set to 1 (*Galdieria* is haploid) and –mbq set to 20 (minimal required Phred score) and annotated using snpEff v4.3i (Cingolani et al., 2012). The called variants were filtered separately for SNPs and InDels using the parameters recommended by GATK (SNPs: "QD < 1.0 | | FS > 30.0 | | MQ < 45.0 | | SOR > 9 | | MQRankSum < −4.0 | | ReadPosRankSum < −10.0," InDels: "QD < 1.0 | | FS > 200.0 | | MQ < 45.0 | | MQRankSum < −6.5 | | ReadPosRankSum < −10.0").

Evolutionary Pattern Analysis

A main goal of this analysis was implementation of a method that enabled discrimination between random variants and variants that may be connected to temperature stress (non-random variants). The following logic was implemented: All variants were transformed to binary code with regards to their haplotype toward the reference genome. When the haplotype was identical to the reference genome, "0" was assigned. Variant haplotypes were assigned "1." Random variants were gained and lost without respect to the sampling succession along the timeline and the different temperature conditions. Consequently, a "fuzzy" pattern of, e.g., "110011| 0000," would indicate a mutation between T_0 and T_1 in the samples taken at 28°C, which was lost in T_3 and regained after T_5. The binary sequence represents the ten samples, six "cold" and four "warm," according to their condition ("cold | warm") and time point of sampling ("28°C_1, 28°C _2, 28°C _3, 28°C _4, 28°C _5, 28°C _6 | 42°C_1, 42°C _3, 42°C _6, and 42°C _9"). Hence, the first six digits denote samples taken at 28°C, the latter four digits those taken at 42°C; "000000| 0101" would represent a mutation in the T_2 sample taken at 42°C that was lost in T_3 and regained in T_4, and "011010| 0101" would represent a variant that does neither with respect to the sampling succession (repeated gain and loss) nor the growth condition of the samples (mutation occurs at both temperatures). By contrast, variants that were gained and fixed in the subsequent samples of a certain growth condition were considered as "non-random variants" that may reflect significant evolutionary patterns. Thus, "111111| 0000" would indicate that a mutation between T_0 and T_1 in the samples taken at 28°C was fixed over the measured period. Similarly, "000000| 0111" would indicate a mutation between T_2 and T_3 in the samples taken at 42°C that was fixed throughout the generations. As such, it was possible to determine all possible pattern combinations for non-random evolutionary patterns. The binary sequence "111111| 1111" represented the case where all ten samples contained a different haplotype when compared to the reference genome. In this specific case, systematic discrepancies between the reference genome and the DNA-Seq reads are the cause of this pattern. Variants following the "111111| 1111" pattern were removed from the dataset.

Data Accession

The DNA sequencing results are described in **Supplementary Table S1**. The Illumina HiSeq3000 raw reads reported in this project have been submitted to the NCBI's Sequence Read Archive (SRA) and are retrievable (FASTQ file format) via BioProject PRJNA513153 and BioSamples SAMN10697271 - SAMN10697280.

Statistical Analysis

Various statistical methods were applied for the different analyses performed in this project. Culture growth was measured for at 28°C ($n = 6$) and at 42°C ($n = 10$). Both datasets failed the Shapiro–Wilk normality test ($p > 0.05$) and showed a visible trend over time. The difference in growth between the populations was tested using the Wilcoxon rank sum test. Further, timepoints along a timeline constitute a dependent sampling approach by which the growth performance of an earlier timepoint is likely to influence the growth performance of a later timepoint.

Trends in growth over the period of this experiment were tested for significance using Jonckheere–Terpstra's test for trends. Enrichment of GO categories as well as *k*-mer enrichment was tested using Fisher's exact test for categorical data, corrected for multiple testing according to Benjamini–Hochberg. The contingency table was set up in such way that the number of times a specific GO was affected by variants was compared with the number of times the same GO was not affected by variants. This category was compared against the "background" consisting of all other GOs affected by variants and all other unaffected GOs. The same methodology was applied for *k*-mer enrichment testing. Differential gene expression based on previously collected data (Rossoni et al., 2018) was calculated with EdgeR (Robinson et al., 2010) implementing the QLF-test in order to address the dispersion uncertainty for each gene (Lun et al., 2016). All samples taken at 28°C were compared against all samples taken at 42°C/46°C.

RESULTS

Culture Growth

Samples grown at 42°C were re-plated 10 times during the 7 months of the experiment due to faster colony growth, whereas cultures growing at 28°C were re-plated only six times (**Figure 1A**). Cultures grown at 42°C achieved an average doubling time of 1.32 days, equivalent to an average growth rate of 0.81/day. Cultures grown at 28°C had an average doubling time of 2.70 days, equivalent to an average growth rate of 0.39/day. This difference in doubling time/growth rate between 28°C and 42°C was significant (non-normal distribution of growth rates, Wilcoxon rank sum test, $p = 0.0002$) (**Figure 1B**). The growth rates reported here were slightly lower than in liquid batch cultures, where growth rates of 0.9/day–1.1/day were measured for heterotrophic cultures grown at 42°C (unpublished data). The changes in growth rate over time were also compared using linear regression (**Figure 1C**). Although the linear regression appears to indicate increasing doubling times in samples grown at 42°C, Jonckheere's test for trends revealed no significant trend in this dataset (Jonckheere-Terpstra, $p > 0.05$). By contrast, samples grown at 28°C gradually adapted to the colder environment

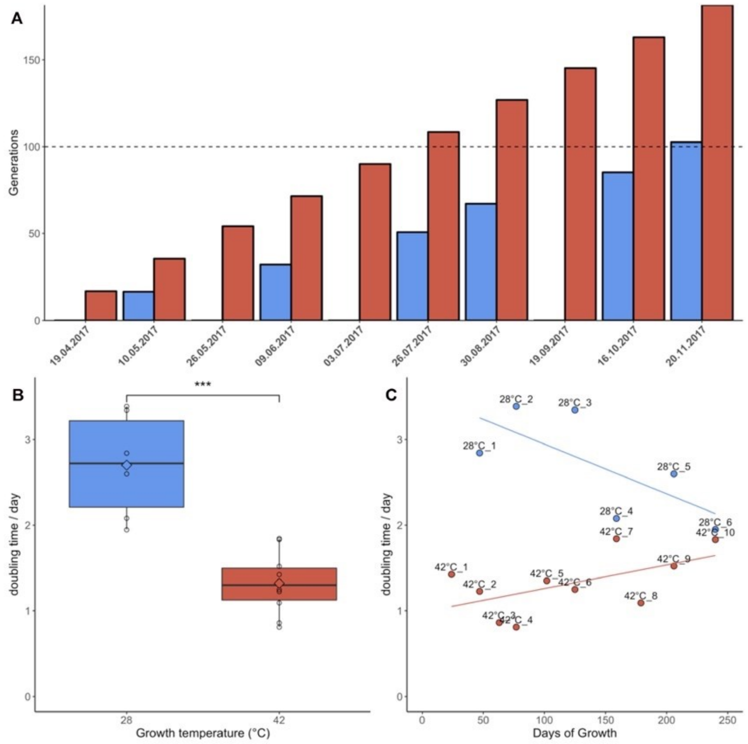
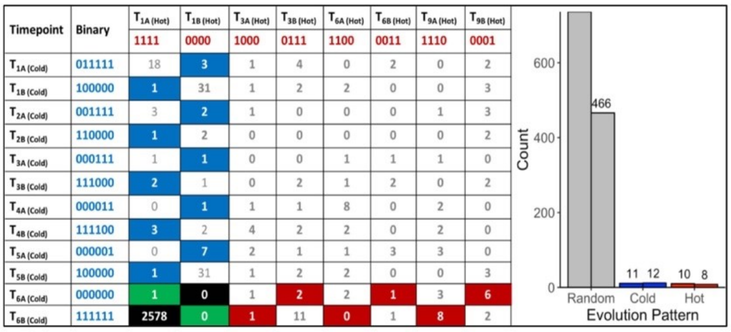


FIGURE 1 | Growth parameters of *Galdieria sulphuraria* RT22. **(A)** Cultures were grown heterotrophically at 28°C (blue) and at 42°C (red) on plates made of 1.5% Gelrite mixed 1:1 with 2× Allen Medium containing 50 mM glucose. The fastest growing colonies were iteratively selected and re-plated over a period of ~7 months until >100 generations were achieved under both conditions. Propagation occurred through picking the five biggest colonies from each plate and transferring them to a new plate. **(B)** The doubling time at 42°C was 1.32 days on average. The doubling time at 28°C was 2.70 days on average. The differences in growth between 42°C and 28°C were significant (Wilcoxon rank sum test, $p = 0.0002$). Cultures grown at 42°C were re-plated 10 times due to faster growth. In comparison, cultures growing at 28°C were re-plated only six times. **(C)** Samples grown at 42°C grew slightly slower over time. By contrast, samples grown at 28°C appeared to decrease their doubling time. While no statistically significant trend could be detected at 42°C, Jonckheere's test for trends reported a significant trend toward faster growth for the populations grown at 28°C.

and significantly (Jonckheere–Terpstra, $p < 0.05$) decreased their doubling time by ~30% during the measured period.

