Palacký University Olomouc

Faculty of Science

Department of Botany

Genome variability in the evolution of microorganisms

Cyanobacteria: Painting the speciation continuum

Doctoral Thesis

By

Mgr. Aleksandar Stanojković

Supervisor: doc. Mgr. Petr Dvořák, Ph.D.

Olomouc 2023

Declaration

I hereby declare that this doctoral thesis has been written solely by myself and without using resources other than those listed in the "References" section. All published results included in this thesis have been approved by the co-authors.

In Olomouc on 31st May 2023

Mgr. Aleksandar Stanojković

Abstract

Genetic variation determines the potential for species to evolve. The complex interplay of evolutionary forces, such as gene flow, mutations, recombination, and selection, shape the extent of variability within a species. The concerted action of these forces dictates the degree of genetic differentiation among different microbial taxa as speciation unfolds, ultimately resulting in their divergence. However, the lack of adequate species concept, promiscuous gene exchange, and high cryptic diversity of microbes hampered the speciation study in cyanobacteria. In this thesis, I studied patterns of speciation in non-model free-living soil cyanobacteria *Microcoleus* and *Laspinema* on a local and global scale by searching for genome-wide hallmarks of ongoing differentiation and divergence. Employing population-level sampling from all continents besides South America, 500 closely related cyanobacterial strains were obtained. They were characterized based on morphology, sequenced markers 16S rRNA and 16S-23S ITS, and genomes (210 strains), which were subsequently used for the population genomics analyses. We found at least 12 distinct species at different points along a continuum of divergence in the global collection of *Microcoleus*, with up to four coexisting in sympatry. A significant influence of abiotic environmental factors (e.g., soil, climate, UV light), homologous recombination, and geography contributed to the diversification of *Microcoleus* in terrestrial soil systems. Furthermore, an ongoing divergence was captured between *Microcoleus* and *Laspinema* species in a sympatric setting. The speciation of these cyanobacteria was likely governed by adaptation to novel yet unexplored microniches in soil systems, particularly adaptation to varied light conditions and stress stimuli. In aggregate, genome-wide signatures of genetic differentiation, homologous recombination, and selection allowed us to place the diverging species on the speciation continuum, although the boundary remains blurry. Further studies are needed to unravel the functions of the genes under selection pressures as well as the nature of emerging barriers to gene flow between the species. Overall, the results of this thesis provide a deeper understanding of the genetic diversity that underlies ongoing speciation in terrestrial cyanobacteria and elucidates mechanisms contributing to the rise of new cyanobacterial species.

Keywords

Abstrakt

Míra genetické variability určuje potenciál pro vývoj druhů. Složitá souhra evolučních sil, jako je tok genů, mutace, rekombinace a selekce, utváří rozsah variability v rámci druhu. Společné působení těchto sil určuje míru genetické diferenciace mezi různými mikrobiálními taxony v průběhu speciace, což nakonec vede k jejich divergenci. Nicméně neadekvátní druhové koncepty, častá výměna genů a vysoká kryptická diverzita mikrobů brzdily studium speciace u sinic. V této práci jsem studoval speciaci u nemodelových volně žijících půdních sinic *Microcoleus* a *Laspinema* v lokálním i globálním měřítku charakterizací probíhající diferenciace a divergence na urovní celého genomu. Na základě vzorkování na úrovni populací ze všech kontinentů, kromě Jižní Ameriky, bylo získáno 500 blízce příbuzných kmenů sinic. Byly charakterizovány na základě morfologie, sekvenovaných markerů 16S rRNA, 16S-23S ITS a genomů (210 kmenů), které byly následně použity pro analýzy populační genomiky. V globální sbírce *Microcoleus* jsme nalezli nejméně 12 odlišných druhů v různých fázích speciačního kontinua, přičemž až čtyři druhy koexistovaly v sympatrii. K diverzifikaci *Microcoleus* v terestrických půdních systémech přispěl významný vliv abiotických faktorů prostředí (např. půda, klima, UV záření), homologní rekombinace a geografická diferenciace. Dále byla zachycena probíhající divergence mezi druhy *Microcoleus* a *Laspinema* v sympatrickém prostředí. Speciace těchto sinic byla pravděpodobně poháněna adaptací na nové, dosud neprozkoumané mikroniky v půdních systémech, zejména adaptací na rozmanité světelné podmínky a stresové podněty. Genetická diferenciace, homologní rekombinace a selekce nám umožnily zařadit divergující druhy do speciačního kontinua, i když hranice zůstává nejasná. K odhalení funkcí genů pod selekčním tlakem a také povahy vznikajících bariér toku genů mezi druhy jsou zapotřebí další studie. Celkově výsledky této práce poskytují hlubší pochopení genetické rozmanitosti, která je základem probíhající speciace u suchozemských sinic, a objasňují mechanismy přispívající ke vzniku nových druhů sinic.

Klíčová slova

Biogeografie, kryptická diverzita, sinice, *Laspinema*, *Microcoleus*, fylogenetika, populační genomika, speciace

List of abbreviations

ANI – Average nucleotide identity

ASR – Ancestral state reconstruction

bp, Mb – Base pair, Mega base

BSC – Biological species concept

DNA – Deoxyribonucleic acid

D_{XY} – Average number of nucleotide differences

F_{ST} – Fixation index

HGT – Horizontal gene transfer

HR – Homologous recombination

ITS – Internal transcribed spacer of the ribosomal operon

LD – Linkage disequilibrium

ML – Maximum-likelihood

rRNA – Ribosomal ribonucleic acid

SNP – Single nucleotide polymorphism

SSU – Small subunit ribosomal gene

π – Nucleotide diversity

List of papers included in the thesis

(the thesis is based on three papers – see Appendices for full reprints/manuscripts)

I. **Stanojković, A.**, Skoupý, S., Hašler, P., Poulíčková, A., & Dvořák, P. **2022**. Geography and climate drive the distribution and diversification of the cosmopolitan cyanobacterium *Microcoleus* (Oscillatoriales, Cyanobacteria). *European Journal of Phycology*, 57, 396-405.

II. **Stanojković, A.**, Skoupý, S., Škaloud, P., & Dvořák, P. **2022**. High genomic differentiation and limited gene flow indicate recent cryptic speciation within the genus *Laspinema* (cyanobacteria). *Frontiers in Microbiology*, 13, 977454.

III. **Stanojković, A.**, Skoupý, S., & Dvořák, P. **2023**. The global speciation continuum of cyanobacterium *Microcoleus*. *Manuscript*.

Additional papers

The following papers were submitted during my doctoral studies but are not included in this thesis.

Skoupý, S., **Stanojković, A**., Pavlíková, M., Poulíčková, A., & Dvořák, P., **2022**. New cyanobacterial genus *Argonema* is hiding in soil crusts around the world. *Scientific Reports*, 12, 7203.

Dvořák, P., Jahodářová, E.*, **Stanojković, A.***, Skoupý, S.*, & Casamatta, D. A. **2023**. Population genomics meets the taxonomy of cyanobacteria. *Algal Research*, 72, 103128.

* These authors contributed equally.

Contents

Introduction

A long time ago, in the galaxy not far away, a group of microorganisms branched off from other bacteria and became harbingers of changes in Earth's history. Our planet took a turn towards habitability around 2.5 billion years ago when these tenacious microorganisms manifested the ability to perform oxygenic photosynthesis (Lyons et al., 2014). This event is known as the Great Oxidation Event, and the microbes responsible for it are cyanobacteria, otherwise known as Cyanophyte, Cyanobacteriota (Oren et al., 2022), and blue-green (micro)algae. Their evolutionary innovation caused oxygen to amass in the atmosphere and begin a cascade of diversification events. That is why we owe our existence to cyanobacteria; now is the time to pay a small tribute to them.

Cyanobacteria are one of the oldest photosynthetic organisms. Although the time of their origin is still being questioned, it is thought that they are almost 3.5 billion years old (Schopf & Packer, 1987). During the billions of years of dominance on the Earth, they spread and inhabited myriad environments such as marine, freshwater, and terrestrial, but also extreme ones (like hot and cold deserts). In these habitats, they represent critical constituents of any microbial assemblage. They have significant roles in the global carbon flux, nitrogen fixation, or as essential plant or animal symbionts (Gaysina et al., 2019). Cyanobacterial significance also lies in their ability to produce and accumulate the most diverse metabolites with antitumor, antibacterial, and antiviral properties (Singh et al., 2011). At the same time, many species produce dangerous, toxic metabolites (cyanotoxins), which deteriorate aquatic ecosystems (van Apeldoorn et al., 2007).

Cyanobacteria are cosmopolitan and small in size, but at the same time, big enough to be visible even from space, particularly in blooms on the water surface (e.g., the Baltic sea; Kahru et al., 1994). Cyanobacteria are also simple in morphology and relatively easy to cultivate; thus, it is not surprising that they are becoming more attractive as studying systems. The long period of evolution throughout which cyanobacteria prevailed and became the most abundant in many habitats made them interesting for taxonomic, ecological, and evolutionary research. Elucidating what factors affect species' adaptability, biodiversity, and distribution is central to understanding species' emergence and cyanobacterial world dominance.

This thesis represents a tribute to the understudied areas of cyanobacterial evolutionary biology. In this thesis, I studied speciation patterns in filamentous soil cyanobacteria (*Microcoleus* and *Laspinema*) on a local and global scale and explored forces governing their diversity and divergence.

A glossary of key terms

Allopatry –a term describing physical factors that limit species dispersal and impose barriers to gene flow between them. These physical barriers could be large (such as the ocean, river, or mountain range) or very small (species are centimeters or micrometers apart).

Barriers to gene flow – various factors limiting the gene flow between species. They could be physical, ecological, and genetic. It is analogous to the reproductive barrier or isolation commonly used in eukaryotes.

Diversification – a process of fine interplay between speciation and extinction, by which taxonomic diversity increases within a species through time.

Ecological divergence – a process in which different habitat types or utilization of different resources drive the populations to become ecologically differentiated.

Gene flow – any exchange of genetic material between genetic units (populations, lineages, species) that leaves a signature in the genome through homologous recombination (HR) and non-homologous recombination.

Genetic divergence – a process in which different evolutionary forces (such as mutation, selection, gene flow) drive the populations to become genetically differentiated by accumulating genetic variation.

Horizontal gene transfer (HGT) – a process of environmental DNA integration into a genome. The mechanisms of genetic exchange include transformation, transduction, or conjugation. It is unidirectional in prokaryotes (from donor to recipient).

Microniches – micrometer to centimeter-scaled ecological niches that have a range of variables (such as biotic, abiotic, or chemical) to which species can be adapted.

Pangenome – a total number of genes that individuals of a particular species possess.

Population – a set of coexisting, closely related individuals of the same species connected by gene flow.

Speciation – a dynamic, continuous process representing any stage of ecological, genetic, morphological, or physiological differentiation of a population; the evolution of barriers to gene flow.

Speciation continuum – a continuum of barriers to gene flow/reproductive isolation.

Sympatry – a term describing the occurrence of species in the same geographic area where barriers to gene flow are low or absent.

Estimates of cyanobacterial biodiversity

Cyanobacteria are a diverse group of organisms in terms of their morphology, ecology, and physiology. Studies on cyanobacterial biodiversity have been popular since their discovery in the 19th century when it was investigated using exclusively morphological characters. Later, other features were considered, like ecology or the genetic background of organisms (so-called polyphasic approach). However, cyanobacteria were central to studies on their role in harmful water bloom formation (see Huisman et al., 2018) and climate change (particularly global warming), while their overall diversity remained less investigated. Consequently, the precise number of cyanobacterial species is still controversial, with some global estimates ranging from 2698 (Nabout et al., 2013) or 2741 (Dvořák et al., 2018) to even up to even 8000 species (Guiry, 2012). According to the AlgalBase (www.algaebase.org; Guiry and Guiry, 2022), there are 5572 formally described cyanobacteria (last checked in May 2023). Uncertainty about the real cyanobacterial diversity advocates the need for more extensive sampling in nature. The biggest challenges in cyanobacterial systematics most certainly pose the lack of adequate species concepts and differential use of either high or low-resolution molecular methods.

Cyanobacteria: life between asexuality and sexuality

Cyanobacteria and other prokaryotes often receive undeserved epithet – primitive. This mainly stems from their simple cell structures, metabolism, and asexual reproduction. Yet, the more we dig into their nature, the more we realize that they are far from simple or primitive. The reproduction in cyanobacteria is accomplished by binary or multiple fission (Kunkel, 1984), thallus fragmentation (Meeks & Elhai, 2002), formation of spores or other specialized structures such as akinetes (Waterbury & Stainer, 1977; Kaplan-Levy et al., 2010). Clonal reproduction is a prominent microbial feature, and why researchers thought microbial individuals do not significantly differ from each other on the intraspecies level. However, many authors endorsed the ability of microbes, including cyanobacteria, to promiscuously exchange the DNA through horizontal (lateral) gene transfer (HGT) (**Figure 1**; Ochman et al., 2000; Thomas & Nielsen, 2005; McDaniel et al., 2010; Dvořák et al., 2014, 2015). HGT is mediated through homologous and non-homologous recombination and represents a process of "the non-genealogical transmission of

genetic material from one organism to another" (Goldenfeld & Woese, 2007). Mechanisms of genetic exchange – transformation (uptake of the DNA from the environment), transduction (uptake of the DNA via phages), conjugation (direct exchange of the DNA between cells), and gene transfer agents (GTAs) introduce foreign DNA to the microbial cell and allow for recombination to act (Popa & Dagan, 2011). While recombination is more potent among closely related species, it has the power even to overcome species boundaries. In other words, it was found to occur between all domains, regardless of evolutionary relatedness, and in all directions, i.e., from Bacteria to Archaea and Eukarya and vice versa (e.g., Rest & Mindell, 2003; Watkins & Gray, 2006).

Recent estimates of homologous recombination (HR) revealed an intriguing aspect of reproduction in prokaryotes (Bobay & Ochman, 2017; Vos & Didelot, 2009). HR was demonstrated to be much more common among bacteria than previously thought, as only less than 15% of investigated species exhibited no signs of genetic exchange (see also Bobay, 2020). Even without the machinery required for sexual reproduction, such frequent HR leaves detectable traces in genetic material, similar to eukaryotes. Hence, cyanobacteria live on the thin line between sexuality and asexuality, often regarded as quasi-sexuality (Rosen et al., 2015). Fraser et al. (2007) even showed that through a simulation of a population divergence with varying levels of HR, confirming that different microbes can be anywhere from clonal (HR occurs less than twice as often as mutation) to sexual (HR occurs more than twice as often as mutation). The concept of quasi-sexuality opened the doors to applying specific species concepts (see chapter Intertwined species concepts and definitions) and definitions in cyanobacteria. Furthermore, it broadened our understanding of evolutionary mechanisms generating fine-scale genetic diversity in microbes (see chapter Speciation).

Interestingly, two terms are often interchangeably used to describe the genetic exchange between individuals – HGT and gene flow. In the Papers of this thesis, we use the term gene flow to describe the exchange of DNA material realized by HR.
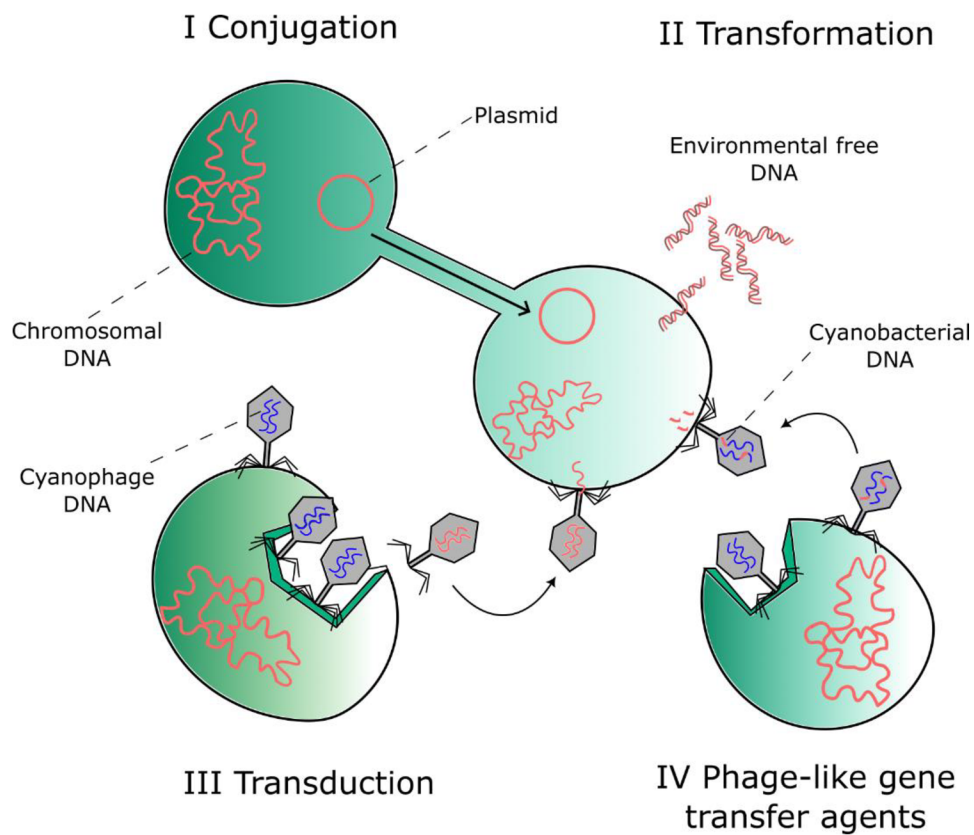
Figure 1. The scheme of horizontal gene transfer from one cyanobacterium to another. (**I**) Conjugation is the transfer of DNA from donor to recipient cyanobacterium and requires cell-to-cell contact via various adhesives or pili. (**II**) Transformation is defined by the uptake of extracellular DNA. (**III**) Transduction is the trasnfer of DNA via cyanophages. They could transfer the DNA from previously infected bacteria or currently infected one. (**IV**) Gene transfer agents (GTAs) are phage-like particles that can carry short fragments of bacterial DNA. The DNA of cyanophage is depicted in blue and cyanobacterial DNA is represented in red.