Variant Calling

A total of 470,680,304 paired-end DNA-Seq reads were generated on an Illumina HiSeq 3000 sequencer. Of these, 462,869,014 were aligned to the genome (98.30%) using BWA (**Supplementary Table S1**). The average concordant alignment rate was 99.71%. The average genome coverage was 444 × (min = 294× and max = 579×). At least 95.5% of the sequence was covered with a depth of >20×. GATK's haplotype caller algorithm reported 6,360 raw SNPs and 5,600 raw InDels. The SNPs and InDels were filtered separately according to GATK's best practice recommendations. A total of 1,864 SNPs and 2,032 InDels passed the filtering process. On average, one SNP occurs every 16,177 nt and one InDel every 44,394 nt. Overall, 66.17% of the filtered variants (2578/3896) were classified as background mutations being at variance with the genome reference ("111111| 1111"). The 1243 remaining variants (966 SNPs + 277 InDels) were sorted according to their evolutionary patterns, here called "Random," "Hot," and "Cold" (**Figure 2**); 486/1243 (36.5%) are located in the intergenic region and the other 757/1243 (63.5%) in the genic region, including 5′UTR, 3′UTR, and introns. In addition, 1202/1243 (96.7%) variants followed random gain and loss patterns that do not exhibit relevant evolutionary trajectories (**Figure 2**). The remaining 41 variants were gained and fixed over time, thus representing non-random, evolutionary relevant variants. Twenty-three variants were fixed at 28°C and 18 variants at 42°C. Consequently,



Timepoint	Binary	$T_{1A (Hot)}$	$T_{1B (Hot)}$	$T_{3A (Hot)}$	$T_{3B (Hot)}$	$T_{6A (Hot)}$	$T_{6B (Hot)}$	$T_{9A (Hot)}$	$T_{9B (Hot)}$
		1111	0000	1000	0111	1100	0011	1110	0001
$T_{1A (Cold)}$	011111	18	3	1	4	0	2	0	2
$T_{1B (Cold)}$	100000	1	31	1	2	2	0	0	3
$T_{2A (Cold)}$	001111	3	2	1	0	0	0	1	3
$T_{2B (Cold)}$	110000	1	2	0	0	0	0	0	2
$T_{3A (Cold)}$	000111	1	1	0	0	1	1	1	0
$T_{3B (Cold)}$	111000	2	1	0	2	1	2	0	2
$T_{4A (Cold)}$	000011	0	1	1	1	8	0	2	0
$T_{4B (Cold)}$	111100	3	2	4	2	2	0	2	0
$T_{5A (Cold)}$	000001	0	7	2	1	1	3	3	0
$T_{5B (Cold)}$	100000	1	31	1	2	2	0	0	3
$T_{6A (Cold)}$	000000	1	0	1	2	2	1	3	6
$T_{6B (Cold)}$	111111	2578	0	1	11	0	1	8	2

FIGURE 2 | "Random" and "Non-random" variant acquisition patterns. "Non-random" variants were defined as mutations gained at some point during growth at 42°C or 28°C, and fixed in the genome of *G. sulphuraria* RT22 during the remaining time points. All variants were translated into binary code according to their haplotype relative to the reference genome. "Cold": variant was obtained and fixed at 28°C. "Hot": variant was obtained and fixed at 42°C. **Left**: Evolutionary patterns and their frequencies. In the specific case of patterns "111111| 0000" and "000000| 1111," a variant was already gained before the first sampling time point. Hence, it was not possible to determine the condition at which the variant was gained. "Background" mutations represent the cases where the sequence of all samples was in disagreement with the reference genome "111111| 1111." The remaining combinations were considered as "random" evolution patterns. Here, variants were gained but not fixated in the subsequent samples of the same growth condition. The numbers in the boxes indicate the count of a specific pattern. **Right**: Count by variant type. The right column of each category indicates the number of variants located in the intergenic space. The left column counts the number of variants located in the coding sequence.

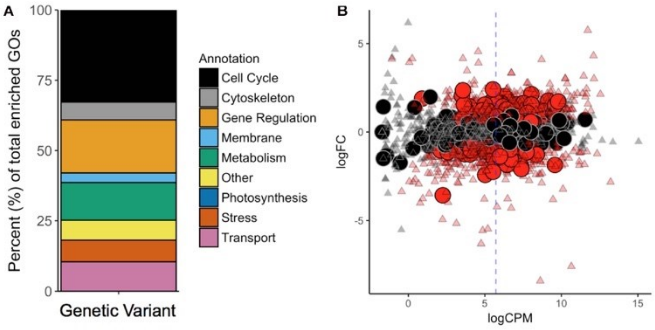


FIGURE 3 | (A) GO-Term (GO) enrichment analysis revealed the cellular functions most affected by mutations. Each GO was manually revised and attributed to one of the nine categories contained in the legend. **(B)** Differential gene expression in *G. sulphuraria* RT22 orthologs in *G. sulphuraria* 074W (reciprocal best blast hit), here measured as log -fold change (logFC) vs. transcription rate (logCPM). Differentially expressed genes are colored red (quasi-likelihood *F*-test, Benjamini-Hochberg, $p <= 0.01$). Genes affected by variants are shown by large circles. Genes without significant differential expression are represented by triangles. The blue dashes indicate the average logCPM of the dataset. The orthologs in *G. sulphuraria* 074W of genes affected by variance in *G. sulphuraria* RT22 did not show more, or less, differential expression under fluctuating temperature.

23 variants (1.9%) were attributed to the "Cold" pattern (11 intergenic, 12 genic) and 18 variants to the "Hot" pattern (1.4%).

GO Enrichment-Based Overview of Cellular Functions Most Affected by Variants

The vast majority of the 757 genic variants was not fixed over time and did not follow consistent evolutionary patterns (**Figure 2**). However, the frequency at which genes and gene functions were affected by mutation can serve as an indicator of the physiological processes most affected by evolutionary pressure at 28°C and 42°C. Here, we analyzed the functional annotations of the 757 variants located on 429 genes (6.1% of the transcriptome) using GO-Term (GO) enrichment analysis. A total of 1602 unfiltered GOs were found within the genes affected by variants (27.3% of all GOs in *G. sulphuraria* RT22), of which 1116 were found at least twice in the variant dataset. Of those, 234 of the GOs were significantly enriched (categorical data, "native" vs. "HGT," Fisher's exact test, Benjamini–Hochberg, $p \leq 0.05$).

To contextualize the function in broader categories, we manually sorted all significantly enriched GOs into the following ten categories: "Cell Cycle," "Cytoskeleton," "Gene Regulation," "Membrane," "Metabolism," "Photosynthesis," "Stress/Signaling," "Transport," "Other," and "NA" (**Figure 3A**). GO terms belonging to the "NA" category were considered meaningless and excluded from the analysis [e.g., "cell part" (GO:0044464), "biological process" (GO:0008150), "binding" (GO:0005488), and "ligase activity" (GO:0016874)].

Cell Cycle

Functions related to the cell cycle accounted for 79/234 (33.8%) of the enriched GOs. Mitosis was affected at every stage: initiation (e.g., "positive regulation of cell proliferation," GO:0008284, p = 0.0024; "re-entry into mitotic cell cycle," GO:0000320, p = 0.0405), DNA replication (e.g., "DNA replication, removal of RNA primer," GO:0043137, p = 0.0001; "ATP-dependent 5′-3′ DNA helicase activity," GO:0043141, p = 0.014), prophase ("preprophase band," GO:0009574, p = 0.0270), metaphase (e.g., "attachment of mitotic spindle microtubules to kinetochore," GO:0051315, p = 0.0142), anaphase (e.g., "mitotic chromosome movement toward spindle pole," GO:0007079, p = 0.0134), and telophase ("midbody," GO:0030496, p = 0.0404). Mutations also accumulated in genes controlling cell cycle checkpoints of mitosis (e.g., "positive regulation of mitotic metaphase/anaphase transition, GO:0045842, p = 0.0270; "mitotic spindle assembly checkpoint," GO:0007094, p = 0.0441).

Genes with functions involved in cell differentiation and maturation of *Galdieria* were also affected significantly by microevolution during organellogenesis (e.g., "regulation of auxin mediated signaling pathway," GO:0010928, p = 0.0012; "phragmoplast," GO:0009524, p = 0.0405; "xylem and phloem pattern formation," GO:0010051, 0.0012), cell polarity (e.g., "establishment or maintenance of epithelial cell apical/basal polarity," GO:0045197, p = 0.0012; "growth cone," GO:0030426, p = 0.0096), and subcellular compartmentalization and localization ("Golgi ribbon formation," GO:0090161, p = 0.0404; "establishment of protein localization," GO:0045184, p = 0.0124). Interestingly, some transcriptional regulators of cell growth seem to be conserved across the eukaryotic kingdom. GOs such as "branching involved in open tracheal system development" (GO:0060446, p = 0.0012) and "eye photoreceptor cell development" (GO:0042462, p = 0.0093) were also found, indicating high amino acid sequence similarity within this category. Further, temperature stress altered genes with functions involved in cell death (e.g., "cell fate determination," GO:0001709, p = 0.0012; "Wnt signalosome," GO:1990909, p = 0.0405).

Gene Regulation

Maintenance of steady and balanced reaction rates across cellular systems is essential for cell survival and poses a major challenge when an organism is confronted with changes in temperature (D'Amico et al., 2002). In this context, the second largest category within the enriched GO terms (49/234, 20.9%) was related to gene regulation. Besides cell cycle control, thermal adaptation and evolution was orchestrated predominantly through mutations in genes involved in controlling the expression profiles of other genes ("gene expression," GO:0010467, p = 0.0118). Also, we found a significant proportion of mutations affecting genes linked to the epigenetic control of gene expression, which can occur through methylation of DNA ("hypermethylation of CpG island," GO:0044027, p = 0.0086), as well as modulation of chromatin density and histone interactions that change the accessibility of whole genomic regions to transcription ("H4 histone acetyltransferase activity," GO:0010485, p = 0.0040) (Jenuwein and Allis, 2001; Bird, 2002). Further, variants may have altered RNA polymerase efficiency (e.g., "RNA polymerase II transcription factor binding," GO:0001085, p = 0.0020), mRNA processing (e.g., "regulation of RNA splicing," GO:0043484, p = 0.0025), post-transcriptional silencing (e.g., "RNA interference," GO:0016246, p = 0.0093) as well as alteration of ribosome structure components (e.g., "structural constituent of ribosome," GO:0003735, p = 0.0336) and rRNA methylation components (e.g., "rRNA methylation," GO:0031167, p = 0.0036). In this regard, GO terms linked to posttranslational protein modification were also enriched ("positive regulation of peptidyl-threonine phosphorylation," GO:0010800, 0.0086; "N-terminal peptidyl-methionine acetylation," GO:0017196, p = 0.0007).

Cytoskeleton

Microtubule are long polymers of tubulin that are constituents of the cytoskeleton of every eukaryote. They play a central role in intracellular organization, stability, transport, organelle trafficking, and cell division (Brouhard and Rice, 2018). Because they associate spontaneously, microtubular assembly (e.g., "microtubule nucleation," GO:0007020, p = 0.0039) and disassembly are mostly driven by tubulin concentrations at the beginning and the end of microtubules once a critical microtubule size is reached (Voter and Erickson, 1984). However, the first steps of microtubule assembly are kinetically unfavorable. Cells solve this issue by using γ-tubulin ring complex as a template (e.g., "tubulin complex," GO:0045298, p = 0.0031). The reaction equilibrium between tubulin polymerization and monomerization is temperature-dependent and requires accurate regulation (e.g., "tau-protein kinase activity," GO:0050321, 0.0086). Shifting temperatures from 37°C to 25°C leads to massive microtubular dissociation in homoeothermic species (Himes and Detrich, 1989). Additionally, tubulin adaptations toward lower temperatures have been observed at the level of DNA sequence as well as at the epigenetic level in psychrophilic organisms (Detrich et al., 2000). Microtubule metabolism and its role in cellular physiology accounted for 13/234 (5.6%) of the enriched GOs.

Membranes and Transport

Another major component that is also influenced by temperature is cell integrity with regards to membrane fluidity (5/234 enriched GOs, 2.1%) and transport (25/234 enriched GOs, 10.7%). Cell membranes are selectively permeable and vital for compartmentation and electric potential maintenance.

In this context, *Galdieria* is able to maintain near-neutral cytosolic pH against a 10^6-fold H^+ gradient across its plasma membrane (Gross, 2000). Membranes maintain a critical range of viscosity to be able to incorporate molecules and transport substrates and nutrients. The fluidity of a membrane is mainly determined by its fatty acid composition. Changes in temperature lead to changes in fatty acid composition, which in turn affect hydrophobic interactions as well as the stability and functionality of membrane proteins and proteins anchored to membranes. Here, we measured a significant enrichment in genes with functions connected to membrane lipid bilayers (e.g., "membrane," GO:0016020, p = 0.0002; "mitochondrial inner membrane," GO:0005743, p = 0.0023) as well as membrane-associated proteins (e.g., "integral component of membrane," GO:0016021, $p < 0.0001$), transporters (e.g., "amino acid transmembrane transporter activity," GO:0015171, p = 5.79568E-06), and transport functions (e.g., "transmembrane transport," GO:0055085, p = 0.0001; "cation transport," GO:0006812, p = 0.0028). Furthermore, temperature imposes significant restrictions to vesicles, which play a central role in molecule trafficking between organelles and in endocytosis. Vesicle formation in particular appears to be affected by temperature (e.g., "vesicle organization," GO:0016050, p = 0.0012; "clathrin-coated endocytic vesicle membrane," GO:0030669, p = 0.0025).

Stress and Signaling

Cell signaling comprises the transformation of information, such as environmental stress, to chemical signals that are propagated and amplified through the system where they contribute to the regulation of various processes (e.g., "response to stress," GO:0006950, p = 0.0051; "hyperosmotic response," GO:0006972, p = 0.0039; and "ER overload response," GO:0006983, p = 0. 0040). Here, we found a total of 18/234 GOs (7.7%) derived from genes involved in cell signaling upon which temperature changes appeared to exhibit significant evolutionary pressure driving the accumulation of variants. A broad array of receptors (G-protein coupled, tyrosine kinases, and guanylate cyclases) performs signal transduction through phosphorylation of other proteins and molecules. The signal acceptors, in turn, influence second messengers and further signaling components that affect gene regulation and protein interactions. GO annotations indicate involvement of temperature in genes coding for receptors (e.g., "activation of protein kinase activity," GO:0032147, p = 4.95227E-05; "protein serine/threonine/tyrosine kinase activity," GO:0004712, p = 0.00045547; and "protein autophosphorylation," GO:0046777, p = 1.53236E-06) as well as in genes coding for the signal acceptors ("stress-activated protein kinase signaling cascade," GO:0031098, p = 6.45014E-06; "cellular response to interleukin-3," GO:0036016, p = 5.05371E-06; and "regulation of abscisic acid-activated signaling pathway," GO:0009787, p = 0.006712687).