Intertwined species concepts and definitions in cyanobacteria

Ever since the classification of organisms became essential in biology, one of the most intrinsic questions emerged. What is a species? Fundamentally, there are common difficulties encountered when addressing this question. One being the species concept and the other being the species definition. The distinction is that species definition represents a set of rules used to identify species, which can be supported by a specific species concept (Gevers et al., 2005). The species concepts are theoretical frameworks for delimitation (Hanage, 2013; Dvořák et al., 2015). Mistakenly, both

---

Box 1. A pragmatic solution to the complicated issue: Sequence thresholds

Before the era of genome sequencing, one of the first criteria to discriminate between two isolates was established based on the possibility of DNA hybridization. If the similarity by DNA-DNA hybridization is lower than 70%, then the isolates are considered different species (Brenner et al., 2000).

With the rise of novel sequencing technologies, various molecular markers able to delimit species surfaced. For cyanobacteria, the molecular marker 16S rRNA (SSU), which encodes small ribosomal subunits, is widely used in species identification and characterization. Its molecular clock properties, the universal presence among cyanobacterial species, and the rich database for sequence comparisons make 16S rRNA a convenient genetic marker (e.g., Woese et al., 1987; Honda et al., 1999). Many authors demonstrated that the threshold of 70% identity based on DNA-DNA hybridization coincides with 97.5% identity of 16S rRNA for delineating isolates (Strackerbrandt & Goebel, 1994). Others recommended a higher threshold of 98-99% (Strackerbrandt & Ebers, 2006; Meier-Kolthoff et al., 2013). However, owing to the high conservatism, it fails to capture the fine resolution at the species level and delineate many, especially cryptic cyanobacteria (Fox et al., 1992; Martinez-Murcia et al., 1992).

Another marker commonly used in cyanobacterial taxonomy is an internally transcribed spacer (ITS) between small and big ribosomal subunits - 16S-23S ITS region (**Figure 2**). By contrast, the ITS region exhibits more variability than 16S rRNA, so its secondary structures were useful in describing many cyanobacterial species (Johansen et al., 2011; Osorio-Santos et al., 2014). Next, one of the proposed cutoffs for species distinction is that when evolutionary lineages share more than 97% sequence similarity, they likely belong to the same species (Pietrasiak et al., 2014). Although the variability makes the 16S-23S ITS region an efficient marker in some cyanobacteria, it often lacks sufficient resolution power to distinguish cryptic ones (Pietrasiak et al., 2019).

With the genomic era, another threshold emerged as the gold standard for species delimitation – average nucleotide identity or ANI, which uses entire genomes for sequence comparisons. For instance, Goris et al. (2007) proposed that evolutionary lineages sharing more than 95-96% ANI are to be one species. However, neither of the sequence thresholds actually reflects the use of a theoretical framework for a species concept (Bobay, 2020).

terms are often synonymously used by many authors. In practice, researchers do not specify the species concept applied to the studied organism. For instance, when arbitrary sequence (dis)similarities are used to define species (**Box 1**), evolutionary causes of species divergence are disregarded; thus, that does not reflect the complexity of evolution (Bobay & Ochman, 2017). The importance of distinguishing between species concept and definition is tightly linked to the adequate characterization and identification of new species and, in that way, to a proper estimation of the species' biodiversity. Over 35 species concepts have already been proposed (Zachos, 2016). Different species concepts and definitions generally used in (cyano)bacteria are outlined below.

The classic, traditional approach to species delimitation in cyanobacteria revolves around their morphology – the phenetic concept (e.g., Sneath & Sokal, 1973). The most important features include cell dimensions, type of cell division, presence/absence of specific structures (e.g., sheath or heterocysts), and color. Phenotypic characters often depend on culturing conditions, and assessments of morphological (dis)similarities between species are greatly influenced by the subjectivity of the observer. There were plenty of instances in cyanobacteria when some taxa were erroneously placed into one species owing to the absence of conspicuous morphological or ecological differences, but they were actually genetically different (e.g., *Prochlorococcus*, Baumdicker et al., 2012; *Synechococcus*, Dvořák et al., 2014; *Microcystis*, Harke et al., 2016). This problem inspired researchers to coin a new term for such organisms - cryptic species, defined as "population systems which were believed to belong to the same species until genetic evidence showed the existence of isolating mechanisms separating them" (Stebbins, 1950; Grant, 1981). Evidently, these organisms represent a challenge for taxonomists and evolutionary biologists as they require more sensitive features than morphological ones to establish clear species boundaries.

Significant progress in species delineation was made with the phylogenetic species concept. In fact, a whole family of such concepts shares some common attributes, like consideration of the organism's evolutionary history (Hennig, 1966; Ridley, 1989; Baum & Shaw, 1995; Wheeler & Meier, 2000). The concepts accommodate the principle of monophyletic species groups (groups sharing a common ancestry) and having differences in some character states (e.g., ecological niche, morphology, evolutionary fate; de Quiroz, 2007). The main advantage of phylogenetic methods is the ability to resolve the evolutionary histories of species on both higher and lower levels (e.g., Hugenholtz et al., 2016). Still, different phylogenetic models frequently provide incongruencies in

phylogenetic trees. Phylogenetic concepts have neither clear, universally accepted criteria for defining a species nor do they consider the speciation model. In searching for a more fitting picture of asexual organisms, the monophyletic species concept emerged as one of today's most prominent and widely used ones in cyanobacteria (adapted by Johansen & Casamatta, 2005[1]). It is highly attractive as it incorporates morphological, ecological, and physiological distinctness of evolutionary lineages with genetic information – for instance, (dis)similarities of 16S rRNA and 16S-23S ITS region (**Figure 2**). When the studied lineages fall into distinct monophyletic groups/clades with the unique feature (autapomorphy), they are "worthy" to be recognized as species. The hunt for autapomorphies can be laborious, and it generally anchors in the researcher's subjectivity since one can choose which feature will be investigated. While this concept has vast potential for asexual organisms, it weakens when cryptic species come forth due to the often absence of a unique character (e.g., morphological, ecological). Finally, we may ask which species property is the most suitable for a lineage to be recognized as species. And, the answer lies in the interests of a researcher.
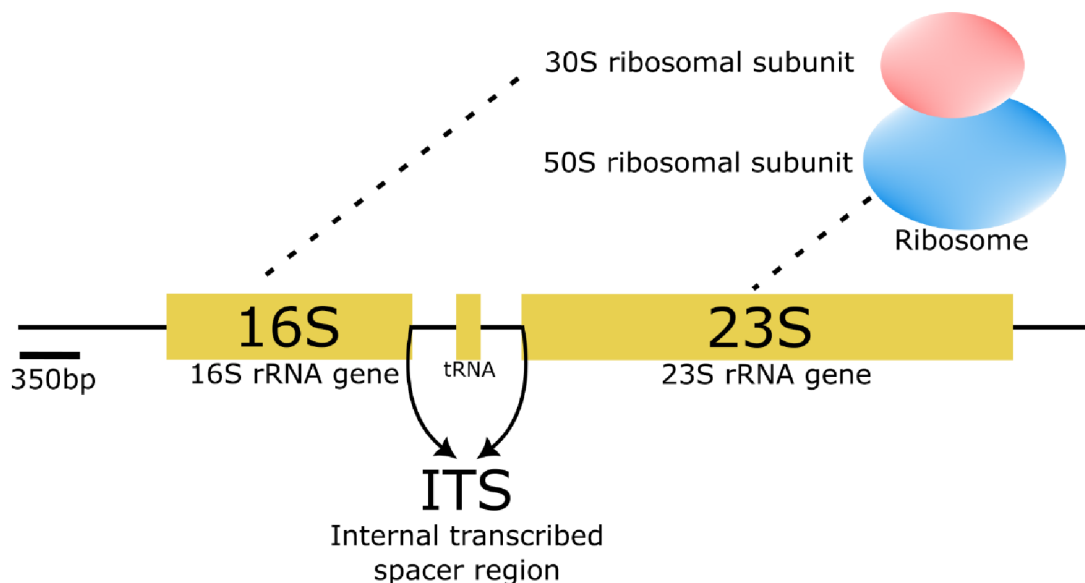
Figure 2. The scheme of the ribosome rRNA composition in prokaryotes. Internally transcribed spacer (ITS) region separates the 16S and 23S rRNA genes.

---

[1] Monophyletic species concept by Johansen & Casamatta (2005) is essentially the phylogenetic species concept *sensu* Mishler & Theriot (2000) modified to be applicable to (asexual) cyanobacteria. A particularly important difference between the two is that the species concept *sensu* Mishler & Theriot recognizes a species if there is at least *some morphological* distinction, but the monophyletic species concept can recognize a species with *any* distinction found between the two groups/clades.

Since the early days, when we all entered the world of the science of life, we have known about the biological species concept (BSC). The BSC had a central role in offering insight into the whole problematics of species. Ernst Mayr (1940, 1942, 1963) tailored and adjusted the revolutionary concept, which kept its essence:

"Species are groups of actually or potentially interbreeding natural populations which are reproductively isolated from other such groups."

Such a definition fits sexual organisms, which engage in recombination during meiosis. Therefore, the reproductive barriers maintain the genetic cohesion of species (Mayr, 1942). But how is this relevant for prokaryotes - (cyano)bacteria?

Given the absence of sex in microbes, defining a "good" species is challenging (Shapiro, 2018). The beauty of well-defined or good species is in the eyes of the observer. Depending on the type of research and dataset, we try to get to the definition of species that bears biological relevance; thus, essentially, species are subjectively determined. In some cases, they can be monophyletic clusters, and in others, clusters of closely related individuals occupying the specific ecological niche.

The BSC assumptions of the existence of a reproductive barrier or sexual isolation between two species mean that it can not be applied to microbes. It appears scientists were merely preoccupied with sex when they thought of defining species through reproductive barriers. The counterpart for reproductive isolation in sexual organisms would, for microbes, be a barrier to gene flow or HR (Bobay & Ochman, 2017; Shapiro, 2018). Mechanisms responsible for the cessation of HR between species can include geography (in allopatric settings), environmental factors (in sympatric settings), a higher level of sequence divergence, and different genetic systems like restriction-modification (e.g., CRISPR-Cas) or mismatch repair systems (Vulić et al., 1997; Jeltsch, 2003; Carrolo et al., 2009; Cordero et al., 2012; Shapiro & Polz, 2014). However, although the rate of HR with foreign DNA material decreases with the sequence divergence (Vulić et al., 1997), microbes can still exchange the DNA across species boundaries (Smith et al., 1993; Doolittle & Zhaxybayeva, 2009; Dvořák et al., 2014). That would always render the species borders "leaky" or "fuzzy" as the genetic exchange would occasionally occur (Hanage, 2013; Bobay, 2020). A recent study by Bobay & Ochman (2017) provided an example of delimiting microbial species on the basis of higher intraspecific than interspecific HR, which appears to be enough to maintain

cohesive genetic units equivalent to species. Although these novel species delimitation methods provided a good understanding of how one can detect species in microbes, the obstacle arises when two species are young and recently diverged (e.g., *Prochlorococcus*; Bobay & Ochman, 2017; *Microcoleus*; Paper I and Paper III; *Laspinema*, Paper II).

Noteworthy is the recent theoretical framework of Kollár et al. (2022) for understanding the species boundaries based on gene flow. Speciation is a dynamic and continuous process during which species can be in various stages of genetic and ecological differentiation (Shapiro & Polz, 2014). However, as the biological and any other species concept requires, species cannot be assigned to discrete units until they have fully diverged (Kollár et al., 2022). Thus, a novel approach has to be applied to diverging species, i.e., evolutionary lineages at the early stages of the speciation spectrum. The universal probabilistic concept of evolutionary lineages (UPCEL) offers the potential to delimit incipient species. Interestingly, the UPCEL puts probability on the lineages' divergence based on gene flow, which is prone to fluctuations. This means that putative species do not have to be discretely defined, but their divergence is described using the probability at a given time. Although an alluring and seemingly promising concept, applicable to all organisms engaging in gene flow, it is still young and requires more attention in the future.

A much more significant influx of ideas and theoretical frameworks for delineating microbes will undoubtedly appear, as it is still unclear how we can define them in a biologically meaningful way. We need creativity coupled with new theoretical frameworks and large datasets to attempt and disentangle this problem.

# Speciation: what drives the genetic variability?

Billion years old, cyanobacteria were engulfed in drastic environmental changes, from the times of the giant snowball Earth to the times of its gradual thawing. Cyanobacteria ensured their survival in the changing environment thanks to their remarkable plasticity, i.e., the ability to adapt to rapid environmental pressures, large population sizes, and short generation times. The secret of cyanobacterial endurance is woven into their genetic material, which underwent many changes over time and space, resulting in their world dominance nowadays. The fundamental part of evolution is the changes in genetic variability and allele frequencies, which have the potential to drive the rise of new species. Speciation, or species emergence, is governed by the concerted interplay of myriad evolutionary forces. Elucidating how species evolve to local and global environmental conditions is paramount to understanding the mechanisms governing life diversity. In this thesis, I explored their contribution to the speciation in recently diverged filamentous cyanobacteria.

Neutral processes

The effect of neutral or random evolutionary processes (neutral mutations and genetic drift) on the emergence of new microbial species has been in focus for a long time (e.g., Kimura, 1983). Microbes were thought to evolve mainly in sympatry by random genetic drift when rare alleles (advantageous or deleterious) have bigger chances of being removed from a gene pool, and selectively neutral alleles are the most frequent (Ohta, 1992). Hence, microbial species continuously diverge and go extinct within the same ecological niche. Conversely, a more recent study by Fraser et al. (2007) demonstrates that the neutral evolution model solely cannot explain the diversity of evolutionary lineages in the face of gene flow. The authors showed that higher interspecific HR is sufficient to maintain clusters of individuals as genetically distinct units. In that way, within such clusters of individuals, HR can generate distinct species-like patterns that are more similar to sexual than clonal populations. Basically, that means that neutral evolution, coupled with other evolutionary processes (e.g., diversifying selection, gene flow), can govern the occurrence of new microbial species (see Shapiro, 2018; Bobay, 2020).

A thought-provoking concept to estimate the frequency of HR within and between populations has recently been inspiring many studies to apply it to prokaryotes. Ultimately, it might bring us one step closer to unraveling the nature of microbial species.

I addressed this recombination-frequency model on the evolution of cyanobacteria on both local (separated by a few centimeters; Papers II and III) and global scale (separated by thousands of kilometers; paper III) (**Figure 3**).
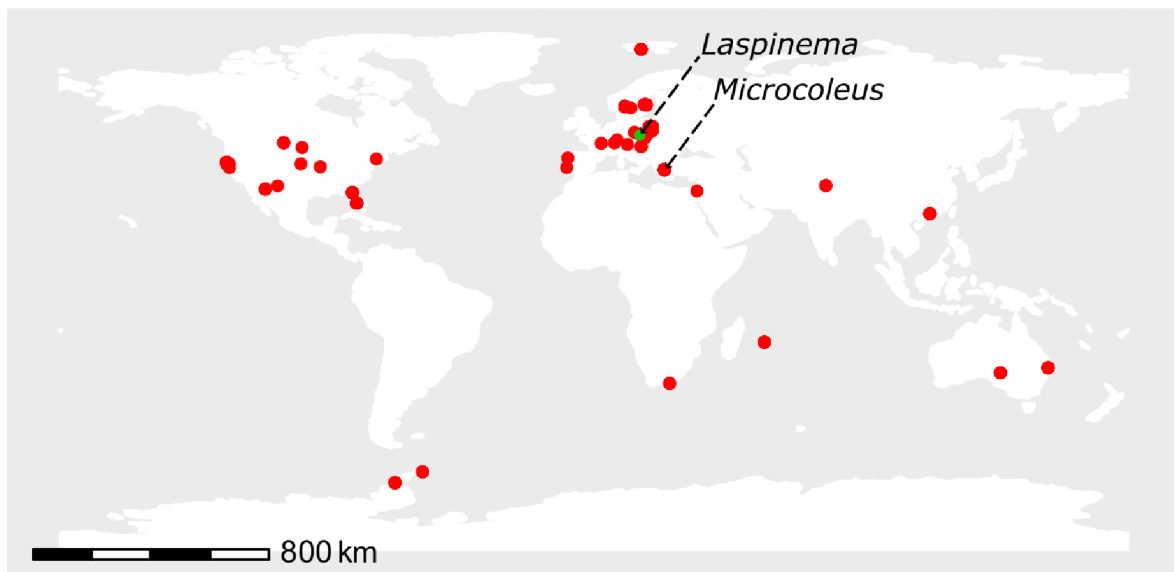
Figure 3. The geographic sampling of *Microcoleus* (our dataset and the GenBank sequences) and *Laspinema*. Every point represents one locality per which 1-11 strains have been isolated. Altogether, our dataset encompassed 10 *Laspinema*, and 291 *Microcoleus* isolates. Sampling sites of *Microcoleus* and *Laspinema* are marked as red and green dots, respectively.

Geographic isolation

The global dominance of microbes is supposedly ensured by their prominent dispersal ability and population sizes (Foissner, 2006); so, microbes were thought to have a predominant sympatric mode of speciation. The idea that geographic differentiation is not evolutionary significant for microorganisms was conceived by Baas-Becking in 1934 when the famous paradigm emerged "everything is everywhere, but the environment selects". Almost a century later, the idea that microbial distribution is ubiquitous and governed by local environmental factors still prevails,

particularly in cyanobacteria. Unfortunately, studies attempting to characterize cyanobacterial biogeographic patterns and estimate the effect geography has on their speciation are still scarce.

Baas-Becking's assertion was supported in studies by Finlay (2002) and Fenchel (2003) but based solely on morphological evidence. As previously mentioned, (cyano)bacteria often exhibit low morphological variability, which can cloud the actual genetic diversity (Leliaert et al., 2014). Employing molecular data, it was confirmed that some cyanobacteria have a cosmopolitan distribution - *Microcoleus/Coleofasciculus chtonoplastes* (16S rRNA, Garcia-Pichel et al., 1996) and *Microcystis* (ITS, van Gremberghe et al., 2011; 16S rRNA, Ribeiro et al., 2020). Conversely, geographic differentiation was a contributing factor in the speciation of *Synechococcus* (16S rRNA, Zwirglmaier et al., 2008), *Prochlorococcus* (16S rRNA, Zwirglmaier et al., 2008; ITS and whole-genomes, Kashtan et al., 2014, 2017), and *Microcoleus* (Dvořák et al., 2012; Paper I, Paper III). Thus, the time for Baas-Becking's paradigm shift arrived. These studies imply the existence of dispersal barriers (often temporary; Dvořák et al., 2012, Papers I and III), which prevent gene flow between isolated cyanobacteria and eventually lead to genetically distinct and independent evolutionary lineages (species).

Owing to the scarcity of biogeographic studies, we have conflicting results about the impact of geography (isolation by distance) on speciation in cyanobacteria. In papers I and III, I investigated the contribution of geographic isolation to the diversification of globally dominating cyanobacterium *Microcoleus*.

Ecological isolation

The pursuit of limited resources in the environment has always been a staple for evolutionary innovations, and in the world of microbes, it sparks a fierce contest. Microbes that venture into different ecological microniches in search of alternative and more abundant resources undergo adaptation to endure the inhospitable conditions in the environment. Although we still lack knowledge of all possible microniches in soil systems, they harbor a myriad of new niches with varying textures, aeration, hydration, and nutrient levels on a micrometer scale (Vos et al., 2013). Diverse ecological niches can affect the genetic diversity of bacteria and lead to distinct genetic groups (species) emerging in sympatry (Wiedenbeck & Cohan, 2011). So what factors could be driving microbial diversification?

Even though physiological experiments offer valuable insights into microbial responses to various environmental changes, they are commonly conducted in model organisms; sometimes, it is labor-intensive to set up the experiments, and certain factors cannot be easily tested under controlled laboratory conditions. However, with the genetic information at hand, we can investigate the factors affecting the genetic variability of specific microbial taxa without great strain. It has been found that present and past climatic conditions (temperature and precipitation) affect the community composition and genetic diversity of cyanobacteria *Microcoleus* (Muñoz-Martín et al., 2019; Papers I and III), *Prochlorococcus* (see Biller et al., 2015), *Raphidiopsis*, and *Microcystis* (Ribeiro et al., 2020). The availability of the resources such as iron, nitrogen, phosphorous, and polysaccharides significantly contributes to the genetic variation that enhances the ability of microbes to acquire and utilize them (e.g., Shapiro et al., 2012; Loza et al., 2014; Chase et al., 2019; Papers II and III). Additionally, ultraviolet (UV) radiation has a detrimental effect on the microbial DNA, so microbes that are able to produce photoprotective pigments or exhibit vertical migration from the surface to deeper soil layers may have a competitive advantage over their sessile counterparts (Garcia-Pichel et al., 2016). The impacts of climate change and human activities extend across every ecosystem, from terrestrial to marine environments, affecting microbial genetic diversity by introducing shifts in selection regime and geographical location (Pauls et al., 2012). These studies suggest that environmental factors can frequently trigger barriers to gene flow, and environmental isolation may be an important driver of microbial diversity.

The environmental parameters can be accessed without the need to individually measure all these variables by leveraging various public databases (e.g., WorldClim, SoilGrids). These databases provide a convenient source from which the required environmental parameters can be extracted for the geographical location of the samples. In Papers I and III, I explored the relative contribution of different environmental factors that may shape the diversification and distribution patterns of *Microcoleus*.

Selection: Interplay of selection and recombination

Another essential evolutionary process driving speciation and the distribution of microbes is environmental selection, encompassing both abiotic and biotic conditions. They can be responsible for selecting genetic variants that contribute to the differential survival of microbes in the

environment, thereby enforcing the rise of distinct ecological and genetic clusters. Two decades ago, Cohan developed a concept of the Stable Ecotype[2] Model of speciation, arguing that periodic environmental selection is the most dominant force in maintaining these distinct genetic clusters of microbes and driving niche specialization (Cohan, 2001; Wiedenbeck & Cohan, 2011). The authors emphasized that the HR rate between microbial populations is so low that adaptive (beneficial) alleles would most likely spread and fix by clonal expansion. Consequently, an adaptive allele at one locus becomes linked to alleles at other loci over the genome. In other words, distinct microbial clusters are genetically differentiated at the majority of loci; thus, genome diversity is reduced, i.e., genome-wide selective sweep (Shapiro & Polz, 2014). Such genome-wide selective sweeps occur every time when microbial population adapts to changing environmental conditions – periodic selection (Cohan, 2001).

However, many have challenged the influence of periodic selection as the dominant force in microbial evolution (e.g., Shapiro et al., 2012; Cadillo-Quiroz et al., 2012). Indeed, it was shown that the spread of beneficial alleles by HR was possible without the loss of diversity genome-wide, i.e., gene-specific selective sweep. This finding implies that distinct microbial clusters are strongly genetically differentiated at one or a few loci, which carry potentially adaptive allele(s). Moreover, the establishment of a barrier to genetic exchange promotes intraspecific HR (within clusters) and impedes interspecific HR (between clusters), hence outpacing the selection and eventually leading to species divergence (Shapiro & Polz, 2014).

Cyanobacteria lag behind other organisms since they have not been the focus organisms of evolutionary studies. Several recent studies that used cyanobacteria as model systems aligned with the hypothesis of recombination and gene flow significantly impacting their recent evolution (e.g., *Microcoleus/Coleofasciculus*, Vos & Didelot, 2009; *Prochlorococcus*, Bobay & Ochman 2017; *Microcystis*; Pérez-Carrascal et al., 2019). Curiously, the effect of recombination differs not only among species of different genera but also among species of the same genus. What implications does this generally have on the speciation in cyanobacteria? What is the underlying genetic structure of cyanobacteria diverging in the face of gene flow? How much can the effect of selection

---

[2] Ecotype represents a group of individuals sharing a high ecological similarity or a group of ecologically distinct genotypes (Cohan, 2006; Shapiro & Polz, 2014).

on genes inform us of cyanobacterial adaptability? I explored these missing pieces of the evolutionary puzzle in Papers II and III.

Horizontal gene transfer (HGT)

HGT is an important evolutionary force shaping the genetic structure underlying the impressive ability of microbes to adapt to a broad spectrum of environmental conditions (Lan & Reeves, 1996; Gogarten & Townsend, 2005; Boto, 2010). It represents a source of new, adaptive alleles or genes from distinct species, which mediate the colonization of novel ecological niches. Furthermore, acquiring new genes via HGT or HR introduces complications in reconstructing and interpreting evolutionary histories among microbes (Hanage et al., 2005; Boto, 2010). Consequently, the genome size and gene content considerably vary between individuals of the same species. The development of the pangenome concept offered insight into the complete set of genes present in microbial genomes (Tettelin et al., 2005) and allowed for comparing individual strains. Within the pangenome, we can differentiate between core genes - ubiquitous in a species and responsible for the basic metabolic functioning of the organism, and flexible/accessory genes - not present ubiquitously in a species and associated with the adaptations to environmental conditions (Medini et al., 2005). Accessory genes are a more common target of HGT. They undergo evolutionary changes more frequently than core genes, which are more conservative and suitable for reconstructing evolutionary relationships between species. Bearing this in mind, many have questioned whether the genetic changes gained via HGT are enough to drive the divergence of bacterial species allowing us to characterize them rationally (Riley & Lizotte-Waniewski, 2009).

The most prominent instances of HGT are acquisitions of genes associated with antibiotic resistance or toxin production in pathogens (Dobrindt & Reidl, 2000; Popa & Dagan, 2011). Although the research on the extent of HGT is still in its infancy in cyanobacteria, some of the best-studied species are *Prochlorococcus* and *Synechococcus* (e.g., Zhaxybayeva et al., 2006; Dvořák et al., 2014). In these species, it was found that HGT is commonly mediated in sympatry and via cyanophages, which introduced genes associated with nutrient uptake (e.g., nitrogen, phosphorous, iron), acclimation to light intensities, regulatory functions, and response to various environmental stress stimuli (Rocap et al., 2003; Stuart et al., 2013; Kashtan et al., 2014). As a result, adaptive alleles governed the divergence of closely related evolutionary lineages, leading to small differences in their fitness. Eventually, that is what allows cyanobacteria to coexist locally

and occupy different, often small ecological niches (microniches). These adaptive alleles or niche-specific variants are in charge of species differentiation that can lead to a full separation of species' evolutionary trajectories (Shapiro & Polz, 2015).

Notably, even if such variants arise in a microbial population, they do not necessarily lead to the end products of speciation that are genetically and ecologically differentiated species. There could also be other evolutionary scenarios (detailed in How can we study species' emergence). For instance, depending on the interplay of evolutionary forces, lineages may stay in the intermediate zones of speciation, often referred to as the "grey zone" (Shapiro & Polz, 2015; Roux et al., 2016; Kollár et al., 2022). Considering the continual and lengthy process of speciation, lineages may remain cryptic for millions of years in this grey zone (Coyne & Orr, 2004); this represents only a fraction of the species' existence length. Alternatively, if the species had not already diverged sufficiently and the barrier to gene flow had not ceased, they could end up merging (Fraser et al., 2009).

Investigating the adaptive benefits of transferred genes and the evolution of pangenome is essential to understand the overall impact of HGT on microbial speciation. In Papers II and III, I investigated the extent of genome diversity among cryptic cyanobacterial species and sought to place them in the corresponding speciation stages, i.e., to paint the speciation continuum.

How can we study the species' emergence?

From the idea of dominant clonality, cosmopolitanism, and sympatric mode of evolution in microbes, our perception of the contributors to vast microbial genetic diversity was drastically altered. Indeed, this was largely possible due to the developments in sequencing technologies and bioinformatic tools, which were sensitive enough to capture speciation in progress. Perhaps, the most promising way to study speciation is through the prism of population genomics, phylogenomics, and gene flow, as outlined by several authors (Shapiro & Polz, 2015; Bobay & Ochman, 2017; Arevalo et al., 2018). The main idea behind such a multifaceted approach is to study microbial species as close as possible to reality, where they can be anywhere on the spectrum of an early to a full divergence.

Challenging opportunities for sampling and sequencing

Although it is the first and the most critical step, the sampling strategy for microbes is a Gordian knot of the population genomic approach. The expected sample should be unbiased, random, large, and diverse enough to account for all representative lineages found in one sampling site (Luikart & Cornuet, 1998). Moreover, closely related individuals should be sampled both locally and globally. Locally sampled individuals would enable us to study the effects of selection and recombination, whereas globally sampled individuals allow us to consider the impact of allopatric divergence on species emergence.

Whether we sample a small patch of soil or a drop of ocean water, we are sure to find a heap of cyanobacterial cells. How can we identify individuals of a specific cyanobacterium that is our model system? How many individual strains do we need to acquire from a population? How different are cyanobacterial strains of the same species in other locations worldwide?

The first step in addressing these questions would be to examine the morphological characteristics of all individuals of a sample under the microscope. However, it is incredibly challenging to distinguish cyanobacteria when they look identical in morphology but, indeed, represent different species – high crypsis. If we were to extract the DNA of individual cultured strains and sequence their genomes right away, we might encounter a case when we have distantly related species in the dataset. That is why good practice is to sequence some molecular markers beforehand, such as 16S rRNA and 16S-23S ITS, which can serve as reasonable proxies for whole-genome analyses and the population structure (e.g., Kashtan et al., 2014, Paper I). With a priori information on genetic population structure, we can carefully select a subset of closely related individual strains for whole genome sequencing; so, the sampling strategy should, in fact, not be completely unbiased (Shapiro & Polz, 2015). As for how many strains from one sample should be included in a population genomic study, it is still unclear because microbes have large population sizes, disperse easily, and occupy an abundance of different ecological niches (Martiny et al., 2006). Until a straightforward recommendation for microbial populations' sample size becomes apparent, we can only sample and sequence as much as our resources allow (e.g., costs, computational limitations, lab conditions).

The breakthrough in sequencing technologies (e.g., Illumina, Pacific BioSciences, Oxford Nanopore) made sequencing on a genome level affordable and widely available only a decade ago. These advances are reflected in the possibility of obtaining longer reads (genome fractions) and higher throughput, which allowed the analysis of much bigger datasets (with thousands of genomes) than was previously possible with Sanger dideoxy sequencing (Sanger et al., 1977). Nevertheless, cyanobacterial genomics is moving at a snail's pace compared to other microbes, especially on a population level. Cyanobacterial genomes account for ~0.8% of all microbial genomes available at this moment (term searched "Cyanobacteria", last checked on April 2023), and surprisingly half of those belong to only two genera (*Prochlorococcus* and *Synechococcus*); hence, we have an incomplete picture of cyanobacterial phylum (**Figure 4**).
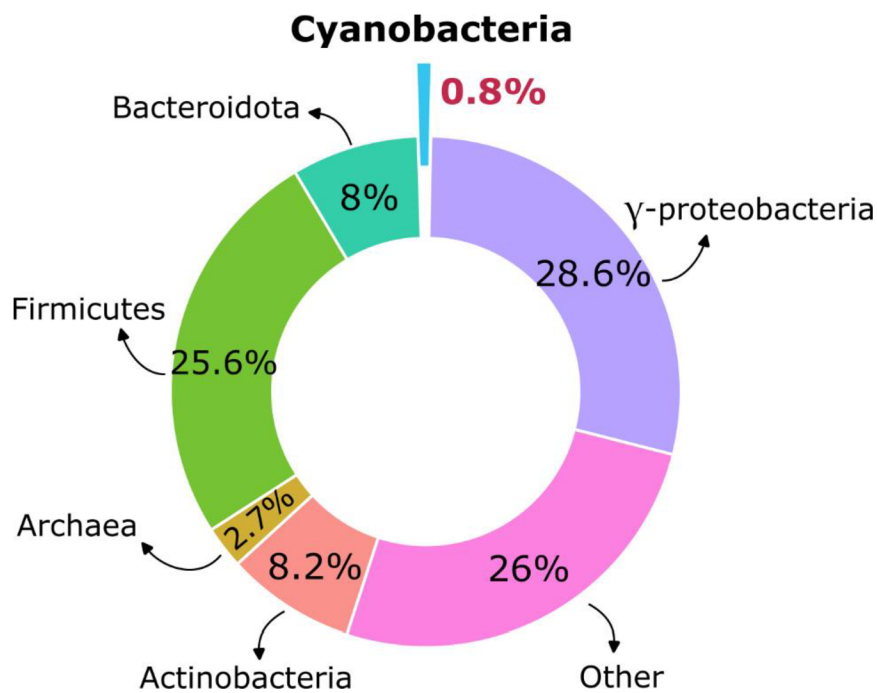
Figure 4. The proportion of prokaryotic genomes currently available in the GenBank database. The number of cyanobacterial genomes accounts for approximately 0.8% of all prokaryotic genomes.

The population genomic approach accommodates both whole-genome and metagenomic sequencing strategies (Arevalo et al., 2018). While both can be advantageous, whole-genome sequencing might be more beneficial for studying the direct effects of selection and recombination among lineages as it yields better accuracy and quality of assemblies (Shapiro & Polz, 2015).

Challenges we encounter in cyanobacterial research, from sampling, and obtaining pure cultures, to computational difficulties, all contributed to the lagging of cyanobacterial population genomics. Nevertheless, the global sampling and sequencing strategy presented in this thesis represents a step forward in cyanobacterial genomics research.

Inferring the evolutionary relationships and population structure

A phylogenetic tree has the power to inform us about the evolutionary descent of different species. It is essential to gain a deeper understanding of the events that occurred during the evolution of an organism. In the era of computational dominance, various bioinformatic pipelines have been developed for the inference of the evolutionary history of a species as reliably as possible. In Papers I and II, I used 16S rRNA and 16S-23S ITS sequences to get insight into individual species' overall relatedness and population structure. By doing so, I set the stage for a careful selection of closely related strains for whole-genome sequencing. Whole-genome phylogeny can then be best inferred with a combination of inputs like single-copy orthologs, individual gene trees, single nucleotide polymorphisms, and core genes (Papers II and III). This allows us to compare the disparities in a species' phylogenetic history based on different approaches and infer which one most probably corresponds to the species' evolutionary history.

How do we decide on what individuals constitute a population in a dataset? Although we can examine the phylogenetic tree for as long as we desire, accurately assigning individual strains to populations that perfectly match reality is not a simple task. Deciding on what microbial individuals constitute a population is challenging due to the inherent variability in the genetic makeup of microbial communities. If we were to infer a population structure based on, for instance, the monophyletic species concept, sole monophyletic clades would not be significant without some autapomorphy. The high cryptic diversity of many cyanobacterial species further exacerbates this issue. The usual practice is to use genetic similarity to define populations, like ANI or DNA-DNA hybridization (**Box 1**), or simply to consider a population as a group of individuals found from the same environmental sample or that occupy the same niche. Another possible solution to this problem is to employ a combination of clustering analyses, either based on various algorithms (Tonkin-Hill et al., 2019) or estimates of HR (Arevalo et al., 2019). Papers II and III delved into this matter.