Metabolism

Maintaining metabolic homeostasis is paramount for organism survival. The efficiency, speed, and equilibrium of metabolic pathways are modulated by enzymes and the specific kinetics of each reaction. Whereas microorganisms are not capable of controlling the amount of free enthalpy in their systems (chemical equilibriums are temperature-dependent, $\Delta G = -RT \ln k$), they are able to actively adjust their metabolic rates by regulating the amount of available enzyme ("Gene Expression"). Passively, mutations can alter enzyme structure, thereby adjusting the affinity of enzymes toward ligands. Variants affecting the genetic code of genes attributed to this category influence a broad variety of metabolic pathways (e.g., "cellular aromatic compound metabolic process," GO:0006725, p = 0.0030 and "amine metabolic process," GO:0009308, $p < 0.0001$) in both anabolism (e.g., "peptidoglycan biosynthetic process," GO:0009252, p = 0.0015 and "glycerol biosynthetic process," GO:0006114, p = 0.0086), and catabolism (e.g., "glycosaminoglycan catabolic process," GO:0006027, p = 0.0011). In spite of pronounced changes in gene expression of metabolic enzymes during short-term cold stress in *G. sulphuraria* 074W (Rossoni et al., 2018) and *Cyanidioschyzon merolae* 10D (Nikolova et al., 2017), microevolution of genes directly involved in metabolic steps appeared to play a minor role in long-term temperature adaptation (34/234 GOs, 14.5%).

Photosynthesis

The majority of photosynthetic light reactions are catalyzed by enzymes located in the photosynthetic thylakoid membranes. Hence, photosynthesis is based upon temperature-dependent proteins located in temperature-dependent membranes (Yamori et al., 2014). Abnormal temperatures affect the electron transport chain between the various components of the photosynthetic process (Hew et al., 1969). If the electron transport chain between photosystem I (PSI) and photosystem II (PSII) is uncoupled, electrons are transferred from PSI to oxygen instead of PSII. This process is also known as PSII excitation pressure and leads to a boost of reactive oxygen species. Long-term microevolution did not appear to significantly affect the photosynthetic apparatus of *G. sulphuraria* RT22 (3/234, 1.3%), likely because the experiment was performed under heterotrophic conditions in continuous darkness.

Variant Hotspots and Non-random Genic Variants

To further investigate the temperature adaptation of *G. sulphuraria* RT22, we selected candidate genes for closer analysis using two different approaches. First, we assumed that high mutation rates in a specific gene reflect increased selective force upon its function and regulation. To identify potential targets of temperature-dependent microevolution, we searched for "variant hotspots," here defined as the 99th percentile of genes most affected by variants. We computed variant number-dependent Z-scores for each gene and extracted genes with a Z-Score > 2.575. This procedure led to identification of seven genes, so-called "variant hotspots," containing at least seven independent variants per gene. Next, we extracted 41 variants that followed non-random evolutionary patterns, here defined as the gain of a variant and its fixation in the subsequent samples that was exclusive to either 28°C or 42°C (**Figure 2**). Eighteen variants followed an evolutionary pattern defined as "Hot" (1.36%) and 23 variants followed an evolutionary pattern defined as "Cold" (1.59%). These non-random evolutionary patterns describe the gain of a variant and its fixation over time either in the 42°C dataset ("Hot," e.g., 000000| 0001, 000000| 0011), or in the 28°C dataset ("Cold," e.g., 000001| 0000, 000011| 0000), respectively. The underlying assumption was that this subset represented beneficial mutations. Synonymous variants were removed from further analysis. As a result, we obtained 13 genes that followed non-random evolutionary patterns (16 non-synonymous variants). An individual functional characterization of each gene is contained in the **Supplementary Material** (**Supplementary Listing S1A** for "Variant Hotspots" and **Supplementary Listing S1B** for "Further Non-random Genic Variants").

The gene function of the selected temperature-dependent gene candidates broadly replicated the results of the GO enrichment analysis. Here, we found multiple enzymes involved in cell cycle control and signaling, e.g., an oxidase of biogenic tyramine (Gsulp_RT22_67_G1995), an armadillo/beta-catenin repeat family protein (Gsulp_RT22_107_G5273), the GTPase-activating ADP-ribosylation factor ArfGAP2/3 (Gsulp_RT22_82_G3036), and a peptidylprolyl *cis-trans* isomerase (Gsulp_RT22_64_G1844). Other candidate genes were involved in transcription and translation, e.g., a NAB3/HDMI transcription termination factor (Gsulp_RT22_83_G3136), or in ribosomal biogenesis (Gsulp_RT22_112_G5896, 50S ribosomal subunit) and required cochaperones (Gsulp_RT22_99_G4499, Hsp40). Three candidate genes were solute transporters (Gsulp_RT22_67_G2013, Gsulp_RT22_118_G6841, Gsulp_RT22_67_G1991). Most interestingly, two genes connected to temperature stress were also affected by mutations. An error-prone iota DNA-directed DNA polymerase (01_Gsulp_RT22_79_G2795), which promotes adaptive point mutation as part of the coordinated cellular response to environmental stress, was affected at 28°C (Napolitano et al., 2000; McKenzie et al., 2001), as well as the 2-phosphoglycerate kinase, which catalyzes the first metabolic step of the compatible solute cyclic 2,3-diphosphoglycerate, which increases the optimal growth temperature of hyperthermophile methanogens (Santos and da Costa, 2002; Roberts, 2005).

HGT Candidates Are Not Significantly Involved in Temperature Microevolution

Horizontal gene transfer has facilitated the niche adaptation of *Galdieria* and other microorganisms by providing adaptive advantages (Schonknecht et al., 2013; Schönknecht et al., 2014; Foflonker et al., 2018). Five of the total 54 HGT gene candidates in *G. sulphuraria* RT22 gained variants (Rossoni et al., 2019). We tested whether a more significant proportion of HGT candidates gained variants in comparison to native genes. This was not the case: HGT candidates did not significantly differ from native genes (categorical data, Fisher's exact test, $p < 0.05$). Of the HGT candidates, only Gsulp_RT22_67_G2013, a bacterial/archaeal APC family amino acid permease potentially involved in the saprophytic lifestyle of *G. sulphuraria*, accumulated a significant number of mutations (12 variants).

Genes Involved in Differential Expression Were Not Targeted by Mutation

We tested if the 6.1% of genes that gained variants were also differentially expressed during a temperature-sensitive RNA-Seq experiment in *G. sulphuraria* 074W, where gene expression was measured at 28°C and 42°C (Rossoni et al., 2018). Of the 6982 sequences encoded by *G. sulphuraria* RT22, 4569 were successfully matched to an ortholog in *G. sulphuraria* 074W (65.4%); 342 were orthologs to a variant-containing gene, representing 79.7% of all genes containing variants in *G. sulphuraria* RT22. The dataset is representative (Wilcoxon rank sum test, Benjamini–Hochberg, $p < 0.05$, no differences in the distribution of variants per gene due to the sampling size). Based on this result, 36.3% of the variant-containing genes were differentially expressed. By contrast, 40.1% of the genes unaffected by variants were differentially expressed (**Figure 3B**). The difference between the two subsets was not significant (categorical data, Fisher's exact test, $p < 0.05$). Hence, genes affected by variance during this microevolution experiment did not react more, or less, pronouncedly to fluctuating temperature.

Intergenic Variant Hotspots

Mutations that affect gene expression strength and pattern are a common target of evolutionary change (Barbosa-Morais et al., 2012). Intergenic DNA encodes *cis*-regulatory elements, such as promoters and enhancers, that constitute the binding sites of transcription factors and, thus, affect activation and transcriptional rate of genes. Promoters are required for transcriptional initiation but their presence alone results in minimal levels of downstream sequence transcription. Enhancers, which can be located either upstream, downstream, or distant from the genes they regulate, are the main drivers of gene transcription intensity and are often thought to be the critical factors of *cis*-regulatory divergence (Wray, 2007). Further, epigenetic changes can lead to heritable phenotypic and physiological changes without the alteration of the DNA sequence (Dupont et al., 2009). As a consequence of its evolutionary history, the genome of *G. sulphuraria* is highly deprived of non-functional DNA (Qiu et al., 2015). Here, we performed variant enrichment analysis of the intergenic space based on *k*-mers ranging from *k*-mer length 1 (4 possible combinations, A| C| G| T) to *k*-mer length 10 (1,048,576 possible combinations) to identify intergenic sequence patterns prone to variant accumulation (**Table 1**). The enriched *k*-mers were screened and annotated against the PlantCARE (Lescot et al., 2002) database containing annotations of plant *cis*-acting regulatory elements. Only partial hits were found, possibly due to the large evolutionary distance between plants and red algae, more specifically the *Galdieria* lineage, which might explain the divergence between *cis*-regulatory sequences (Wittkopp and Kalay, 2012). The sequence "CG," which is the common denominator of CpG islands (Deaton and Bird, 2011), was found enriched within the *k*-mer set of length 2. In addition, partial hits to the PlantCARE database with a *k*-mer length >5 were considered as potential hits. Using this threshold, we found three annotated binding motifs, the OCT (octamer-binding motif) (Zhao, 2013), RE1 (Repressor Element 1) (Paonessa et al., 2016), and 3-AF1 (accessory factor binding sites) (Scott et al., 1996; Rhen and Cidlowski, 2005).