Given that the pangenome represents a comprehensive set of all genes found in individuals within a dataset, it harbors a tremendous amount of helpful information on genetic variation (Tettelin et al., 2005). The usage of pangenome in species delimitation is based on the expectation that strains of one species share more similar gene repertoires than strains of distantly related ones (Bobay, 2020). A promising approach was recently developed by Moldovan & Gelfand (2018), whose proposed species delineation method focused on identifying genes that are lineage-specific. However, the main caveat in this approach is the flexible genome and the extensive evolutionary changes it undergoes, primarily via HGT. Moreover, many cyanobacterial genes still have unknown functions (e.g., ~76% of the *Laspinema* pangenome, Paper II). Therefore, we would need deeper genome sampling and further physiological experiments to understand the overall evolutionary dynamics of the pangenome.
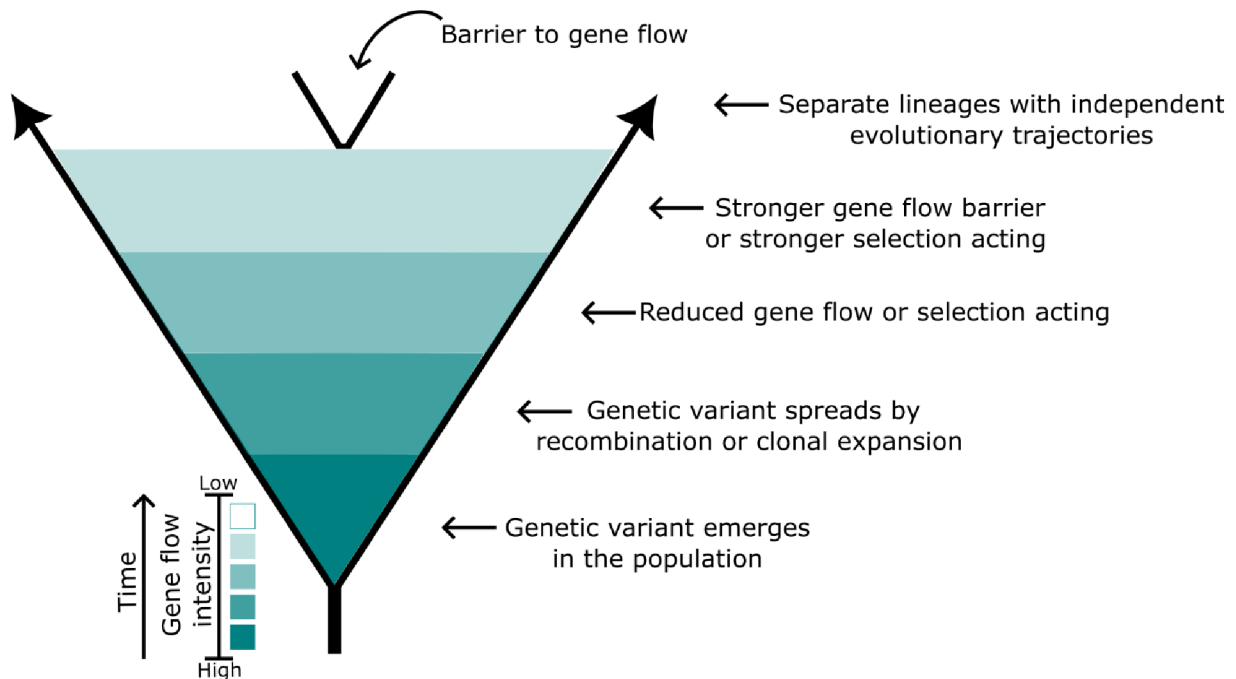
Painting the speciation continuum

After assembling the genome dataset and designating individual strains to populations, we are left to investigate phylogenetic signals in single nucleotide polymorphisms (SNPs) and analyze genome-wide patterns of divergence and gene flow among lineages. The wide array of population genomic measures such as nucleotide diversity (π), fixation index (F_{ST}), absolute divergence (d_{XY}), and linkage disequilibrium (r^2) can be utilized to estimate intra- and interpopulation-genetic differentiation between sequenced individuals (Krause & Whitaker, 2015). Coupling these metrics with gene flow and selection estimates allows us to capture speciation in progress while having a slightly clearer picture of the dynamic structure of cryptic species (Papers II and III). Eventually, we can paint the grey zone and link all the observed patterns of genomic diversity with the specific stages of the speciation continuum.

It is still thoroughly debated what the speciation continuum is and whether the concept is pragmatic enough (Stankowski & Ravinet, 2021). According to the authors, we can define it as "a continuum of reproductive isolation". Moreover, the concept of speciation continuum provides an opportunity to track the progress of speciation and inform us about mechanisms affecting it, including genetic, ecological, spatial, and morphological divergence (Stankowski & Ravinet, 2021; Bolnick et al., 2022). A recent rise of this notion that speciation does not have a distinct starting- or endpoint but is instead a process signifies a profound shift in how we conceptualize and observe species (Nosil & Feder, 2012; Kollár et al., 2022). Namely, populations accumulate genetic variation at a fairly

gradual pace for millions of years, which triggers a continual growth of barriers to gene flow (Coyne & Orr, 2004). Observing speciation as a process accentuates the rare occurrence of discrete events, although rapid speciation has been recorded in populations of Darwin's finches, cichlid fish, apple flies, silverswords, and sunflowers (see Marques et al., 2019). Accordingly, we need to observe species through the perspective of the probabilistic framework; higher genetic distances and restricted gene flow between units (population/lineage/species) indicate a higher likelihood of their divergence (Kollár et al., 2022). By performing genome scans between closely related co-occurring and incipient species, we can obtain so-called snapshots of the genomic changes, as the genomic regions show a strong signature of diversifying selection early in speciation (Nosil & Feder, 2012). Hence, it seems more worthwhile to consider placing these units along the speciation continuum, from the point of the emergence of a genetic variant and the absence of a barrier to gene flow to the complete cessation of gene flow between distinct species.

Following Shapiro & Polz (2015) and Kollár et al. (2022), we can represent a speciation continuum through a simplified scheme, as shown in **Figure 5**, and explain the different stages at which

Figure 5. Graphical depiction of the speciation continuum and the grey zone of speciation. The scheme represents two cyanobacterial species diverging through time from one population. The intensity of gene flow between individuals of a population varies from the highest at the beginning (teal) to the lowest at the end (white). Lineages go through several intermediate phases, from acquiring a genetic variant to being on separate evolutionary paths.

lineages/species can be. Nevertheless, the boundaries between the stages are blurry and certainly not obvious as they are continuous. Also, lineages/species can stay at certain stages for millions of years without ever reaching a full divergence, or they can even back merge (Fraser et al., 2009).

Stage 1 is characterized by a group of closely related individuals (population) of cyanobacteria that coexist and freely engage in genetic exchange, i.e., there are no barriers to gene flow. For instance, the appearance of a new ecological (micro)niche with an abundance of new resources can trigger the genetic (adaptive) variant's emergence via mutation or HGT. No phenotypic characters would distinguish individuals having the new variant at this stage.

Stage 2 is the spread of the adaptive variant from one individual to another with a minor barrier to gene flow. Suppose individuals have a more clonal nature. In that case, the adaptive variant would spread by clonal expansion, purging the diversity genome-wide in a population. Alternatively, if individuals are frequently subjected to HR, then the adaptive variant would spread by it and, thus, not purge the diversity genome-wide. These are the early stages of speciation when nascent populations are still connected by gene flow and cannot be differentiated from the ancestral population (Shapiro & Polz, 2015).

Stage 3 occurs when two populations "bud off" from the ancestral one by occupying distinct ecological microniches. They differentially respond to some environmental factors but are still connected via intermediate levels of gene flow. With the emergence of such an ecological barrier to gene flow, genetic separation synchronously arises in neutral loci. A potential example from microbes could be *Vibrio* (Shapiro et al., 2012) and *Microcoleus* (Paper III), which are at the 2-3 stage of divergence. While in the grey zone, populations can divide, merge, and go extinct, sometimes never reaching terminal stages of speciation (Coyne & Orr, 2004).

Patterns of genetic differentiation typical for stages in the early process of speciation occurring in eukaryotes include, for instance, butterflies (Martin et al., 2013), mussels (Roux et al., 2016), corals (Roux et al., 2016), fish (see Hendry et al., 2009), sunflowers (Andrew & Rieseberg, 2013).

Stage 4 is characterized by the emergence of a strong genetic barrier (e.g., sufficient sequence divergence), and HR becomes even more reduced over the whole genomes. We recognize two populations as species heading toward different evolutionary trajectories between the third and

fourth stages (Shapiro & Polz, 2015). Potential examples of microbes at stages 3-4 include *Sulfolobus* (Cadillo-Quiroz et al., 2012) and *Laspinema* (Paper II).

Stage 5 is the end of a specific speciation process (although note that speciation is not a definite process), characterized by the absence of gene flow between genetically and ecologically distinct lineages. Based on the extent of gene flow, they have a very high probability of splitting in two and will probably never merge again (but they can go extinct). It is important to underline here that these stages are not any concepts or definitions of species but rather helpful descriptions of states in which intermediate species can be.

Taking over the world – the heroes of the world dominance

Throughout this thesis, it was mentioned that cyanobacteria could rapidly adapt to changing environments, which allows them to inhabit the most diverse and extreme ecosystems at all altitudes and latitudes. This ability is repeatedly emphasized here as this factor makes cyanobacteria the most successful and essential organisms ever to have lived in Earth's history. This thesis focused on filamentous cyanobacteria thriving in soil crusts – *Microcoleus* and *Laspinema*.

Terrestrial habitats harbor many different microenvironments. Biological soil crusts covering approximately 12% of Earth's land area are particularly interesting (Rodriguez-Caballero et al., 2018). They represent assemblages of sediment or soil particles with bacteria, cyanobacteria, fungi, lichens, bryophytes, and microfauna (Belnap and Lange, 2003). Biocrusts are crucial components of arid and semi-arid regions, where they play a key role in soil development, water retention, nutrient cycling, soil temperature, and plant community development (Weber et al., 2016; Rodríguez-Caballero et al., 2012). All of the roles above could be almost entirely assigned to cyanobacteria, which tend to dominate biocrusts. Therefore, it is no surprise that they are often referred to as small engineers of ecosystems. Integral cyanobacteria that make soil crusts are mostly filamentous or unicellular species; for instance, *Nostoc commune*, *Microcoleus vaginatus*, *Coleofasciculus chtonoplastes*, *Trichocoleus sociatus*, *"Oscillatoria" acuminata* (*Laspinema*

acuminatum), *Scytonema*, *Leptolyngbya*, *Chrococcidiopsis*, *Gloeocapsa* (Garcia-Pichel et al., 2001; Büdel et al., 2016).

Microcoleus is a well-known genus first described in the 19th century by Desmazières and later redefined by Gomont (1892). It is a terrestrial, filamentous, with a calyptra, bundle-forming cyanobacterium occurring in biocrusts (Garcia-Pichel & Wojciechowski, 2009). Although at first glance, *Microcoleus* appears to be an ordinary green filament among billions and billions of other entangled filaments, looks can be deceiving. In the grand scheme, *Microcoleus* plays a crucial role in shaping the dynamics of local microniches, soil properties, nutrient cycling, and sustaining the life of various organisms (Belnap & Gardner, 1993). Specifically, it binds and holds soil particles, stabilizing the substrate by producing a fine matrix of exopolysaccharides. In that way, it protects the soil from erosion. It also excretes many photosynthates directly into the soil, augmenting its quality and promoting the growth of other bacteria and plants (Baran et al., 2015). *Microcoleus* also plays a role in water retention, which is extremely important in arid and semi-arid habitats (Belnap & Gardner, 1993). According to the molecular evidence, the genus is polyphyletic (Siegesmund et al., 2008; Hašler et al., 2012) and underwent a series of taxonomic revisions over the last two decades. From the existence of two species within *Microcoleus* (Boyer et al., 2002) to the establishment of a new family – *Microcoleaceae* and more than a dozen *Microcoleus* species, according to Strunecký et al. (2013). The same authors defined the type species for *Microcoleus vaginatus*. Currently, *Microcoleus* species are only vaguely defined, so the framework for their delimitation is bound to be proposed in future studies.

Microcoleus can serve as a perfect model system for studying distribution and divergence patterns in cyanobacteria. The first reason is its representation in all parts of the world. Then, it is easily cultivated and distinguished from other cyanobacteria by morphological features. The genus includes many recently diverged groups, ideal for studying the species' emergence.

Laspinema is a newly erected genus within the recently established family of *Laspinemaceae* (Heidari et al., 2018). The genus occurs in diverse terrestrial habitats, from ones with extremely high temperatures and radioactive contamination to microbial mats of puddles, e.g., hyposaline puddles (Dadheech et al., 2013), saline ponds (Casamatta et al., 2005), thermal springs and muds (Heidari et al., 2018; Duval et al., 2020). The type species is *Laspinema thermale*. The genus *Laspinema* includes only several species transferred from *Phromidium* and *Oscillatoria* genera

(Heidari et al., 2018; Zimba et al., 2020). This genus is still undersampled and consists of recently diversified species, making it an ideal organism for delimitation and evolutionary studies.

In Papers I and III, I explored phylogenetic relationships between the species and uncovered the evolutionary forces driving the speciation of *Microcoleus*, while in Paper II, I used *Laspinema* for those purposes.

Research aims

Nearly five years of my doctoral studies were dedicated to bringing the focus of evolutionary studies on remarkable less-studied cyanobacteria. The underlying goal of my thesis has been to characterize evolutionary forces governing genetic variation that might lead to the diversification of microbial species from a global to a local scale. Specifically, I have studied speciation patterns by centering on genome-wide signatures of diversity and divergence. My focus was on cryptic filamentous cyanobacteria - *Microcoleus* and *Laspinema*.

For **Paper I**, we aimed to establish a large collection culture of globally distributed strains on a population level, a rarely used approach for cyanobacteria. I sequenced 16S rRNA and 16S-23S ITS to investigate drivers of diversification and biogeographic distribution of *Microcoleus*. Additionally, the inferred global population structure of *Microcoleus* served as a priori information for subsequent genome analyses in Paper III.

In **Paper II**, I shifted focus to explore patterns of genetic variability and evolutionary mechanisms generating it on a local scale. Specifically, I scanned the genome for signatures of selection and recombination to seek specific adaptive genes responsible for the differentiation of diverging species of *Laspinema*.

Using the population genomics approach developed in Paper II and the dataset from Paper I, I sought to investigate local and global patterns of genomic diversity and the emergence of the continuum of *Microcoleus* species. I examined genome-wide inter- and intraspecific signatures of local selection and HR to provide a better understanding of the potential cause of species' divergence. Ultimately, I estimated whether *Microcoleus* lineages are on separate evolutionary trajectories, i.e., placed them on the speciation continuum (**Paper III**).

Summary of the results

Paper I. Geography and climate drive the evolutionary and biogeographic patterns in *Microcoleus*

Baas-Becking's postulate "everything is everywhere, but the environment selects" has sparked many investigations on microbial ubiquity over the last century. While some investigations relied on microscopy to identify patterns of diversity and distribution (Finlay, 2002; Fenchel, 2003), others utilized various molecular and sequencing methods, which improved our pursuit of fine diversity among prokaryotes. These molecular studies uncovered that different cyanobacteria exhibit different distribution and evolutionary patterns (e.g., *Chrococcidiopsis*, Bahl et al., 2011; *Microcoleus*, Dvořák et al., 2012; *Raphidiopsis* and *Microcystis*, Ribeiro et al., 2020), rejecting or affirming the above-mentioned tenet.

Extending from previous findings, for **Paper I,** I used a large sequence dataset (16S rRNA and 16S-23S ITS) of 495 *Microcoleus* strains originating from all continents besides South America to investigate the phylogeographic structure and the underlying factors governing the diversification. I found tremendous genetic diversity and observed 13 lineages within *Microcoleus vaginatus*, which might be distinct species. Phylogenetic inference and ancestral state reconstruction (habitat and geographic area) revealed fine phylogeographic and structuring by the habitat of strains within respective lineages. However, some strains still deviated from this structuring pattern, implying the existence of gene flow between lineages, e.g., between North America and Europe (the intensities of gene flow between them were reconstructed in Paper III). Moreover, the sampling design on a population level allowed us to identify the cooccurrence of up to four distinct lineages at one sampling site. Such a coexistence could be explained by fine differences in (micro) niche exploitation between the lineages (Kashtan et al., 2014; Shapiro & Polz, 2015; Papers II and III).

By examining the congruence between genetic diversity and geography, dispersal limitation was found to affect the distribution and diversification of *Microcoleus* lineages. Concretely, lineages tend to be genetically more similar as the physical distance decrease, although geographic barriers might be temporary (Bahl et al., 2011; Dvořák et al., 2012). Therefore, these observations stand at odds with Baas-Becking's assumption. Additionally, phylogenetic signal analyses and the Mantel

test revealed that climate also affects the diversification of *Microcoleus*, specifically temperature and precipitation.

This study contributes to a significant question in microbial ecology and evolution: what mechanisms elicit and maintain diversification in soil-dwelling cyanobacteria? Collectively, the findings in this paper indicate that the genetic structure of *Microcoleus* is affected by both isolations by distance and by the environment.

Paper II. Indications of recent, cryptic speciation within cyanobacterium *Laspinema*

The coexistence of multiple closely related but distinct lineages in sympatry appears to be a characteristic feature of free-living (cyano)bacteria (e.g., Kashtan et al., 2014). Molecular markers traditionally used in the study of cyanobacteria (16S rRNA and 16S-23S ITS) are not sensitive enough to capture fine genetic diversity between cryptic, closely related species and infer drivers of variability. In **Paper II**, I utilized the population genomics approach to explore the extent of local genomic diversity in a centimeter-scaled habitat (puddle) and the mechanisms responsible for the sympatric speciation of *Laspinema thermale*.

From the two samples, nine *Laspinema thermale* strains were obtained. Their 16S rRNA and 16S-23S ITS were sequenced to confirm that strains belong to one species and inspect the population structure, which was then reproduced using whole genome data. I found two cryptic species of *Laspinema*, highly genetically differentiated and separated by the barrier to gene flow, which could be of genetic and ecological origin.

Before inferring the drivers of their divergence, cyanobacterial species must be properly delimited first. A multifaceted approach[3] was used to address this issue, from sequence thresholds (ANI and 16S rRNA), species-specific genes (the pangenome), and phylogenomics to gene flow. I found that consistent monophyly of two clades and high HR levels within them (see below for details) best delineated *Laspinema* into two species (for reference, ANI yielded five, p-distance of 16S one

---

[3] We have also checked the strains' morphological features and predicted their ITS secondary structures. Delineation of *Laspinema* strains was not possible with these approaches. (data not shown)

population/species). Notably, both had their own block (set) of genes within the pangenome, but close to 77% of the genes had an unknown function.

By measuring population genomic metrics (F_{ST}, d_{XY}, π), a high intra- and inter-population genetic differentiation of the species was revealed, suggesting that they form cohesive genetic units with a substantial degree of divergence. In particular, in the areas of high divergence, I found 26 putative adaptive genes associated with the response to high/low light and stress conditions as well as with nutrient uptake. However, there was an indication of only a weak positive selection acting on them. This implies a potential role of ecological selection in driving speciation, where species are differentially adapted to the conditions in microniches (Shapiro et al., 2012; Cadillo-Quiroz et al., 2012).

Next, I estimated whether HR contributes to the maintenance of such cohesive genetic groups. Calculated HR rates (r/m) of individual strains varied from 0.1-1.6, confirming previous analyses that free-living prokaryotes engage in less recombination than mutation (Vos & Didelot, 2009). Even though HR occurred less frequently, it still introduced almost 1.5 times more substitutions than mutations in one lineage, indicating its importance in shaping genetic diversity (lower HR effect in another could be due to the small sampling size). More importantly, a pattern of preferential HR within the lineages (r/m = 1.3 and 0.8) than between them (r/m = 0.6) highlights the existence of a barrier to gene flow, which, if we observe as reproductive isolation, then the BSC concept can apply to our dataset.