DISCUSSION

Growth Rates Adapt to Temperature

In this study, we subjected two populations of *G. sulphuraria* RT22 to a temperature-dependent microevolution experiment for 7 months. One culture was grown at 28°C, representing cold stress, and a control culture was grown at 42°C. This experiment aimed to uncover the genetic acclimation response to persistent stress, rather than the short-term acclimation response of *G. sulphuraria* to cold stress (Rossoni et al., 2018). We performed genomic re-sequencing along the timeline to measure changes in the genome sequence of *G. sulphuraria* RT22. After 7 months, corresponding to ~170 generations of growth at 42°C and ~100 generations of growth at 28°C, the cold-adapted cultures decreased their doubling time by ~30%. The control cultures maintained constant growth, although a trend to slower growth might occur (**Figure 1**). A similar increase in the growth rate was also observed in the photoautotrophic sister lineage of *Cyanidioschyzon*, where cultures of *Cyanidioschyzon merolae* 10D were grown at 25°C for a period of ~100 days, albeit under photoautotrophic conditions. This study found that the cold-adapted cultures outgrew the control culture at the end of the experiment (Nikolova et al., 2017). While faster doubling times at 28°C can be attributed to gradual adaptation to the suboptimal growth temperatures, we may only speculate about the causes leading to slower growth in the control condition (42°C). Perhaps *G. sulphuraria* RT22, which originated from the Rio Tinto river near Berrocal (Spain), may be able to thrive at high temperatures, but not for such a prolonged period.

Cultures Grown at 28°C Accumulate Twice the Number of Mutations as Compared to Controls

We identified 1243 filtered variants (966 SNPs + 277 InDels), of which 757 (63.5%) were located on the coding sequence of 429 genes and 486 (36.5%) in the intergenic region. The mutation rate was estimated to be 2.17×10^{-6}/base/generation for samples grown at 28°C and 1.10×10^{-6}/base/generation for samples grown at 42°C, which we interpret as an indication of greater evolutionary stress at 28°C. Hence, suboptimal growth temperatures constitute a significant stress condition and promote the accumulation of mutations. In comparison, mutation rates in other microevolution experiments were 1.53×10^{-8}/base/generation–6.67×10^{-11}/base/generation for the unicellular green freshwater alga *Chlamydomonas reinhardtii* and 5.9×10^{-9}/base/generation in the green plant *Arabidopsis thaliana* (Ness et al., 2012; Sung et al., 2012; Perrineau et al., 2014). The 100-fold higher evolutionary rates in comparison to *C. reinhardtii* might result from the selective strategy employed in this experiment (only the five biggest colonies were selected to start the next generations). Although the cold-stressed samples accumulated twice as many mutations per generation in comparison to the control condition, the number of gained variants over the same period was higher in the 42°C cultures due to faster growth rates.

Cell Cycle and Transcription Factors Are the Main Drivers of Temperature Adaptation

The impact of temperature-driven microevolution on the cellular functions of *G. sulphuraria* RT22 was analyzed using GO enrichment analysis. More than 75% of the 234 significantly enriched GOs affected genes functions involved in the processes of cell division, cell structure, gene regulation, and signaling. In short, the cellular life cycle appears to be targeted by variation at any stage starting with mitosis, morphogenesis, and finishing with programmed cell death. By contrast, genes directly affecting metabolic processes were less affected by mutation and made up only 10% of the enriched GOs. These observations were also confirmed through the functional annotation of the seven genes most affected by variants ("variant hotspots") as well as the 13 genes carrying non-synonymous variants with non-random evolutionary patterns.

The Intergenic Space in Galdieria Is Equally Affected as Coding Regions

Historically, intergenic DNA has frequently been considered to represent non-functional DNA. It is now generally accepted that mutations affecting intergenic space can heavily influence the expression intensity and expression patterns of genes. Variants altering the sequence of *cis*-regulatory elements are a common source of evolutionary change (Wittkopp and Kalay, 2012). Due to two phases of genome reduction (Qiu et al., 2015), the genome of *Galdieria* is highly streamlined and the intergenic space accounts for only 36% of its sequence. As a consequence, it is assumed that *G. sulphuraria* lost non-functional intergenic regions that are affected by high random mutation rates in other organisms. In this experiment, variants accumulated proportionally between the genic and the intergenic space, which we interpret as an indication of high relevance of the non-coding regions in *Galdieria*. K-mer analysis revealed significant enrichment of variants occurring in CpG islands. CpG islands heavily influence transcription on the epigenetic level through methylation of the cytosines. In mammals, up to 80% of the cytosines in CpG islands can be methylated, and heavily influence epigenetic gene expression regulation. Furthermore, they represent the most common promotor

TABLE 1 | *K*-mer screen of intergenic regions.

K-mer size	Sequence	Variant	Non-variant	Fisher's p (BH)	Annotation	Sequence	PlantCARE Comments
2	CG	45	81329	5.8819E-06	CpG Island	CG	NA
					>30 Hits	Various	Various
2	TT	71	483528	3.4026E-05	>100 Hits	Various	Various
3	CG	13	11612	0.0117	Part of: JERE	AGACCGCC	Jasmonate and elicitor-responsive element
					Part of: ABRE	Various	ACGT-containing ABA Response Element
					Part of: C-box	ACGAGCACCGCC	*Cis*-aclinq regulatory-element involved in light responsiveness
					Part of: Chs-unit	Various	Various
					Part of: RbcS-CMA7c	Various	Various
					Part of:F2Fb	TTTGCCGC	G1-M transitton of cell cycle
					Part of: GC-motiv	Various	Various
					Part of: GC-repeat	GGCCTCGCCACG	?
					Part of: Box-C	TATTACCTGGTCACG CTTTCATA	*Cis*-actinq element involved in the basal expression of me PR1 genes
					Part of: GRA	CACTGGCCGCCC	Important for transcription in leaves
					Part of: OCT	CGCGGATC	Part of the histone H4 gene promoter, which can express H4C7 under inducing or non-inducing conditions. Cell division is accompanied by a concomitant activation of histone genes which produces equivalent amounts of core histones to be incorporated wilh newly replicated DNA into chromatin
					Part of: RE1	GGGCGCGGAACA AGGATCGGC GCGCCACGCC	Repressiny element
3	TTT	33	183031	0.0117	Part of >30 Elements	Various	Various
3	GCG	12	11649	0.0260	Part of: RE1	GGGCGCGGAACA AGGATCGGC GCGCCACGCC	Repressing element
					Part of:Unnamed_7	TTTCTTGCGTT TTTTG GCATAT	?
					Part of:GTGGC-motif	Variuos	Part of the rbcA conserved DNA module array (rbcA-CMA1) involved in light responsiveness
					Part of: E2F	AGTGGCGGNN NNNTTTGAA	G1-M transition of cell cycle
					Partof:As-1-Box	TGACGAATG CGATGACC	Involved in various stress-responses correlated with auxin: salicylic acid and methyl jasmonate

(Continued)

TABLE 1 | Continued

K-mer size	Sequence	Variant	Non-variant	Fisher's p (BH)	Annotation	Sequence	PlantCARE Comments
					Part of: ABRE	Various	ACGT-containing ABA Response Element
					Part of: GC-motiv	Various	Enhancer-like element involved in anoxic specific inducibility
					Part of: Sp 1 Motif I	CGCCGG	Involved in light responsiveness
					Part of: Re2f-1	GCGGGAAA	Putative E2F binding sties in the rice PCNA promoter mediate activation in actively dividing cells
					Pan of: ACE	GCGACGTACC	*Cis*-acting element in promoter and enhancer; involved in light responsiveness
					Part of: I-Box	Various	Part of a light responsive element
					Part of: OCT	CGCGGATC	Part of the histone H3 gene promoter, which tan express H3C4 under inducing or non-inducing conditions. Cell division is accompanied by a concomitant activation of histone genes which produces equivalent amounts of core histones to be incorporated with newly replicated DNA into chromatin
					Part of:CHS unit 11	AGTCGTGGCCA TCCATCCTCCCGTCA ATGGACCTAACCCGC	sequence consisting of three modules: enough to make light inducing possible
					Part of:RbcS-CMA7c	ACGCAGTGTGTG GAGGAGCA	Part of a light responsive element
5	CGCGG	4	230	0.0139	Pan of: RE1	GGGCGCGGAACA AGGATCGG CGCGCCACGCC	Repressing element
					Part of: OCT	CGCGGATC	Part of the histone H4 gene promoter, which can express H4C7 under inducing or non-inducing conditions. Cell division is accompanied by a concomitant activation of histone genes which produces equivalent amounts of core histones to be incorporated with newly replicated0020DNA into chromatin
5	CATAT	14	6311	0.0196	Part of:I-Box	cCATATCCAAT	Part of a liqhl responsive element
					Part of:Unnamed_7	TTTCTTGCGTTTTTT GGCATAT	?
5	GAGAG	11	4444	0.0336	Part of: 3-AF1	TAAGAGAGGAA	Light responsive element
6	CGCGGA	4	57	0.0005	Part of: RE1	GGGCGCGGAACAA GGATCGGC GCGCCACGCC	Repressing element
					Part of: OCT	CGCGGATC	Part of the histone H4 gene promoter, which can express H4C7 under inducing or non-inducing conditions. Cell division is accompanied by a concomitant activation of histone genes which produces equivalent amounts of core histones to be incorporated with newly replicated DNA into chromatin
6	AGAGAG	10	2182	0.0067	Part of: 3-AF1	TAAGAGAGGAA	Liqht responsive element
6	AGCGCG	3	46	0.0067	Part of: GC-motif	AGCGCGCCG	?
6	CGGGAT	4	195	0.0112	NA	NA	NA

(Continued)

TABLE 1 | Continued

K-mer size	Sequence	Variant	Non-variant	Fisher's p (BH)	Annotation	Sequence	PlantCARE Comments
6	GAGAGA	11	2273	0.0024	NA	NA	NA
6	GCGCGG	3	66	00112	Part of RE1	GGGCGG	Repressing element
6	TCGGGA	4	210	0.0114	NA	NA	NA
6	GAGCGC	3	110	0.0409	NA	NA	NA
7	GAGAGAG	11	1071	<0 0001	NA	NA	NA
7	GCGCGGA	3	3	<0.0001	Part of:RE1	GGG{CGCGG} AACAAGG...	Repressing element
7	CGCGGAC	3	6	0.0003	NA	NA	NA
7	AGCGCGG	3	7	0.0003	NA	NA	NA
7	AGAGAGA	10	1284	0 0007	NA	NA	NA
7	GAGCGCG	3	13	0.0009	NA	NA	NA
7	TCGGGAT	4	80	0.0020	NA	NA	NA
7	*CCTTCCC*	4	101	0.0042	NA	NA	NA
7	TTACGAG	4	103	0 004 2	NA	NA	NA
7	CGAGACC	3	33	0.0066	NA	NA	NA
7	TAGAGAG	5	288	0.0084	NA	NA	NA
7	AATCAAG	6	523	0.0092	NA	NA	NA
7	TGAGCGC	3	41	0.0094	NA	NA	NA
7	CGGGATT	3	64	0.0190	NA	NA	NA

The non-coding sequence of Galdieria sulphuraria RT22 was screened using k-mers spanning 1–10 nucleotides. The k-mers of each length were tested for variant enrichment (Fisher's exact test). Only significantly enriched k-mers are shown here. K-mers longer than eight nucleotides did not produce any database hits and are not shown. K-mer size: length of the analyzed k-mer. Sequence: the sequence of the k-mer. Variant: Number of k-mers with specific sequence affected by variants. Non-Variant: Number of k-mers with a specific sequence not affected by variants. Fisher's p (BH): Benjamini–Hochberg post hoc corrected p-value of Fisher's enrichment test. Annotation: PlantCARE identifier (ID) of regulatory element. "Part of:" indicates a partial hit of the k-mer sequence to the database entry. ID Sequence: Full sequence of the regulatory element. PlantCARE Comments: Additional information provided by PlantCARE.

type in the human genome, affecting transcription of almost all housekeeping genes and the portions of developmental regulator genes (Jabbari and Bernardi, 2004; Saxonov et al., 2006; Zhu et al., 2008). Hence, temperature adaptation is not only modulated through accumulation of mutations in the genetic region but equally driven by the alteration of gene expression through epigenetics and mutations affecting the non-coding region.

CONCLUSION

We show here that the significant growth enhancement of samples grown at 28°C over more than 100 generations was driven mainly by mutations in genes involved in the cell cycle, gene regulation, and signal transfer, as well as mutations that occurred in the intergenic regions, possibly changing the epigenetic methylation pattern and altering the binding specificity to *cis*-regulatory elements. Our data indicate the absence of a few specific "key" temperature switches. Rather, it appears that the evolution of temperature tolerance is underpinned by a systems response which requires the gradual adaptation of an intricate gene expression network and deeply nested regulators (transcription factors, signaling cascades, and *cis*-regulatory elements). Our results also emphasize the difference between short-term acclimation and long-term adaptation with regard to temperature stress, highlighting the multiple facets of adaptation that can be measured using different technologies. The short-term stress response of *G. sulphuraria* and the long-term stress response in *C. merolae* were quantified using transcriptomic and proteomic approaches, respectively (Nikolova et al., 2017; Rossoni et al., 2018). At the transcriptional and translational levels, both organisms reacted toward maintaining energetic and metabolic homeostasis by increased protein concentrations, adjusting the protein folding machinery, changing degradation pathways, regulating compatible solutes, remodeling of the photosynthetic machinery, and tuning the photosynthetic capacity. SNP and InDel calling revealed underlying regulators mostly affected by variation which are potential drivers of altered transcript and protein concentrations and ultimately determine physiology and phenotype. Some issues, however, remained unresolved. Is the observed growth phenotype permanent, or is it mostly derived from epigenetic modification which could be quickly reversed? We also did not investigate the temperature-dependent differential splicing (Bhattacharya et al., 2018; Qiu et al., 2018) apparatus in *Galdieria*, or the impact of non-coding RNA elements, both of which may provide additional layers for adaptive evolution (van Bakel et al., 2010).

DATA AVAILABILITY

The datasets generated for this study can be found in NCBI Small Read Archive, PRJNA513153.

AUTHOR CONTRIBUTIONS

AR and AW conceived the study and developed the experimental design, edited and drafted the final version of the manuscript. AR did all the experimental work, data analysis, and drafted the manuscript.

FUNDING

This work was supported by the Deutsche Forschungsgemeinschaft under Germany's Excellence Strategy – EXC-2048/1 – Project ID: 390686111 and DFG-ANR grant Mo-MiX (WE 2231/21-1).

ACKNOWLEDGMENTS

We thank the "Genomics & Transcriptomics Laboratory" of the "Biologisch-Medizinischen Forschungszentrum" (BMFZ) at the Heinrich Heine University Düsseldorf (Germany) for technical support and conducting the Illumina sequencing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2019.00927/full#supplementary-material

REFERENCES

Barbosa-Morais, N. L., Irimia, M., Pan, Q., Xiong, H. Y., Gueroussov, S., Lee, L. J., et al. (2012). The evolutionary landscape of alternative splicing in vertebrate species. *Science* 338, 1587–1593. doi: 10.1126/science.1230612

Barcytė, D., Elster, J., and Nedbalová, L. (2018). Plastid-encoded rbcL phylogeny suggests widespread distribution of *Galdieria phlegrea* (Cyanidiophyceae, Rhodophyta). *Nord. J. Bot.* 36, e01794. doi: 10.1111/njb.01794

Benton, M. J. (2008). *When Life Nearly Died: The Greatest Mass Extinction of all Time.* London: Thames & Hudson.