Taking all the patterns of genetic divergence, selection, and HR, we painted the speciation continuum and positioned our *Laspinema* species at stages 3 or 4. Potential niche-specific variants have already emerged at these speciation stages, followed by a weak diversifying selection and a barrier to gene flow restricting HR over most genomic regions. Hence, the diversification of *Laspinema* species might have been driven by genomic and ecological processes. Future studies should aim to investigate the nature of barriers to gene flow more closely.

Paper III. Exploring the emergence of the global continuum of *Microcoleus* species

Microbial species emerge through a dynamic and continuous process called speciation, which involves the evolution of genetic and ecological differentiation across the genome (Shapiro & Polz, 2014). The often gradual acquisition of genetic variants within a microbial population is mediated by the complex coaction of evolutionary forces such as selection, gene flow, mutations, and genetic drift (Reno et al., 2009). However, understanding the relative importance of mechanisms by which genetic differences accumulate over time and subsequently lead to the emergence of new species remains unclear. **Paper III** illustrates the largest speciation continuum observed and the most comprehensive examination of genomic differentiation on a local and global scale in free-living terrestrial prokaryotes, especially cyanobacteria.

In **Paper III**, I utilized the population genomics approach established in Paper II and the global dataset of *Microcoleus* strains from Paper I. The population structure of *Microcoleus* from Paper I was reproduced by sequencing 202 genomes with an addition of 89 *Microcoleus* genomes obtained from the GenBank database and eight genomes from herbarium specimens. I found a continuum of at least 13 species, varying in genetic and ecological differentiation levels, reflecting the global speciation continuum.

The dating analysis of the 16S rRNA and Bayesian skyline plots were used to investigate species-splitting patterns through time and species demographic histories. The initial radiation of *Microcoleus* commenced in the Eocene/Oligocene (before 29.6 million years ago) when the climate was significantly warmer and wetter than it is today. Interestingly, this split concurs with the global expansion of arid regions, which began 34 million years ago (Sun & Windley, 2015). Furthermore, the diversification transpired throughout the middle Miocene and Pliocene (13.7-4.7 million years ago), coinciding with simultaneous significant population expansions of several *Microcoleus* species. In contrast, other species underwent demographic fluctuations in the Pleistocene (less than 500 thousand years ago). Therefore, higher global aridity and climate shifts affected the demographic patterns and diversification of *Microcoleus*.

The significance of Mantels' and phylogenetic signal tests confirmed that the environment is responsible for the diversification patterns of *Microcoleus*. Selective pressures were likely introduced by the precipitation levels, soil properties like the volume of coarse fragments or

organic carbon levels, and, more recently, by human activities (such as changes in net primary production of ecosystems). Additionally, the differences in their climatic preferences had already emerged as some of them occupied environments with high precipitation and wide temperature ranges, while others favored cooler and drier habitats. Besides the adaptation to specific microniches, the Mantel test showed that geographic separation also played a significant role in generating the diversity within *Microcoleus* (r = 0.44, p < 0.0001), endorsing the patterns observed in Paper I.

The signature of ecology on the genetic differentiation of *Microcoleus* is reflected in its huge flexible genome, which encompasses almost 95% of its total pangenome. This suggests that differences in the fitness of *Microcoleus* lineages are dictated by their ability to respond quickly to environmental pressures (Polz et al., 2013) and colonize novel ecological niches in inhospitable soil systems. In fact, by performing genome scans, I discovered genomic regions of elevated genetic differentiation (99th percentile of F_{ST} and D_{XY}) containing 28 genes related to stress response and biosynthesis. In addition, the McDonald-Kreitman test showed a strong effect of positive selection on these loci, confirming the ongoing ecological differentiation along the *Microcoleus* speciation continuum.

The extensive ecological differentiation was followed by the vast genetic differentiation across the whole genome, with F_{ST} spanning from 0.2 to 0.93. The wide range of calculated F_{ST} values indicates that *Microcoleus* species are at various stages across the continuum of genetic divergence. Despite the high levels of genetic differentiation, gene flow was still substantial, with up to 70% of the genome subjected to recombination among them. While recombination was not occurring often within *Microcoleus* (mean ρ/θ, 0.012-0.041), it still significantly contributed to its genetic diversity (mean r/m, 0.38-2). Remarkably, six of the 12 species had one to a few highly recombinant genotypes (n =23, r/m>2) that might be responsible for the removal of deleterious mutations by accelerating the rate of adaptation through fitness-associated recombination (Hadany & Becker, 2003). Similar to *Laspinema* (Paper II), *Microcystis* (Pérez-Carrascal et al., 2019), and *Sulfolobus* (Cadillo-Quiroz et al., 2012), *Microcoleus* species underwent preferential gene flow within species rather than between, underscoring the existence of a barrier to gene flow.

The overall pattern of high genome-wide divergence, gene flow, and selection across the speciation continuum allowed me to paint it by utilizing a novel approach to species delineation known as

UPCEL (Kollár et al., 2022). Together with the microbial speciation model of Shapiro & Polz (2015), I estimated that five species were on stages 4-5 with a high probability (>93%) of becoming fully separate species, three remained in the grey zone (stages 2-3) with a >85% probability of separating, and the least diverged four species remained in the early speciation stages (1-2), with >73% probability of becoming fully separate species. Collectively, findings from Paper III suggest that the speciation of *Microcoleus* species entails a complex interplay of gene flow, selection, and ecological and geographic factors. *Microcoleus* species provide a full continuum of taxa pairs at varying stages of reproductive isolation – from freely engaging in gene flow to species with largely established barriers to gene flow.

Conclusions and the outlook for the future

The broad view of this dissertation was to open the doors into the wondrous world of cyanobacteria – what factors governed the species' emergence and the global competitive success in highly multiplexed terrestrial ecosystems? Until only a decade ago, sequencing technologies, molecular techniques, and bioinformatic tools were not adequate to address these key evolutionary questions in prokaryotes. Since then, many have utilized comparative and population genomics and provided exciting ideas on species concepts suitable for microbes, delineation of cryptic species, and how microbes evolve in once inhospitable habitats. The chapters of this thesis can also be seen as a contribution toward guiding cyanobacteriologists and microbiologists to apply various approaches and concepts and delineate the microbial species based on the relative importance of evolutionary forces to their divergence. While the work presented here tackles these questions and contributes to the overall debate on molecular mechanisms of microbial evolution, many questions are getting raised.

What is the extent of cyanobacterial diversity in heterogeneous soil systems?

The traditional approach of investigating cyanobacterial genetic diversity is still based on phenotypic features and sequencing conservative molecular markers like 16S and 16S-23S ITS, often from a single or few strains per sampling site. In recent years, multiple papers focused on the local diversity of aquatic model cyanobacteria, *Prochlorococcus* (Kashtan et al., 2013, 2017), and *Microcystis* (Pérez-Carrascal et al., 2019) and showed how the actual fine genomic diversity could be immense. However, cryptic diversity on a global scale and in terrestrial cyanobacteria remains unexplored. Our population-level sequencing revealed a vast genetic diversity of filamentous cyanobacteria, which is reflected in the discovery of at least 12 recently diversified distinct *Microcoleus* species on a global scale. Alternatively, at least two *Laspinema* and four *Microcoleus* genetically differentiated species coexisted in sympatry. Further, we captured *Microcoleus* species at all stages of genetic differentiation, which allowed us to fill the grey zone of the speciation continuum and follow speciation from the start until the near finish. Naturally, including even more genomes of closely related strains from unexplored habitats is an evident continuation of this work (for instance, from tropical parts). Additionally, utilizing a population

genomic approach to other cryptic and non-model cyanobacterial species is necessary to understand the commonality of intermediate speciation stages in natural microbial populations.

What factors cause genetic variability and, ultimately, the emergence of cyanobacteria?

The analyses of congruencies between global and local genetic diversity of soil cyanobacteria *Microcoleus* and *Laspinema* with various environmental (temperature, precipitation, soil variables, UV light) and physical forces revealed them as operating in tandem, influencing the species divergence. Globally distributed *Microcoleus* species are still genetically connected, and the gene flow maintains them at various stages of speciation. Conversely, HR was more frequent within than between *Laspinema* species, which indicates that they might be on the way to being two separate species following the BSC concept, while barriers to gene flow maintain them on separate evolutionary trajectories. Investigating the nature of emerging barriers to gene flow and what causes them is important for future detailed studies on mechanisms driving diversification. Moreover, by estimating gene flow rates, one could investigate genetic connectivity between diverging, closely related species and trace the impact of evolutionary forces backward and forward on the speciation continuum. Collecting multivariate data for a different model and non-model species will provide a platform for future investigations of species' evolutionary trajectories and the relative contribution of forces affecting them.

What is the function of the genes associated with the global adaptive potential of cyanobacteria?

Our work on the genes found in arenas of high genetic differentiation and recombination suppression addresses the potential role of niche-specific genes in the diversification of cyanobacteria. For instance, genes associated with light could drive one *Laspinema* species to burrow deeper in soil, or perhaps one species is more efficient in uptaking certain nutrients. Alternatively, important for *Microcoleus* were the genes linked to stress response and biosynthesis (proteins, cofactors), possibly allowing these cyanobacteria to quickly respond to the changing high-stress environmental conditions in drylands, including periods of extreme draughts and heavy precipitations. Still, further characterization of genes is needed, e.g., via transcriptomic, epigenomic, and experimental physiological studies, given the high percentage of hypothetical

cyanobacterial genes. Besides, recent advances in techniques commonly used in eukaryotic models, like genome-wide association study (GWAS), could be enforced to identify genomic regions associated with prokaryote adaptive traits (Sheppard et al., 2013). Understanding the function of adaptive genes will aid our exploration of cyanobacterial responses to conditions in various unknown microniches in the soil as well as of pangenome dynamics.

In the current era of genomics dominance, we have already made critical steps forward in the research of prokaryotic evolution. Now it is time for cyanobacteria to be in the spotlight. With this Ph.D. thesis finished, it can serve as a starting point for an even more thorough exploration of the evolutionary secrets embedded deep into the cyanobacterial genome. After all, we have much work ahead to find the answer to the "Great Question" of microbial "Life, the Universe, and Everything".

Acknowledgments

Before anyone else, I would like to thank and express my deepest appreciation to my supervisor, **Petr Dvořák**. Thank you for your selfless guidance and the great understanding you showed all these years. Your seemingly endless consideration always managed to steer me back on the path whenever I felt the pressure of work or personal challenges. You knew when to apply pressure, offer a good laugh, and extend a helping hand. Thank you for creating such an inspirational and comforting atmosphere in our group. You have been an outstanding scientific inspiration and I have no doubt that you will continue to push the frontiers of science. I could not wish for a better and more honest bond with a supervisor. Thank you.

Then, I would like to thank the members of our small yet incredible algological group. I am highly grateful to **Aloisie Poulíčková** for introducing me to the Department, connecting me with my supervisor and other colleagues, and always bringing positivity into the office despite being so busy at times. Also, I would like to thank **Petr Hašler** for sharing his expertise on cyanobacteria with me, as well as for hours of intriguing conversations on Czech culture, history, and music. Special thanks for both of your support and efforts so that I could improve and start speaking Czech more freely. Although I spent most of my PhD working on a computer alone, I absolutely enjoyed the hours I spent in the lab working with **Svát'a** on our humongous collection culture. Thank you for all the great laughs and talks in the lab and outside of it and also for the delicious treats you were baking for us all. And **Eliška**, I am very grateful for the lovely pep talks you gave me when I was just completely absorbed in my work and writing. Thank you both for the occasional reality checks, and I have my fingers crossed that you will write fantastic PhD theses soon and go through academic pressures with ease.

I would like to thank the whole Department of Botany (UP) and the people who helped me settle in nicely at the beginning of my PhD and later for all our events and parties. I had the luck and the most incredible pleasure of working at the Department of Organismal Biology (UU) in Sweden, where I met many wonderful people and exceptional scientists. Many thanks to **Hanna Johannesson**, who hosted me in her lab for six months in very turbulent pandemic times and offered a glimpse into the amazing world of evolution in fungi. Also, I am deeply grateful to

Markus for guiding my lab and computational work on *Marasmius*, who did a lot so I could feel extra comfortable in their team. Thank you both for your patience and inspiration.

The road to becoming a PhD is filled not only with constant professional endeavors but also countless battles with one's inner self. I soon came to the realization that I could not undertake those alone, and I couldn't be happier that I befriended many extraordinary people along the way. **Pablo** has been a pillar of my strength and sanity, and I owe you the deepest gratitude for sticking with me all the way through. Thank you for all the cozy late-night balcony talks, for listening to my scientific outbursts, and for showing me how an honest and loving friendship can truly be. This thesis is a bit of yours as well. **Tijana**, zaslužuješ jedno preogromno hvala, što si bezrezervno bila uz mene svih ovih divnih godina i što smo nekako zajedno uspeli da prebrodimo sve napore akademije. Hvala ti na prijateljstvu i nadam se da ćemo posle doktorata da se vratimo na zajednička putovanja po svetu. **Andrea** and **Karito**, thank you both for being there for me since the very first days of my PhD journey. Having two of you over our dinners and travels together was one of the most therapeutic things that kept me going through rough Olo days. **Jovanka**, thank you for being an amazing support during this road of mine as well as the company over the past couple of years, which actually seem like decades. Za još puno nezaboravnih uspomena! Draga **Anita**, tebi hvala za sve trenutke provedene zajedno i sve žurke koje smo organizovali, a naročito hvala za kave i naše dugo kratke pauze od dnevnih bioloških drama. Also, thank you, **Veera**, for going through a significant period of PhD journey with me, and thanks for your unwavering support and loving words that pushed and made me a better version of myself. I would also like to acknowledge all my dear friends who inspired me along the way and made a significant spot engraved in my heart for them. Thank you **Kateřina**, **Milanče**, **Duje**, **Yulia**, **Kiril**, **Muha**, **Sanja**, **Karel**, **Tanya**, **Marko**, **Magda**… A vas dve, **Mila** i **Jeko**, vas čekam na mrkim paradnim medvedima da nastavimo naše zajedničke pohode.

Finally, this thesis would not be possible without my **parents** and **aunt,** to whom I am indebted for all the invaluable support they provided along the road. **Marija**, thanks for being there for our honest and lovely life talks and rants every Thursday, which kept me sane during swirling periods. To my beloved brother **Ljuba**, I dedicate this thesis as a tribute to your memory; now, it will never fade.

References

Andrew, R. L. & Rieseberg, L. H. 2013. Divergence is focused on few genomic regions early in speciation: incipient speciation of sunflower ecotypes. *Evolution*, 67, 2468-2482. doi: 10.1111/evo.12106

Arevalo, P., VanInsberghe, D., & Polz, M. F. 2018. A reverse ecology framework for bacteria and archaea. In: Polz, M. F. & Rajora, O. P. (Eds.) *Population Genomics: Microorganisms*. Springer, Cham, 77-96. doi: 10.1007/978-3-030-04756-6

Arevalo, P., VanInsberghe, D., Elsherbini, J., Gore, J., & Polz, M. F. 2019. A reverse ecology approach based on a biological definition of microbial populations. *Cell*, 178, 820-834. doi: 10.1016/j.cell.2019.06.033

Baas-Becking, L. G. M. 1934. Geobiologie of Inleiding tot de Milieukunde. W. P. Van Stockum & Zoon, The Hague.

Bahl, J., Lau, M. C. Y., Smith, G. J. D., Vijaykrishna, D., Cary, S. C., Lacap, D. C., Lee, C. K., Papke, R. T., WarrenRhodes, K. A., Wong, F. K. Y., McKay, C. P., & Pointing, S. B. 2011. Ancient origins determine global biogeography of hot and cold desert cyanobacteria. *Nature Communications*, 2, 1–6. doi: 10.1038/ncomms1167

Baran, R., Brodie, E. L., Mayberry-Lewis, J., Hummel, E., Da Rocha, U. N., Chakraborty, R., Bowen, B. P., Karaoz, U., Cadillo-Quiroz, H., Garcia-Pichel, F., & Northen, T.R. 2015. Exometabolite niche partitioning among sympatric soil bacteria. *Nature Communications*, 6, 8289. doi: 10.1038/ncomms9289

Baum, D. A. & Shaw, K. L. 1995. Genealogical perspectives on the species problem. *Experimental and molecular approaches to plant biosystematics*, 53, 123-124.

Baumdicker, F., Hess, W. R., & Pfaffelhuber, P. 2012. The infinitely many genes model for the distributed genome of bacteria. *Genome biology and evolution*, 4, 443-456. doi: 10.1093/gbe/evs016

Belnap, J. & Gardner, J. S. 1993. Soil microstructure in soils of the Colorado Plateau: the role of the cyanobacterium Microcoleus vaginatus. *The Great Basin Naturalist*, 53, 40-47.

Belnap, J. & Lange, O. L. 2003. Structure and functioning of biological soil crusts: a synthesis. In: Belnap, J. & Lange, O. L. (Eds.) *Biological Soil Crusts: Structure, Function, And Management*. Springer, Cham, 471–479.

Biller, S. J., Berube, P. M., Lindell, D., & Chisholm, S. W. 2015. Prochlorococcus: the structure and function of collective diversity. *Nature Reviews Microbiology*, 13, 13-27. doi: 10.1038/nrmicro3378

Bobay, L. M. & Ochman, H. 2017. The evolution of bacterial genome architecture. *Frontiers in genetics*, 8, 72. doi: 10.3389/fgene.2017.00072

Bobay, L. M. 2020. The prokaryotic species concept and challenges. In: Tettelin, H. & Medini, D. (Eds.) *The Pangenome*. Springer, Cham, 21-49. doi: 10.1007/978-3-030-38281-0

Bolnick, D. I., Hund, A. K., Nosil, P., Peng, F., Ravinet, M., Stankowski, S., Subramanian, S., Wolf, J., & Yukilevich, R. 2022. A multivariate view of the speciation continuum. *Evolution*. doi: 10.1093/evolut/qpac004

Boto, L. 2010. Horizontal gene transfer in evolution: facts and challenges. *Proceedings of the Royal Society B: Biological Sciences*, 277, 819-827. doi: 10.1098/rspb.2009.1679

Boyer, S. L., Johansen, J. R., Flechtner, V. R., & Howard, G. L. 2002. Phylogeny and genetic variance in terrestrial *Microcoleus* (Cyanophyceae) species based on sequence analysis of the 16S rRNA gene and associated 16S–23S ITS region. *Journal of Phycology*, 38, 1222–1235. doi: 10.1046/j.1529-8817.2002.01168.x

Brenner, D. J., Staley, J., & Krieg, N. 2000. Classification of prokaryotic organisms and the concept of bacterial speciation. In: Boone, D. R., Castenholz, R. W., & Garrity, G. M. (Eds.) *Bergey's Manual of Systematic Biology*. 2nd edition. Springer, Boston, MA, 27-32

Büdel, B., Dulić, T., Darienko, T., Rybalka, N., & Friedl, T. 2016. Cyanobacteria and algae of biological soil crusts. In: Weber, B., Büdel, B., & Belnap, J. (Eds.) *Biological soil crusts: An Organizing Principle in Drylands*. Springer, Cham, 226, 55-80. doi: 10.1007/978-3-319-30214-0_4

Cadillo-Quiroz, H., Didelot, X., Held, N., Herrera, A., Darling, A., Reno, M., Krause, D. J., & Whitaker, R. J. 2012. Patterns of gene flow define species of thermophilic archaea. *PLoS Biology*, 10, e1001265. doi: 10.1371/journal.pbio.1001265

Carrolo, M., Pinto, F. R., Melo-Cristino, J., & Ramirez, M. 2009. Phenotypes are driving genetic differentiation within *Streptococcus pneumoniae*. *BMC Microbiology*, 9, 1–10. doi: 10.1186/1471-2180-9-191

Casamatta, D. A., Johansen, J. R., Vis, M. L., & Broadwater, S. T. 2005. Molecular and morphological characterization of ten polar and near-polar strains within the Oscillatoriales (Cyanobacteria) 1. *Journal of Phycology*, 41, 421-438. doi: 10.1111/j.1529-8817.2005.04062.x

Casamatta, D. A., Vis, M. L., & Sheath, R. G. 2003. Cryptic species in cyanobacterial systematics: a case study of *Phormidium retzii* (Oscillatoriales) using RAPD molecular markers and 16S rDNA sequence data. *Aquatic Botany*, 77, 295-309. doi: 10.1016/j.aquabot.2003.08.005

Chase, A. B., Arevalo, P., Brodie, E. L., Polz, M. F., Karaoz, U., & Martiny, J. B. 2019. Maintenance of sympatric and allopatric populations in free-living terrestrial bacteria. *MBio*, 10, e02361-19. doi: 10.1128/mBio.02361-19

Cohan, F. M. 2001. Bacterial species and speciation. *Systematic Biology*, 50, 513-524. doi: 10.1080/10635150118398

Cohan, F. M. 2006. Towards a conceptual and operational union of bacterial systematics, ecology, and evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361, 1985-1996. doi: 10.1098/rstb.2006.1918

Cordero, O. X., Ventouras, L. A., DeLong, E. F., & Polz, M. F. 2012. Public good dynamics drive evolution of iron acquisition strategies in natural bacterioplankton populations. *Proceedings of the National Academy of Sciences*, 109, 20059–20064. doi: 10.1073/pnas.1213344109

Coyne, J. A. & Orr, H. A. 2004. *Speciation*. Sunderland, MA.

Dadheech, P. K., Glöckner, G., Casper, P., Kotut, K., Mazzoni, C. J., Mbedi, S., & Krienitz, L. 2013. Cyanobacterial diversity in the hot spring, pelagic and benthic habitats of a tropical soda lake. *FEMS Microbiology Ecology*, 85, 389-401. doi: 10.1111/1574-6941.12128

De Queiroz, K. 2007. Species concepts and species delimitation. *Systematic Biology*, 56, 879-886. doi: 10.1080/10635150701701083

Dobrindt, U. & Reidl, J. 2000. Pathogenicity islands and phage conversion: Evolutionary aspects of bacterial pathogenesis. *International Journal of Medical Microbiology,* 290, 519–27. doi: 10.1016/S1438-4221(00)80017-X

Doolittle, W. F. & Zhaxybayeva, O. 2009. On the origin of prokaryotic species. *Genome Research*, 19, 744-756. doi: 10.1101/gr.086645.108

Duval, C., Hamlaoui, S., Piquet, B., Toutirais, G., Yepremian, C., Reinhart, A., Duperron, S., Marie, B., Demay, J., & Bernard, C. 2020. Characterization of cyanobacteria isolated from thermal muds of Balarucles-Bains (France) and description of a new genus and species *Pseudochroococcus couteii*. bioRxiv. doi: 10.1101/2020.12.12.422513

Dvořák, P., Casamatta, D. A., Poulíčková, A., Hašler, P., Ondřej, V., & Sanges, R. 2014. *Synechococcus*: 3 billion years of global dominance. *Molecular Ecology*, 23, 5538-5551. doi: 10.1111/mec.12948

Dvořák, P., Hašler, P., & Poulíčková, A. 2012. Phylogeography of the *Microcoleus vaginatus* (cyanobacteria) from three continents–a spatial and temporal characterization. *PLoS One*, 7, e40153. doi: 10.1371/journal.pone.0040153

Dvořák, P., Jahodářová, E., Casamatta, D. A., Hašler, P., & Poulíčková, A. 2018. Difference without distinction? Gaps in cyanobacterial systematics; when more is just too much. *Fottea*, 18, 130-136. doi: 10.5507/fot.2017.023

Dvořák, P., Poulíčková, A., Hašler, P., Belli, M., Casamatta, D. A., & Papini, A. 2015. Species concepts and speciation factors in cyanobacteria, with connection to the problems of diversity and classification. *Biodiversity and Conservation*, 24, 739-757. doi: 10.1007/s10531-015-0888-6

Fenchel, T. 2003. Biogeography for bacteria. *Science*, 301, 925–926. doi: 10.1126/science.1089242

Finlay, B. J. 2002. Global dispersal of free-living microbial eukaryote species. *Science*, 296: 1061–1063. doi: 10.1126/science.1070710

Foissner, W. 2006. Biogeography and dispersal of microorganisms: a review emphasizing protists. *Acta Protozoologica*, 45, 111–136.

Fox, G. E., Wisotzkey, J. D., & Jurtshuk Jr, P. 1992. How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. *International Journal of Systematic and Evolutionary Microbiology*, 42, 166-170. doi: 10.1099/00207713-42-1-166

Fraser, C., Alm, E. J., Polz, M. F., Spratt, B. G., & Hanage, W. P. 2009. The bacterial species challenge: making sense of genetic and ecological diversity. *Science*, 323, 741-746. doi: 10.1126/science.1159388

Fraser, C., Hanage, W. P., & Spratt, B. G. 2007. Recombination and the nature of bacterial speciation. *Science*, 315, 476-480. doi: 10.1126/science.1127573

Garcia-Pichel, F. & Wojciechowski, M. F. 2009. The evolution of a capacity to build supra-cellular ropes enabled filamentous cyanobacteria to colonize highly erodible substrates. *PLoS ONE*, 4, e7801. doi: 10.1371/journal.pone.0007801

Garcia-Pichel, F., Felde, V. J. M. N. L., Drahorad, S. L., & Weber, B. 2016. Microstructure and weathering processes within biological soil crusts. In: Weber, B., Büdel, B., & Belnap, J. (Eds.) *Biological soil crusts: An Organizing Principle in Drylands*. Springer, Cham, 226, 237-255.

Garcia-Pichel, F., Lopez-Cortez, A., & Nubel, U. 2001. Phylogenetic and morphological diversity of cyanobacteria in soil deserts crusts from the Colorado Plateau. *Applied and Environmental Microbiology*, 67, 1902–1910. doi: 10.1128/AEM.67.4.1902-1910.2001

Garcia-Pichel, F., Prufert-Bebout, L., & Muyzer, G. 1996. Phenotypic and phylogenetic analyses show *Microcoleus chthonoplastes* to be a cosmopolitan cyanobacterium. *Applied and Environmental Microbiology*, 62: 3284–3291. doi: 10.1128/aem.62.9.3284-3291.1996

Gaysina, L.A., Saraf, A., & Singh, P. 2019. Cyanobacteria in diverse habitats. In: *Cyanobacteria: From Basic Science to Applications*, 1st edition, Academic Press, London, 1–28. 10.1016/B978-0-12-814667-5.00001-5

Gevers, D., Cohan, F. M., Lawrence, J. G., Spratt, B. G., Coenye, T., Feil, E. J., Stackerbrandt, E., Van de Peer, Y., Vandamme, P., Thompson, F. L., & Swings, J. 2005. Re-evaluating prokaryotic species. *Nature Reviews Microbiology*, 3, 733-739. doi: 10.1038/nrmicro1236

Gogarten, J. P. & Townsend, J. P. 2005. Horizontal gene transfer, genome innovation and evolution. *Nature Reviews Microbiology*, 3, 679-687. doi: 10.1038/nrmicro1204

Goldenfeld, N. & Woese, C. 2007. Biology's next revolution. *Nature*, 445, 369-369. doi: 10.1038/445369a

Gomont, M. 1892. Monographie des Oscillariées (Nostocacées homocystées). Annales des Sciences Naturelles, Botanique, 7, 263–368.

Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P., & Tiedje, J. M. 2007. DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *International Journal of Systematic and Evolutionary Microbiology*, 57, 81–91. doi: 10.1099/ijs.0.64483-0

Grant, V. 1981. Plant speciation. New York: Columbia university press.

Guiry, M. D. & Guiry, G. M. 2022. AlgaeBase. At: www.algaebase.org (last accessed November 18, 2022).

Guiry, M. D. 2012. How many species of algae are there?. *Journal of Phycology*, 48, 1057-1063. doi: 10.1111/j.1529-8817.2012.01222.x

Hadany, L. & Beker, T. 2003. On the evolutionary advantage of fitness-associated recombination. *Genetics*, 165, 2167-2179. doi: 10.1093/genetics/165.4.2167

Hanage, W. P. 2013. Fuzzy species revisited. *BMC biology*, 11, 1-3. doi: 10.1186/1741-7007-11-41

Hanage, W. P., Fraser, C., & Spratt, B. G. 2005. Fuzzy species among recombinogenic bacteria. *BMC Biology*, *3*, 1-7. doi: 10.1186/1741-7007-3-6

Harke, M. J., Steffen, M. M., Gobler, C. J., Otten, T. G., Wilhelm, S. W., Wood, S. A., & Paerl, H. W. 2016. A review of the global ecology, genomics, and biogeography of the toxic cyanobacterium, *Microcystis* spp. *Harmful Algae*, 54, 4-20. doi: 10.1016/j.hal.2015.12.007

Hašler, P., Dvořák, P., Johansen, J. R., Kitner, M., Ondřej, V., & Poulíčková, A. 2012. Morphological and molecular study of epipelic filamentous genera *Phormidium*, *Microcoleus* and *Geitlerinema* (Oscillatoriales, Cyanophyta/Cyanobacteria). *Fottea*, 12, 341–356.

Heidari, F., Hauer, T., Zima, J. R. H., & Riahi, H. 2018. New simple trichal cyanobacterial taxa isolated from radioactive thermal springs. *Fottea*, 18, 137–149. doi: 10.5507/fot.2017.024

Hendry, A.P. 2009. Ecological speciation! Or the lack thereof?. *Canadian Journal of Fisheries and Aquatic Sciences*, 66, 1383-1398. doi: 10.1139/F09-074

Hennig, W. 1999. *Phylogenetic systematics*. University of Illinois Press.

Honda, D., Yokota, A., & Sugiyama, J. 1999. Detection of seven major evolutionary lineages in cyanobacteria based on the 16S rRNA gene sequence analysis with new sequences of five marine *Synechococcus* strains. *Journal of Molecular Evolution*, 48, 723-739. doi: 10.1007/PL00006517

Hugenholtz, P., Skarshewski, A., & Parks, D. H. 2016. Genome-based microbial taxonomy coming of age. *Cold Spring Harbor Perspectives in Biology*. 8, a018085. doi: 10.1101/cshperspect.a018085

Huisman, J., Codd, G. A., Paerl, H. W., Ibelings, B. W., Verspagen, J. M., & Visser, P. M. 2018. Cyanobacterial blooms. *Nature Reviews Microbiology*, 16, 471-483. doi: 10.1038/s41579-018-0040-1

Jeltsch, A. 2003. Maintenance of species identity and controlling speciation of bacteria: a new function for restriction/modification systems? *Gene*, 317, 13–16. doi: 10.1016/S0378-1119(03)00652-8

Johansen, J. R. & Casamatta, D. A. 2005. Recognizing cyanobacterial diversity through adoption of a new species paradigm. *Algological Studies*, 117, 71-93. doi: 10.1127/1864-1318/2005/0117-0071

Johansen, J. R., Kovacik, L., Casamatta, D. A., Iková, K. F., & Kastovský, J. 2011. Utility of 16S-23S ITS sequence and secondary structure for recognition of intrageneric and intergeneric limits within cyanobacterial taxa: *Leptolyngbya corticola* sp. nov. (Pseudanabaenaceae, Cyanobacteria). *Nova Hedwigia*, 92, 283. doi: 10.1127/0029-5035/2011/0092-0283

Kahru, M., Horstmann, U., & Rud, O. 1994. Satellite detection of increased cyanobacteria blooms in the Baltic Sea: Natural fluctuation or ecosystem change?. *Ambio*, 23, 469-472.

Kaplan-Levy, R. N., Hadas, O., Summers, M. L., Rücker, J., & Sukenik, A. 2010. Akinetes: dormant cells of cyanobacteria. In: Lubzens E., Cerda, J., & Clark, M. (Eds.) *Dormancy and Resistance in Harsh Environments. Topics in Current Genetics*. Springer, Berlin, 5–27. doi: 10.1007/978-3-642-12422-8_2

Kashtan, N., Roggensack, S. E., Berta-Thompson, J. W., Grinberg, M., Stepanauskas, R., & Chisholm, S. W. 2017. Fundamental differences in diversity and genomic population structure between Atlantic and Pacific *Prochlorococcus*. *The ISME Journal*, 11, 1997-2011. doi: 10.1038/ismej.2017.64

Kashtan, N., Roggensack, S. E., Rodrigue, S., Thompson, J. W., Biller, S. J., Coe, A., Ding, H., Marttinen, P., Malmstrom, R. R., Stocker, R., Follows, M. J., Stepanauskas, R., & Chisholm, S. W. 2014. Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. *Science*, 344, 416-420. doi: 10.1126/science.1248575

Kimura, M. 1983. *The neutral theory of molecular evolution*. Cambridge University Press.

Kollár, J., Poulíčková, A., & Dvořák, P. 2022. On the relativity of species, or the probabilistic solution to the species problem. *Molecular Ecology*, 31, 411–418. doi: 10.1111/mec.16218

Krause, D. J. & Whitaker, R. J. 2015. Inferring speciation processes from patterns of natural variation in microbial genomes. *Systematic Biology*, 64, 926-935. doi: 10.1093/sysbio/syv050

Kunkel, D. D. 1984. Cell division in baeocyte producing cyanobacteria. *Protoplasma*, 123, 104–115. doi: 10.1007/BF01283581

Lan, R. & Reeves, P. R. 1996. Gene transfer is a major factor in bacterial evolution. *Molecular Biology and Evolution*, 13, 47-55. doi: 10.1093/oxfordjournals.molbev.a025569

Leliaert, F., Verbruggen, H., Vanormelingen, P., Steen, F., López-Bautista, J. M., Zuccarello, G. C., & De Clerck, O. 2014. DNA-based species delimitation in algae. *European Journal of Phycology*, 49, 179–196. doi: 10.1080/09670262.2014.904524

Loza, V., Perona, E., & Mateo, P. 2014. Specific responses to nitrogen and phosphorus enrichment in cyanobacteria: Factors influencing changes in species dominance along eutrophic gradients. *Water Research*, 48, 622-631. doi: 10.1016/j.watres.2013.10.014

Luikart, G. & Cornuet, J. M. 1998. Empirical evaluation of a test for identifying recently bottlenecked populations from allele frequency data. *Conservation Biology*, 12, 228-237.

Lyons, T. W., Reinhard, C. T., & Planavsky, N. J. 2014. The rise of oxygen in Earth's early ocean and atmosphere. *Nature*, 506, 307-315. doi: 10.1038/nature13068

Marques, D. A., Meier, J. I., & Seehausen, O. 2019. A combinatorial view on speciation and adaptive radiation. *Trends in Ecology & Evolution*, 34, 531-544. doi: 10.1016/j.tree.2019.02.008

Martin, S. H., Dasmahapatra, K. K., Nadeau, N. J., Salazar, C., Walters, J. R., Simpson, F., Blaxter, M., Manica, A., Mallet, J. & Jiggins, C. D. 2013. Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Research*, 23, 1817-1828. doi: 10.1101/gr.159426.113

Martinez-Murcia, A. J., Benlloch, S., & Collins, M. D. 1992. Phylogenetic interrelationships of members of the genera Aeromonas and Plesiomonas as determined by 16S ribosomal DNA sequencing: lack of congruence with results of DNA-DNA hybridizations. *International Journal of Systematic and Evolutionary Microbiology*, 42, 412-421. doi: 00207713-42-3-412

Martiny, J. B. H., Bohannan, B. J., Brown, J. H., Colwell, R. K., Fuhrman, J. A., Green, J. L., Horner-Devine, M. C., Kane, M., Krumins, J. A., Kuske, C. R., Morin, P. J., Naeem, S., Øvreås, L., Reysenbach, A., Smith, V. H., & Staley, J. T. 2006. Microbial biogeography: putting microorganisms on the map. *Nature Reviews Microbiology*, 4, 102-112. doi: 10.1038/nrmicro1341

Mayr, E. 1940. Speciation phenomena in birds. *The American Naturalist*, 74, 249-278.

Mayr, E. 1942. Systematics and the origin of species. Columbia University Press.

Mayr, E. 1963. Animal species and evolution. Harvard University Press.