Bhattacharya, D., Qiu, H., Lee, J., Yoon, H. S., Weber, A. P. M., and Price, D. C. (2018). When less is more: red algae as models for studying gene loss and genome evolution in eukaryotes. *Crit. Rev. Plant Sci.* 37, 81–99. doi: 10.1080/07352689.2018.1482364

Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes Dev.* 16, 6–21. doi: 10.1101/gad.947102

Brouhard, G. J., and Rice, L. M. (2018). Microtubule dynamics: an interplay of biochemistry and mechanics. *Nat. Rev. Mol. Cell. Biol.* 19, 451–463. doi: 10.1038/s41580-018-0009-y

Burgess, S. D., Bowring, S., and Shen, S.-Z. (2014). High-precision timeline for Earth's most severe extinction. *Proc. Natl. Acad. Sci. U.S.A.* 111, 3316–3321. doi: 10.1073/pnas.1317692111

Castenholz, R. W., and McDermott, T. R. (2010). "The Cyanidiales: ecology, biodiversity, and biogeography," in *Red Algae in the Genomic Age*, ed. D. J. Chapman (Berlin: Springer), 357–371. doi: 10.1007/978-90-481-3795-4_19

Cingolani, P., Platts, A., Wang, L. L, Coon, M., Nguyen, T., Wang, L., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* 6, 80–92. doi: 10.4161/fly.19695

Ciniglia, C., Yoon, H. S., Pollio, A., Pinto, G., and Bhattacharya, D. (2004). Hidden biodiversity of the extremophilic Cyanidiales red algae. *Mol. Ecol.* 13, 1827–1838. doi: 10.1111/j.1365-294X.2004.02180.x

D'Amico, S., Claverie, P., Collins, T., Georlette, D., Gratia, E., Hoyoux, A., et al. (2002). Molecular basis of cold adaptation. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 357, 917–925. doi: 10.1098/rstb.2002.1105

Deaton, A. M., and Bird, A. (2011). CpG islands and the regulation of transcription. *Genes Dev.* 25, 1010–1022. doi: 10.1101/gad.2037511

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43, 491–498. doi: 10.1038/ng.806

Detrich, H. W. III, Parker, S. K., Williams, R. C. Jr., Nogales, E., and Downing, K. H. (2000). Cold adaptation of microtubule assembly and dynamics. Structural interpretation of primary sequence changes present in the alpha- and beta-tubulins of Antarctic fishes. *J. Biol. Chem.* 275, 37038–37047. doi: 10.1074/jbc.M005699200

Doemel, W. N., and Brock, T. (1971). The physiological ecology of *Cyanidium caldarium*. *Microbiology* 67, 17–32. doi: 10.1099/00221287-67-1-17

Dupont, C., Armant, D. R., and Brenner, C. A. (2009). Epigenetics: definition, mechanisms and clinical perspective. *Semin. Reprod. Med.* 27, 351–357. doi: 10.1055/s-0029-1237423

Foflonker, F., Mollegard, D., Ong, M., Yoon, H. S., and Bhattacharya, D. (2018). Genomic analysis of picochlorum species reveals how microalgae may adapt to variable environments. *Mol. Biol. Evol.* 35, 2702–2711. doi: 10.1093/molbev/msy167

Gross, W. (2000). Ecophysiology of algae living in highly acidic environments. *Hydrobiologia* 433, 31–37. doi: 10.1023/A:1004054317446

Gross, W., Oesterhelt, C., Tischendorf, G., and Lederer, F. (2002). Characterization of a non-thermophilic strain of the red algal genus Galdieria isolated from Soos (Czech Republic). *Eur. J. Phycol.* 37, 477–483.

Hendry, A. P., and Kinnison, M. T. (1999). Perspective: the pace of modern life: measuring rates of contemporary microevolution. *Evolution* 53, 1637–1653. doi: 10.1111/j.1558-5646.1999.tb04550.x

Hew, C. S., Krotkov, G., and Canvin, D. T. (1969). Effects of temperature on photosynthesis and Co2 evolution in light and darkness by green leaves. *Plant Physiol.* 44, 671–677. doi: 10.1104/pp.44.5.671

Himes, R. H., and Detrich, H. W. (1989). Dynamics of Antarctic fish microtubules at low-temperatures. *Biochemistry* 28, 5089–5095. doi: 10.1021/bi00438a028

Hsieh, C. J., Zhan, S. H., Liao, C. P., Tang, S. L., Wang, L. C., Watanabe, T., et al. (2018). The effects of contemporary selection and dispersal limitation on the community assembly of acidophilic microalgae. *J. Phycol.* 54, 720–733. doi: 10.1111/jpy.12771

Huertas, I. E., Rouco, M., Lopez-Rodas, V., and Costas, E. (2011). Warming will affect phytoplankton differently: evidence through a mechanistic approach. *Proc. R. Soc. B Biol. Sci.* 278, 3534–3543. doi: 10.1098/rspb.2011.0160

Iovinella, M., Eren, A., Pinto, G., Pollio, A., Davis, S. J., Cennamo, P., et al. (2018). Cryptic dispersal of Cyanidiophytina (Rhodophyta) in non-acidic environments from Turkey. *Extremophiles* 22, 713–723. doi: 10.1007/s00792-018-1031-x

Jabbari, K., and Bernardi, G. (2004). Cytosine methylation and CpG, TpG (CpA) and TpA frequencies. *Gene* 333, 143–149. doi: 10.1016/j.gene.2004.02.043

Jenuwein, T., and Allis, C. D. (2001). Translating the histone code. *Science* 293, 1074–1080. doi: 10.1126/science.1063127

Kolbert, E. (2014). *The Sixth Extinction: An Unnatural History*, 1st Edn. New York, NY: Henry Holt and Company.

Lescot, M., Dehais, P., Thijs, G., Marchal, K., Moreau, Y., Van de Peer, Y., et al. (2002). PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. *Nucleic Acids Res.* 30, 325–327. doi: 10.1093/nar/30.1.325

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

López-Rodas, V., Costas, E., Maneiro, E., Marvá, F., Rouco, M., Delgado, A., et al. (2009). Living in Vulcan's forge: algal adaptation to stressful geothermal ponds on vulcano island (southern Italy) as a result of pre-selective mutations. *Phycol. Res.* 57, 111–117. doi: 10.1111/j.1440-1835.2009.00527.x

Lun, A. T. L., Chen, Y., and Smyth, G. K. (2016). "It's DE-licious: a recipe for differential expression analyses of rna-seq experiments using quasi-likelihood methods in edgeR," in *Statistical Genomics: Methods and Protocols*, eds E. Mathe and S. Davis (New York, NY: Springer), 391–416. doi: 10.1007/978-1-4939-3578-9_19

McElwain, J. C., and Punyasena, S. W. (2007). Mass extinction events and the plant fossil record. *Trends Ecol. Evol.* 22, 548–557. doi: 10.1016/j.tree.2007.09.003

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit: a map reduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110

McKenzie, G. J., Lee, P. L., Lombardo, M.-J., Hastings, P., and Rosenberg, S. M. (2001). SOS mutator DNA polymerase IV functions in adaptive mutation and not adaptive amplification. *Mol. Cell* 7, 571–579. doi: 10.1016/s1097-2765(01)00204-0

Napolitano, R., Janel-Bintz, R., Wagner, J., and Fuchs, R. (2000). All three SOS-inducible DNA polymerases (Pol II, Pol IV and Pol V) are involved in induced mutagenesis. *EMBO J.* 19, 6259–6265. doi: 10.1093/emboj/19.22.6259

Ness, R. W., Morgan, A. D., Colegrave, N., and Keightley, P. D. (2012). Estimate of the spontaneous mutation rate in *Chlamydomonas reinhardtii*. *Genetics* 192, 1447–1454. doi: 10.1534/genetics.112.145078

Nikolova, D., Weber, D., Scholz, M., Bald, T., Scharsack, J. P., and Hippler, M. (2017). Temperature-induced remodeling of the photosynthetic machinery tunes photosynthesis in the Thermophilic Algae *Cyanidioschyzon merolae*. *J. Plant Physiol.* 174, 35–46. doi: 10.1104/pp.17.00110

Osundeko, O., Dean, A. P., Davies, H., and Pittman, J. K. (2014). Acclimation of microalgae to wastewater environments involves increased oxidative stress tolerance activity. *Plant Cell Physiol.* 55, 1848–1857. doi: 10.1093/pcp/pcu113

Palumbi, S. R. (2001). Evolution - Humans as the world's greatest evolutionary force. *Science* 293, 1786–1790. doi: 10.1126/science.293.5536.1786

Paonessa, F., Criscuolo, S., Sacchetti, S., Amoroso, D., Scarongella, H., Pecoraro Bisogni, F., et al. (2016). Regulation of neural gene transcription by optogenetic inhibition of the RE1-silencing transcription factor. *Proc. Natl. Acad. Sci. U.S.A.* 113, E91–E100. doi: 10.1073/pnas.1507355112