McDaniel L. D., Young, E., Delaney, J., Ruhnau, F., Ritchie, K. B., & Paul, J. H. 2010. High frequency of horizontal gene transfer in the oceans. *Science*, 330, 50. doi: 10.1126/science.1192243

Medini, D., Donati, C., Tettelin, H., Masignani, V., & Rappuoli, R. 2005. The microbial pan-genome. *Current Opinion in Genetics & Development*, 15, 589-594. doi: 10.1016/j.gde.2005.09.006

Meeks, J. C. & Elhai, J. 2002. Regulation of cellular differentiation in filamentous cyanobacteria in free-living and plant-associated symbiotic growth states. *Microbiology and Molecular Biology Reviews*, 66, 94–121. doi: 10.1128/MMBR.66.1.94-121.2002

Meier-Kolthoff, J. P., Auch, A. F., Klenk, H. P., & Göker, M. 2013. Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics*, 14, 60. doi: 10.1186/1471-2105-14-60

Mishler, B. D. & Theriot, E. C. 2000. The phylogenetic species concept (sensu Mishler and Theriot): monophyly, apomorphy, and phylogenetic species concepts. In: Wheeler, Q. D. & Meier, R. (Eds.) *Species concepts and phylogenetic theory, a Debate*. Columbia University Press, New York, 44–54.

Moldovan, M. A. & Gelfand, M. S. 2018. Pangenomic definition of prokaryotic species and the phylogenetic structure of *Prochlorococcus* spp. *Frontiers in Microbiology*, 9, 428. doi: 10.3389/fmicb.2018.00428

Muñoz-Martín, M. Á., Becerra-Absalón, I., Perona, E., Fernández-Valbuena, L., Garcia-Pichel, F., & Mateo, P. 2019. Cyanobacterial biocrust diversity in Mediterranean ecosystems along a latitudinal and climatic gradient. *New Phytologist*, 221, 123–141. doi: 10.1111/nph.15355

Nabout, J. C., da Silva Rocha, B., Carneiro, F. M., & Sant'Anna, C. L. 2013. How many species of Cyanobacteria are there? Using a discovery curve to predict the species number. *Biodiversity and Conservation*, 22, 2907-2918. doi: 10.1007/s10531-013-0561-x

Nosil, P. & Feder, J. L. 2012. Genomic divergence during speciation: causes and consequences. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367, 332-342. doi: 10.1098/rstb.2011.0263

Ochman, H., Lawrence, J. G., & Groisman, E. A. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405, 299–304. doi: 10.1038/35012500

Ohta, T. 1992. The nearly neutral theory of molecular evolution. *Annual Review of Ecology and Systematics*, 263-286.

Oren, A., Mareš, J., & Rippka, R. 2022. Validation of the names Cyanobacterium and Cyanobacterium stanieri, and proposal of Cyanobacteriota phyl. nov. *International Journal of Systematic and Evolutionary Microbiology*, 72, p. 005528. doi: 10.1099/ijsem.0.005528

Osorio-Santos, K., Pietrasiak, N., Bohunická, M., Miscoe, L. H., Kováčik, L., Martin, M. P., & Johansen, J. R. 2014. Seven new species of *Oculatella* (Pseudanabaenales, Cyanobacteria): taxonomically recognizing cryptic diversification. *European Journal of Phycology*, 49, 450-470. doi: 10.1080/09670262.2014.976843

Pauls, S. U., Nowak, C., Bálint, M., & Pfenninger, M. 2013. The impact of global climate change on genetic diversity within populations and species. *Molecular Ecology*, 22, 925-946. doi: 10.1111/mec.12152

Pérez-Carrascal, O. M., Terrat, Y., Giani, A., Fortin, N., Greer, C. W., Tromas, N., & Shapiro, B. J. 2019. Coherence of Microcystis species revealed through population genomics. *The ISME Journal*, 13, 2887-2900. doi: 10.1038/s41396-019-0481-1

Pietrasiak, N., Mühlsteinová, R., Siegesmund, M. A., & Johansen, J. R. 2014. Phylogenetic placement of *Symplocastrum* (Phormidiaceae, Cyanophyceae) with a new combination *S. californicum* and two new species: *S. flechtnerae* and *S. torsivum*. *Phycologia*, 53, 529-541. doi: 10.2216/14-029.1

Pietrasiak, N., Osorio-Santos, K., Shalygin, S., Martin, M. P., & Johansen, J. R. 2019. When is a lineage a species? A case study in *Myxacorys* gen. nov. (Synechococcales: Cyanobacteria) with the description of two new species from the Americas. *Journal of Phycology*, 55, 976-996. doi: 10.1111/jpy.12897

Popa, O. & Dagan, T. 2011. Trends and barriers to lateral gene transfer in prokaryotes. *Current Opinion in Microbiology*, 14, 615–623. doi: 10.1016/j.mib.2011.07.027

Reno, M. L., Held, N. L., Fields, C. J., Burke, P. V., & Whitaker, R. J. 2009. Biogeography of the Sulfolobus islandicus pan-genome. *Proceedings of the National Academy of Sciences*. 106, 8605–8610. doi: 10.1073/pnas.0808945106

Rest, J. S. & Mindell, D. P. 2003. Retroids in archaea: phylogeny and lateral origins. *Molecular Biology and Evolution*, 20, 1134-1142. doi: 10.1093/molbev/msg135

Ribeiro, K. F., Ferrero, A. P., Duarte, L., Turchetto-Zolet, A. C., & Crossetti, L. O. 2020. Comparative phylogeography of two free-living cosmopolitan cyanobacteria: Insights on biogeographic and latitudinal distribution. *Journal of Biogeography*, 47, 1106-1118. doi: 10.1111/jbi.13785

Ridley, M. 1989. The cladistic solution to the species problem. *Biology and Philosophy*, 4, 1-16. doi: 10.1007/BF00144036

Riley, M. A. & Lizotte-Waniewski, M. 2009. Population genomics and the bacterial species concept. In: Gogarten, M. B., Gogarten, J. P., & Olendzenski, L. C. (Eds.) *Horizontal Gene Transfer, Methods in Molecular Biology*. Humana Press, New York, 367-377. doi: 10.1007/978-1-60327-853-9_21

Rocap, G., Larimer, F. W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N. A., Arellano, A., Coleman, M., Hauser, L., Hess, W. R., Johnson, Z. I., Land, M., Lindell, D., Post, A. F., Regala, W., Shah, M., Shaw, S. L., Steglich, C., Sullivan, M. B., Ting, C. S., Tolonen, A., Webb, E. A., Zinser, E. R., & Chisholm, S. W. 2003. Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature*, 424, 1042-1047. doi: 10.1038/nature01947

Rodriguez-Caballero, E., Belnap, J., Büdel, B., Crutzen, P. J., Andreae, M. O., Pöschl, U., & Weber, B. 2018. Dryland photoautotrophic soil surface communities endangered by global change. *Nature Geoscience*, 11, 185-189. doi: 10.1038/s41561-018-0072-1

Rodríguez-Caballero, E., Cantón, Y., Chamizo, S., Afana, A., & Solé-Benet, A. 2012. Effects of biological soil crusts on surface roughness and implications for runoff and erosion. *Geomorphology*, 145, 81-89. doi: 10.1016/j.geomorph.2011.12.042

Rosen, M. J., Davison, M., Bhaya, D., & Fisher, D. S. 2015. Fine-scale diversity and extensive recombination in a quasisexual bacterial population occupying a broad niche. *Science*, 348, 1019-1023. doi: 10.1126/science.aaa4456

Roux, C., Fraisse, C., Romiguier, J., Anciaux, Y., Galtier, N., & Bierne, N. 2016. Shedding light on the grey zone of speciation along a continuum of genomic divergence. *PLoS Biology*, 14, e2000234. doi: 10.1371/journal.pbio.2000234

Sanger, F., Nicklen, S., & Coulson, A. R. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74, 5463-5467. doi: 10.1073/pnas.74.12.546

Schopf, J. W. & Packer, B. M. 1987. Early Archean (3.3-billion to 3.5-billion-year-old) microfossils from Warrawoona Group, Australia. *Science*, 237, 70-73. doi: 10.1126/science.11539686

Shapiro, B. J. & Polz, M. F. 2014. Ordering microbial diversity into ecologically and genetically cohesive units. *Trends in Microbiology*, 22, 235-247. doi: 10.1016/j.tim.2014.02.006

Shapiro, B. J. & Polz, M. F. 2015. Microbial speciation. *Cold Spring Harbor Perspectives in Biology*, 7, a018143. doi: 10.1101/cshperspect.a018143

Shapiro, B. J. 2018. What microbial population genomics has taught us about speciation. In: Polz, M. F. & Rajora, O. P. (Eds.) *Population Genomics: Microorganisms*. Springer, Cham, 31-47. doi: 10.1007/978-3-030-04756-6

Shapiro, B. J., Friedman, J., Cordero, O. X., Preheim, S. P., Timberlake, S. C., Szabó, G., Polz, M. F., & Alm, E. J. 2012. Population genomics of early events in the ecological differentiation of bacteria. *Science*, 336, 48–51. doi: 10.1126/science.1218198

Sheppard, S. K., Didelot, X., Meric, G., Torralbo, A., Jolley, K. A., Kelly, D. J., Bentley, S. D., Maiden, M. C. J., Parkhill, J., & Falush, D. 2013. Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in *Campylobacter. Proceedings of the National Academy of Sciences*, 110, 11923-11927. doi: 10.1073/pnas.1305559110

Siegesmund, M. A., Johansen, J. R., Karsten, U., & Friedl, T. 2008. *Coleofasciculus* gen. nov. (Cyanobacteria): morphological and molecular criteria for revision of the genus *Microcoleus* Gomont. *Journal of Phycology*, 44, 1572–1585. doi: 10.1111/j.1529-8817.2008.00604.x

Singh, R. K., Tiwari, S. P., Rai, A. K., & Mohapatra, T. M. 2011. Cyanobacteria: an emerging source for drug discovery. *The Journal of antibiotics*, 64, 401-412. doi: 10.1038/ja.2011.21

Smith, J. M., Smith, N. H., O'Rourke, M., & Spratt, B. G. 1993. How clonal are bacteria? *Proceedings of the National Academy of Sciences*, 90, 4384–4388. doi: 10.1073/pnas.90.10.4384

Sneath, P. H. & Sokal, R. R. 1973. *Numerical Taxonomy. The Principles and Practice of Numerical Classification*. 1st edition. Freeman, San Francisco.

Stackebrandt, E. & Goebel, B. M. 1994. Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *International Journal of Systematic and Evolutionary Microbiology*, 44, 846-849. doi: 10.1099/00207713-44-4-846

Stankowski, S. & Ravinet, M. 2021. Defining the speciation continuum. *Evolution*, 75, 1256-1273. doi: 10.1111/evo.14215

Stebbins, G. L. 1950. Variation and evolution in plants. New York: Columbia University Press.

Strackerbrandt E. & Ebers J. 2006. Taxonomic parameters revisited: Tarnished gold standards. *Microbiology Today*, 33, 152– 155.

Strunecký, O., Komárek, J., Johansen, J., Lukešová, A., & Elster, J. 2013. Molecular and morphological criteria for revision of the genus *Microcoleus* (Oscillatoriales, Cyanobacteria). *Journal of Phycology*, 49, 1167–1180. doi: 10.1111/jpy.12128

Stuart, R. K., Brahamsha, B., Busby, K., & Palenik, B. 2013. Genomic island genes in a coastal marine *Synechococcus* strain confer enhanced tolerance to copper and oxidative stress. *The ISME Journal*, 7, 1139-1149. doi: 10.1038/ismej.2012.175

Sun, J. & Windley, B. F. 2015. Onset of aridification by 34 Ma across the Eocene-Oligocene transition in Central Asia. *Geology*, 43, 1015-1018. doi: 10.1130/G37165.1

Polz, M. F., Alm, E. J., & Hanage, W. P. 2013. Horizontal gene transfer and the evolution of bacterial and archaeal population structure. *Trends in Genetics*. 29, 170–175. doi: 10.1016/j.tig.2012.12.006

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., et al. 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences*, 102, 13950-13955. doi: 10.1073/pnas.0506758102

Thomas, C. M. & Nielsen, K. M. 2005. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nature Reviews Microbiology*, 3, 711-721. doi: 10.1038/nrmicro1234

Tonkin-Hill, G., Lees, J. A., Bentley, S. D., Frost, S. D., & Corander, J. 2019. Fast hierarchical Bayesian analysis of population structure. *Nucleic Acids Research*, 47, 5539-5549. doi: 10.1093/nar/gkz361

Van Apeldoorn, M. E., Van Egmond, H. P., Speijers, G. J., & Bakker, G. J. 2007. Toxins of cyanobacteria. *Molecular Nutrition & Food Research*, 51, 7-60. doi: 10.1002/mnfr.200600185

Van Gremberghe, I., Leliaert, F., Mergeay, J., Vanormelingen, P., Van der Gucht, K., Debeer, A. E., Lacerot, G., Meester, L. D., & Vyverman, W. 2011. Lack of phylogeographic structure in the freshwater cyanobacterium *Microcystis aeruginosa* suggests global dispersal. *PLoS ONE*, 6, e19561. doi: 10.1371/journal.pone.0019561

Vos, M. & Didelot, X. 2009. A comparison of homologous recombination rates in bacteria and archaea. *The ISME Journal*, 3, 199-208. doi: 10.1038/ismej.2008.93

Vos, M., Wolf, A. B., Jennings, S. J., & Kowalchuk, G. A. 2013. Micro-scale determinants of bacterial diversity in soil. *FEMS Microbiology Reviews*, 37, 936–954. doi: 10.1111/1574-6976.12023

Vulić, M., Dionisio, F., Taddei, F., & Radman, M. 1997. Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. *Proceedings of the National Academy of Sciences*, 94, 9763-9767. doi: 10.1073/pnas.94.18.9763

Waterbury, J. & Stanier, R. 1977. Two unicellular cyanobacteria which reproduce by budding. *Archives of Microbiology*. 115, 249–257. doi: 10.1007/BF00446449

Watkins, R. F. & Gray, M. W. 2006. The frequency of eubacterium-to-eukaryote lateral gene transfers shows significant cross-taxa variation within amoebozoa. *Journal of Molecular Evolution*, 63, 801-814. doi: 10.1007/s00239-006-0031-0

Weber, B., Bowker, M., Zhang, Y., & Belnap, J. 2016. Natural recovery of biological soil crusts after disturbance. In: Weber, B., Büdel, B., & Belnap, J. (Eds.) *Biological soil crusts: An Organizing Principle in Drylands*. Springer, Cham, 479-498. doi: 10.1007/978-3-319-30214-0

Wheeler Q. D. & Meier R. 2000. Species Concepts and Phylogenetic Theory: a Debate. Columbia University Press, New York.

Wiedenbeck, J. & Cohan, F. M. 2011. Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. *FEMS Microbiology Reviews*, 35, 957–976. doi: 10.1111/j.1574-6976.2011.00292.x

Woese, C. R. 1987. Bacterial evolution. *Microbiological Reviews*, 51, 221-271.

Zachos, F. E. 2016. Species concepts in biology. Springer, Cham. doi: 10.1007/978-3-319-44966-1

Zhaxybayeva, O., Gogarten, J. P., Charlebois, R. L., Doolittle, W. F., & Papke, R. T. 2006. Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. *Genome Research*, 16, 1099-1108. doi: 10.1101/gr.5322306

Zimba, P. V., Shalygin, S., Huang, I. S., Momčilović, M., & Abdulla, H. 2020. A new boring toxin producer – *Perforafilum tunnelli* gen. & sp. nov. (Oscillatoriales, Cyanobacteria) isolated from Laguna Madre, Texas, USA. *Phycologia*, 60, 10–24. doi: 10.1080/00318884.2020.1808389

Zwirglmaier, K., Jardillier, L., Ostrowski, M., Mazard, S., Garczarek, L., Vaulot, D., Not, F., Massana, R., Ulloa, O., & Scanlan, D. J. 2008. Global phylogeography of marine *Synechococcus* and *Prochlorococcus* reveals a distinct partitioning of lineages among oceanic biomes. *Environmental Microbiology*, 10, 147-161. doi: 10.1111/j.1462-2920.2007.01440.x

Manuscript contributions

Paper I*: Geography and climate drive the distribution and diversification of the cosmopolitan cyanobacterium Microcoleus (Oscillatoriales, Cyanobacteria)*. AS designed the research idea with the co-authors, isolated and maintained the strains, performed experimental and statistical analyses, and led the writing process with contributions from all co-authors.
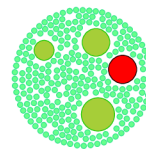
Paper II: *High genomic differentiation and limited gene flow indicate recent cryptic speciation within the genus Laspinema (cyanobacteria)*. AS designed the research idea with the co-authors, wrote all code, processed and analyzed the data, and led the writing process with contributions from all co-authors.

Paper III: *The global speciation continuum of cyanobacterium Microcoleus*. AS designed the research idea with the co-authors, wrote code for statistical and population genetic analyses, produced all plots and figures, and led the writing process with contributions from co-authors.

Papers

I. **Stanojković, A.**, Skoupý, S., Hašler, P., Poulíčková, A., & Dvořák, P. **2022**. Geography and climate drive the distribution and diversification of the cosmopolitan cyanobacterium *Microcoleus* (Oscillatoriales, Cyanobacteria). *European Journal of Phycology*, 57, 396-405. https://doi.org/10.1080/09670262.2021.2007420. (IF_{2022} − 2.667)

II. **Stanojković, A.**, Skoupý, S., Škaloud, P., & Dvořák, P. **2022**. High genomic differentiation and limited gene flow indicate recent cryptic speciation within the genus *Laspinema* (cyanobacteria). *Frontiers in Microbiology*, 13, 977454. https://doi.org/10.3389/fmicb.2022.977454. (IF_{2022} − 6.064)

III. **Stanojković, A.**, Skoupý, S., & Dvořák, P. **2023**. The global speciation continuum of cyanobacterium *Microcoleus*. *Manuscript*.