Perrineau, M. M., Gross, J., Zelzion, E., Price, D. C., Levitan, O., Boyd, J., et al. (2014). Using natural selection to explore the adaptive potential of *Chlamydomonas reinhardtii*. *PLoS One* 9:e92533. doi: 10.1371/journal.pone.0092533

Qiu, H., Price, D. C., Weber, A. P., Reeb, V., Yang, E. C., Lee, J. M., et al. (2013). Adaptation through horizontal gene transfer in the cryptoendolithic red alga *Galdieria phlegrea*. *Curr. Biol.* 23, R865–R866. doi: 10.1016/j.cub.2013.08.046

Qiu, H., Price, D. C., Yang, E. C., Yoon, H. S., and Bhattacharya, D. (2015). Evidence of ancient genome reduction in red algae (Rhodophyta). *J. Phycol.* 51, 624–636. doi: 10.1111/jpy.12294

Qiu, H., Rossoni, A. W., Weber, A. P. M., Yoon, H. S., and Bhattacharya, D. (2018). Unexpected conservation of the RNA splicing apparatus in the highly streamlined genome of *Galdieria sulphuraria*. *BMC Evol. Biol.* 18:41. doi: 10.1186/s12862-018-1161-x

Reeb, V., and Bhattacharya, D. (2010). "The thermo-acidophilic cyanidiophyceae (Cyanidiales)," in *Red Algae in the Genomic Age*, ed. D. J. Chapman (Berlin: Springer), 409–426. doi: 10.1007/978-90-481-3795-4_22

Rhen, T., and Cidlowski, J. A. (2005). Antiinflammatory action of glucocorticoids - New mechanisms for old drugs. *N. Engl. J. Med.* 353, 1711–1723. doi: 10.1056/NEJMra050541

Roberts, M. F. (2005). Organic compatible solutes of halotolerant and halophilic microorganisms. *Saline Systems* 1:5.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Romero-Lopez, J., Lopez-Rodas, V., and Costas, E. (2012). Estimating the capability of microalgae to physiological acclimatization and genetic adaptation to petroleum and diesel oil contamination. *Aquat. Toxicol.* 12, 227–237. doi: 10.1016/j.aquatox.2012.08.001

Rossoni, A. W., Price, D. C., Seger, M., Lyska, D., Lammers, P., Bhattacharya, D., et al. (2019). The genomes of polyextremophilic Cyanidiales contain 1% horizontally transferred genes with diverse adaptive functions. *bioRxiv* [Preprint]. doi: 10.1101/526111

Rossoni, A. W., Schönknecht, G., Lee, H. J., Rupp, R. L., Flachbart, S., Mettler-Altmann, T., et al. (2018). Cold Acclimation of the Thermoacidophilic Red Alga *Galdieria sulphuraria* - Changes in gene expression and involvement of horizontally acquired genes. *Plant Cell Physiol.* 60, 702–712. doi: 10.1093/pcp/pcy240

Sahney, S., and Benton, M. J. (2008). Recovery from the most profound mass extinction of all time. *Proc. R. Soc. B Biol. Sci.* 275, 759–765. doi: 10.1098/rspb.2007.1370

Santos, H., and da Costa, M. S. (2002). Compatible solutes of organisms that live in hot saline environments. *Environ. Microbiol.* 4, 501–509. doi: 10.1046/j.1462-2920.2002.00335.x

Saxonov, S., Berg, P., and Brutlag, D. L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl. Acad. Sci. U.S.A.* 103, 1412–1417. doi: 10.1073/pnas.0510310103

Schonknecht, G., Chen, W. H., Ternes, C. M., Barbier, G. G., Shrestha, R. P., Stanke, M., et al. (2013). Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote. *Science* 339, 1207–1210. doi: 10.1126/science.1231707

Schönknecht, G., Weber, A. P., and Lercher, M. J. (2014). Horizontal gene acquisitions by eukaryotes as drivers of adaptive evolution. *Bioessays* 36, 9–20. doi: 10.1002/bies.201300095

Scott, D. K., Mitchell, J. A., and Granner, D. K. (1996). The orphan receptor COUP-TF binds to a third glucocorticoid accessory factor element within the phosphoenolpyruvate carboxykinase gene promoter. *J. Biol. Chem.* 271, 31909–31914. doi: 10.1074/jbc.271.50.31909

Seckbach, J. (1972). On the fine structure of the acidophilic hot-spring alga *Cyanidium caldarium*: a taxonomic approach. *Microbios* 5, 133–142.

Shen, S. Z., Crowley, J. L., Wang, Y., Bowring, S. A., Erwin, D. H., Sadler, P. M., et al. (2011). Calibrating the end-permian mass extinction. *Science* 334, 1367–1372. doi: 10.1126/science.1213454

Stockwell, C. A., Hendry, A. P., and Kinnison, M. T. (2003). Contemporary evolution meets conservation biology. *Trends Ecol. Evol.* 18, 94–101. doi: 10.1016/s0169-5347(02)00044-7

Sung, W., Ackerman, M. S., Miller, S. F., Doak, T. G., and Lynch, M. (2012). Drift-barrier hypothesis and mutation-rate evolution. *Proc. Natl. Acad. Sci. U.S.A.* 109, 18488–18492. doi: 10.1073/pnas.1216223109

van Bakel, H., Nislow, C., Blencowe, B. J., and Hughes, T. R. (2010). Most "dark matter" transcripts are associated with known genes. *PLoS Biol.* 8:e1000371. doi: 10.1371/journal.pbio.1000371

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., et al. (2013). From FastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* 43, 11.10.1–11.10.33. doi: 10.1002/0471250953.bi1110s43

Voter, W. A., and Erickson, H. P. (1984). The kinetics of microtubule assembly. Evidence for a two-stage nucleation mechanism. *J. Biol. Chem.* 259, 10430–10438.

Wittkopp, P. J., and Kalay, G. (2012). Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat. Rev. Genet.* 13, 59–69. doi: 10.1038/nrg3095

Wray, G. A. (2007). The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet.* 8, 206–216. doi: 10.1038/nrg2063

Yamori, W., Hikosaka, K., and Way, D. A. (2014). Temperature response of photosynthesis in C-3, C-4, and CAM plants: temperature acclimation and temperature adaptation. *Photosynth. Res.* 119, 101–117. doi: 10.1007/s11120-013-9874-6

Zhao, F. Q. (2013). Octamer-binding transcription factors: genomics and functions. *Front. Biosci.* 18:1051–1071.

Zhu, J., He, F., Hu, S., and Yu, J. (2008). On the nature of human housekeeping genes. *Trends Genet.* 24, 481–484. doi: 10.1016/j.tig.2008.08.004

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

IV. Danksagung

An dieser Stelle möchte ich die Gelegenheit nutzen, mich bei den Menschen und Institutionen zu bedanken, die diese Dissertation ermöglicht haben:

Prof. Dr. Andreas P.M. Weber für die Bereitstellung des Themas, der Betreuung und Begutachtung der Arbeit, sowie für das Vertrauen, die Unterstützung und die erfolgreiche Zusammenarbeit.

Prof. Dr. Michael Feldbrügge reibungslose Übernahme der Rolle als Zweitgutachter.

Prof. Dr. Debashish Bhattacharya für die tatkräftige Unterstützung, seinen Rat und insbesondere seine Gastfreundschaft und Herzlichkeit während meiner Zeit in New Brunswick. Vielen Dank auch an Susanne, Udi, Alex, Dana und Nicole.

Allen Schnick-Schnack-Schnuck-Experten, Kaffeetrinker, Fitness-Freaks, Basketballer, Pokerstars, Zeit-Quizzer und Food-Fanatiker im Institut für Biochemie der Pflanzen. Insbesondere den dauerhaften Mitgliedern des Metal-Labs für die zahlreichen unterhaltsamen Momente.

Meinen Eltern für ihren Rückhalt während des gesamten Studiums.

Dr. Matthias Nolte, ohne dessen Inspiration für das Fach Biologie ich wahrscheinlich „etwas Vernünftiges“ studiert hätte.

Heinrich Heine Consulting e.V. den Paradigmenwechsel.

Den Verlagen eLife, Science, Frontiers, Springer Nature, Oxford University Press und Elsevier für die freundliche Bereitstellung veröffentlichter Inhalte.

Julika. We made a language for us two, we don't need to describe.