Paper I

British Phycological Society
Understanding and using algae

Taylor & Francis
Taylor & Francis Group

Check for updates

Geography and climate drive the distribution and diversification of the cosmopolitan cyanobacterium *Microcoleus* (Oscillatoriales, Cyanobacteria)

Aleksandar Stanojković ⓘ, Svatopluk Skoupý ⓘ, Petr Hašler ⓘ, Aloisie Poulíčková ⓘ and Petr Dvořák ⓘ

Palacký University Olomouc, Faculty of Science, Department of Botany, Šlechtitelů 27, 78371 Olomouc, Czech Republic

ABSTRACT
Despite the extensive diversity of bacteria and their importance to the fundamental functioning of terrestrial ecosystems, their distribution patterns are still not fully known. To fill the gap and further understand the biogeographic patterns in bacteria, we investigated the phylogeographic structure and the underlying drivers of diversification among populations of the cyanobacterium *Microcoleus* spp. The phylogenetic history was reconstructed using 16S rRNA genes and the 16S–23S internal transcribed spacer (ITS) of 495 *Microcoleus* spp. isolates. Ancestral area and state reconstruction was employed to investigate the distributional and ecological patterns within *Microcoleus*. Both isolation by distance and isolation by environment were tested with distance matrices analysis. The phylogenetic signal tests were conducted in order to assess the influence of the climatic preferences on the diversification of *Microcoleus* isolates. The distribution and phylogenetic diversification of *Microcoleus* are driven by both isolation by distance and environment, leading to at least 13 distinct lineages that could represent novel cyanobacterial species. *Microcoleus* spp. exhibited a distinct phylogeographic structure within the respective lineages. The ancestral area and state reconstruction revealed that *Microcoleus* most likely arose in Europe in terrestrial habitats. The phylogenetic signal showed that the phylogeny significantly affects the climatic preferences of *Microcoleus* strains. Geographic distance and contemporary climatic conditions play significant roles in shaping the distribution and diversification of *Microcoleus*. The observed patterns of distribution may shift in the future due to the impact of climate change.

HIGHLIGHTS
- *Microcoleus* exhibited distinct phylogeographic structure within the respective lineages.
- Geographic and environmental heterogeneity affect *Microcoleus* distribution and diversification.
- Genetically distinct lineages coexist at the same site.

Introduction

The number of microbial species worldwide is immense, and we have only a limited understanding of their dispersal and distribution patterns. Articulating microbial biogeographic patterns is hampered by the inconsistent use of species concepts, understudied microhabitats, underestimated diversity and unknown rates of dispersal (Fierer, 2008), and this is further complicated by different methodological approaches (e.g. sequencing strategy as reviewed in Hanson *et al.*, 2012). With the development of next-generation sequencing, it has become possible to better detect biogeographic patterns and mechanisms of speciation (selection, dispersal, genetic drift and mutations; Hanson *et al.*, 2012). Different microbial species exhibit distinct distribution and evolutionary patterns; therefore, describing the biogeographic patterns in non-model organisms is an essential step to better understanding microbial biogeography in a broader context.

Microorganisms differ from most macroorganisms because they have larger dispersal ability, population sizes, and may form resistant propagules within their life cycles (Foissner, 2006; Fontaneto & Brodie, 2011). Initially, microbiologists assumed that spatial differentiation was not evolutionarily significant – as noted by Baas-Becking's (1934) tenet 'everything is everywhere, but the environment selects'. Recent studies (e.g. Finlay, 2002; Fenchel, 2003) support this assertion, but employ only morphological data as evidence. However, molecular studies have shown that the dispersal of free-living microorganisms may actually be limited (e.g. Miller *et al.*, 2007; Bates *et al.*, 2013; Aguilar *et al.*, 2014; Ribeiro *et al.*, 2020), but this is variable because dispersal barriers may only be temporary (Bahl *et al.*, 2011; Dvořák *et al.*, 2012). Reno *et al.* (2009) demonstrated the existence of dispersal barriers on a continental scale using genome

data in archaeon *Sulfolobus*. Thus, morphological data seem to have limited resolution. This may be exacerbated by cryptic speciation – low morphological diversity which conceals much larger genetic diversity (Leliaert *et al.*, 2014).

In cyanobacteria, studies focused on phylogeographic structure seem to provide ambiguous results. For instance, *Microcystis* cyanobacteria did not exhibit any phylogeographic pattern according to Van Gremberghe *et al.* (2011) and Ribeiro *et al.* (2020), but *Raphidiopsis raciborskii* populations have distinct structure (Ribeiro *et al.*, 2020) as do those of *Microcoleus vaginatus* (Dvořák *et al.*, 2012). While these studies indicate the complex nature of the diversification and distribution of cyanobacteria, they had only limited population-level resolution since each locality was mainly represented by a single strain and sequence.

Phylogenetic signal is a measure of the significance of the relationship between species traits and genetic relatedness (Blomberg *et al.*, 2003) and has been used to highlight the importance of ecology on the distribution of microorganisms (e.g. Aguilar *et al.*, 2014; González-Rocha *et al.*, 2017). Even without a clear consensus on measuring the phylogenetic signal, this concept needs consideration when detecting biogeographic species patterns (Losos, 2008; Crisp & Cook, 2012). In cyanobacteria, the phylogenetic signal has previously been determined with morphological, physiological (e.g. Uyeda *et al.*, 2016) and some environmental factors including nutrients, pH and sea surface temperature (e.g. Larkin *et al.*, 2016).

Microcoleus vaginatus is a cosmopolitan, filamentous bundle-forming cyanobacterium that inhabits benthic and subaerial habitats (Garcia-Pichel *et al.*, 1996). It is an essential part of biological soil crusts in both hot and cold deserts, where it plays an important role in the biogeochemical cycle and stabilizing soil particles with its exopolysaccharides (Garcia-Pichel & Wojciechowski, 2009). *Microcoleus* is polyphyletic (e.g. Siegesmund *et al.*, 2008; Hašler *et al.*,

2012), and Strunecký *et al.* (2013) defined a new type material for *M. vaginatus*. They note that *M. vaginatus* is composed of at least six species, but they were only vaguely defined, and thus revision of the genus *Microcoleus* is needed. In this paper, we seek to shed some new light on the biogeography and the evolution of free-living organisms using phylogenetic reconstruction of cyanobacterium *Microcoleus*. Here, we recognize the existence of many lineages within *M. vaginatus* and we will refer to them as '*Microcoleus* spp.'.We investigated patterns of diversity from a single *Microcoleus* spp. population to continent-scale distribution, on both an inter- and intraspecies level. Moreover, we identify the relationship between climate, geography and genetic diversity.

Materials and methods

Data collection and cultivation

Collections of *Microcoleus* spp. were made during 2019. Once collected, samples were placed directly into plastic bags. Strains were obtained from 75 environmental samples originating from all continents except for South America. Sampling locations had diverse climates and habitats: soil (dry convex surface), puddles (ephemeral concave water bodies 5–10 cm in depth), concrete, moss vegetation and rocks (Fig. 1; Supplementary table S1). A small portion of each sample was placed in 10 ml capped tubes with liquid Zehnder medium (Staub, 1961). Part of the grown biomass was then transferred to Petri dishes on agar solidified (1.5%) Zehnder medium. Unialgal cultures were isolated following techniques described by Hašler *et al.* (2012). From each environmental sample, one to 11 cultures from one filament were obtained, and altogether 495 clonal cultures of *Microcoleus* spp. were established. The entire culture collection is currently maintained at the Department of Botany, Palacký University in Olomouc, Czech
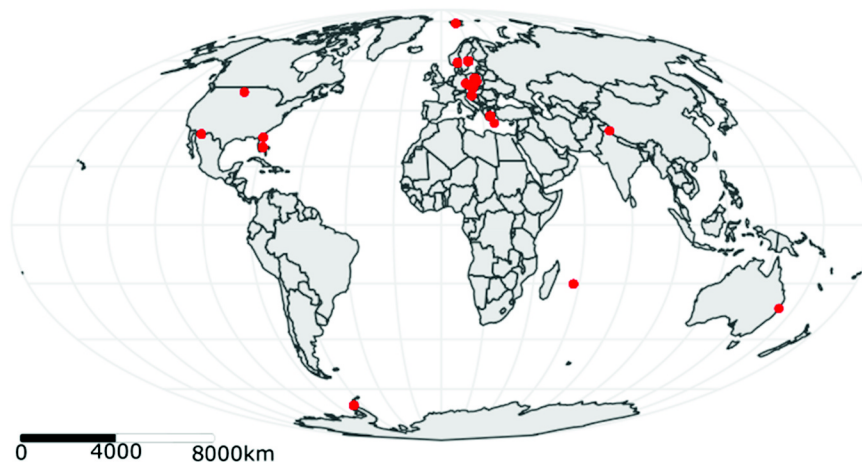

Fig. 1. Map showing the position of *Microcoleus* spp. sampling sites (Mollweide projection).

Republic. Morphology of the strains was inspected under 400× magnification using a Zeiss Primo Star (Oberkochen, Germany) light microscope and they were identified following the taxonomic system *sensu* Komárek & Anagnostidis (2005). We evaluated morphological characters of filaments: division, shape, size and the presence of calyptra. All isolates were grown in 10 ml capped tubes with liquid Zehnder medium and maintained at 22 ± 1°C with an average photon flux density of 20 μmol photons $m^{-2}s^{-1}$ under 12h light/12h dark light regime.

DNA extraction, PCR amplification and sequencing

DNA was extracted from ~50 mg of fresh biomass using an UltraClean Microbial DNA Isolation Kit (MOBIO, Carlsbad, California, USA) following the manufacturer's recommendations. Isolated DNA was eluted in 30 μl of TE buffer and kept at −20°C. PCR amplification of partial 16S rRNA and the whole 16S–23S ITS region was performed with forward P2 (5′-GGGGAATTTTCCGCA ATGGG-3′) and reverse P1 (5′-CTCTGTGTGCCTAG GTATCC-3′) primers, according to Boyer *et al.* (2002). Each 40 μl PCR reaction had 17 μl of sterile water, 1 μl of each primer (0.01 mM concentration), 1 μl of template DNA (50 ng $μl^{-1}$), and 20 μl EmeraldAmp Max HS PCR Master Mix (Takara Bio Europe S.A.S., France). PCR amplification was carried out under conditions previously described in Dvořák *et al.* (2012). The concentration and quality of PCR products were inspected on 1.5% agarose gels, stained with 1.5 ul ethidium bromide. Purification of all PCR products was done using GenElute™ PCRClean-Up Kit (Sigma-Aldrich, Co., Saint Louis, Missouri, USA) and E.Z.N.A.® Cycle Pure Kit (Omega Bio-tek, Inc., Norcross, Georgia, USA).

The PCR products were sequenced using the aforementioned P1 and P2 primers with additional P5 (5′-TGTACACACCGCCCGTG-3′) and P8 (5′-AAGGAGGTGATCCAGCCACA-3′), described in Boyer *et al.* (2002). The purified PCR products were Sanger sequenced at Macrogen Europe, Inc. (Amsterdam, the Netherlands, http://dna.macrogen-europe.com). 16S rRNA and 16S–23S ITS sequences were trimmed and assembled in Sequencher 5.0 (Gene Codes Corporation, Ann Arbor, Michigan, USA) and then identified using the BLAST nucleotide search (http://blast.ncbi.nlm.nih.gov/Blast.cgi) to confirm they belonged to *Microcoleus* spp.

Phylogenetic analysis

In addition to 495 *Microcoleus* spp., four *Kamptonema animale* clonal cultures were obtained from the sample N3 (Norway, Europe) and used as an outgroup in the phylogenetic tree. All 16S rRNA and 16S–23S ITS sequences have been deposited in GenBank (www.ncbi.nlm.nih.gov/genbank; see

accession numbers in Supplementary table S1). Both multiple sequence alignment of 16S rRNA and 16S–23S ITS genes and alignment processing were performed by the Muscle algorithm (Edgar, 2004) in AliView (Larsson, 2014). The multiple sequence alignment was trimmed using trimAl 1.4.22 (Capella-Gutiérrez *et al.*, 2009) with the option –automated1. The phylogenetic tree was inferred the maximum-likelihood tree in IQ-TREE 1.6.1 (Nguyen *et al.*, 2015). The most appropriate model –Trn+I+G was selected based on Bayesian Information Criterion (Schwarz, 1978). The tree topology was tested using ultrafast bootstrapping, implemented within the same software, with 2000 replications (Hoang *et al.*, 2018).

Ancestral area and state reconstruction

Ancestral state reconstruction (ASR) of geographic origin and habitats of *Microcoleus* spp. was performed using Bayesian binary MCMC (BBM) analysis that is implemented within RASP v4.2 (Yu *et al.*, 2015, 2020). The geographic origin of *Microcoleus* spp. strains was divided into seven regions: Europe, Arctic, Antarctic, Australia, North America, Asia and Africa. The ancestral area reconstruction and the estimation of the spatial patterns of geographic diversification within *Microcoleus* spp. were inferred using the Bayesian binary method (BBM) which was selected due to its capability of presenting single distribution areas. The BBM was run with the fixed state frequency model (Jukes–Cantor) with equal among-site rate variation for 50 000 generations and 10 chains each. The maximum number of ancestral areas was set to seven for geographic origin and five for ancestral habitats.

Statistical analysis

To assess the effect of the climate on the emergence of *Microcoleus* spp. strains, we measured the phylogenetic signal. Firstly, 19 bioclimatic variables for locations of sampled strains were downloaded from the WorldClim v2.1 database (Fick & Hijmans, 2017) at 2.5 arc minutes resolution and then extracted in R (version 4.0.0, R Development Core Team) using the package 'raster' v3.3-13 (Hijmans, 2020) (Supplementary table S2). Then, we measured the phylogenetic signal of bioclimatic variables using three signal indices. Pagel's λ (Pagel, 1999) was estimated with the function phylosig (package 'phytools' v0.7-47; Revell, 2012), Abouheif's C_{mean} (Abouheif, 1999) was calculated with the function abouheif.moran and the method oriAbouheif (package 'adephylo' v1.1-11; Jombart *et al.*, 2010) and Moran's I (Moran, 1950) was calculated with the function phyloSignal (package 'phylosignal' v1.3; Keck *et al.*, 2016).

The Pagel's λ value of zero corresponds to traits, which do not have a phylogenetic signal (independent

evolution of climatic niche from the phylogeny), and one corresponds to traits having a strong phylogenetic signal (dependent evolution of climatic niche from the phylogeny) (Pagel, 1999). To test the significance of the estimated lambda value when there is no phylogenetic signal ($\lambda = 0$, null hypothesis), the likelihood-ratio test ($-2[\log L0 - \log L1]$) was employed. This test estimates the difference between the two models, where $\log L0$ represents the likelihood of $\lambda = 0$ and $\log L1$ represents the likelihood of $\lambda = 1$. Its statistical significance was analysed using the χ^2 test.

Abouheif's C_{mean} and Moran's I, with 999 simulations, were also used to detect the phylogenetic signal in the climatic niche space. Both indices' values are ranging from -1 to 1, where the value of -1 represents a lack of phylogenetic signal and the value of 1 represents the presence of phylogenetic signal in traits. If there was statistically significant high phylogenetic autocorrelation (Moran's I and Abouheif's C_{mean} greater than zero), then the null hypothesis of no phylogenetic autocorrelation in the dataset could be rejected.

The Mantel test (Mantel, 1967) was used to investigate the correlation between climate, geographic and genetic distance. It was performed in R with the function mantel (package 'vegan' v2.5-6; Oksanen et al., 2016). First, the genetic distance matrix was obtained from Mega7 (Kumar et al., 2016) using Tamura–Nei substitution model (Tamura & Nei, 1993). Second, the spatial matrix was built using the geographic coordinates of localities where we sampled our strains and inferred using the function distm and the function distGeo (package 'geosphere' v1.5-10; Hijmans et al., 2017) in R. Lastly, using the function findCorrelation (package 'caret' v6.0-86; Kuhn, 2008), highly correlated bioclimatic variables were eliminated ($r < 0.9$). We kept the following bioclimatic variables: precipitation seasonality, precipitation of driest quarter, mean temperature of warmest quarter, precipitation of coldest quarter, mean diurnal range, isothermality, temperature seasonality, maximum temperature of warmest month, temperature annual range, mean temperature of wettest quarter and mean temperature of the driest quarter. Prior to inference of the environmental matrix using the function dist and the euclidean method in R, all kept bioclimatic variables were centred and scaled. Significances of correlations between genetic and two other matrices were calculated on the basis of 9999 randomized permutations with the Pearson correlation coefficient.

Results

The diversity within Microcoleus spp.

Phylogenetic analysis of 16S rRNA and 16S–23S ITS sequences revealed 13 monophyletic lineages of *Microcoleus* spp. (Figs 2, 3, Supplementary fig. S1).

Some lineages were more diverse than others in our collection. Lineage 1 was the most diverse one containing 149 strains, followed by lineages 6, 10 and 11, which all had $\geq$ 50 strains. Lineages 5, 7, 9, 12 and 13 were least diverse containing <10 strains and in the rest of lineages, the number of strains ranged from 10 to 50 (details in Supplementary table S3).

We detected different lineages of *Microcoleus* spp. coexisting within a sampling site (Supplementary table S4). In 45 samples we found just one lineage, in 21 samples we found two lineages and in eight samples we discovered three. In one sample from Europe (AT16), we found four lineages.

Phylogeography and evolution of habitat preference in Microcoleus spp.

The phylogeny revealed that isolates mostly followed a clustering pattern according to their geographic origin within the respective lineages (Fig. 2). Within lineage 1, we observed a highly supported subclade (bootstrap values $\geq$ 99) composed of African and European strains. The exceptions from the clustering pattern were two North American strains that grouped with European isolates. Lineage 2 included a well-supported North American and European subclades. Lineage 3 included a highly supported and diverse subclade of Arctic strains, a European subclade and a subclade composed of strains from Europe and Asia. Within lineage 4, among two European subclades was a highly supported one encompassing North American strains. Six lineages (5, 7, 9, 10, 12, 13) were composed of only European strains. Lineage 6 was comprised of several European, North American, and Asian subclades. One African isolate had the longest branch in the tree and it formed a highly supported subclade with European strains. Moreover, two isolates from Asia and North America clustered with European strains, revealing incongruence to the clustering pattern according to geographic origin. Lineage 8 had well-defined subclades with significant bootstrap supports that included North American and European isolates. Yet, three North American strains appeared to be more closely related to the group of European ones. Within lineage 11 there were very well-supported subclades with Antarctic and European strains that were closely related to the Australian strains. Four strains forming three singleton nodes were not included in any of thirteen lineages, but they were counted as an additional lineage isolated from the respective samples.

Isolation by distance was tested using a correlation analysis of genetic and geographical distance with the Mantel test. Geographical and genetic distances between the strains were significantly correlated ($r = 0.2971$, $p = 0.0001$) (Fig. 4). This provides additional
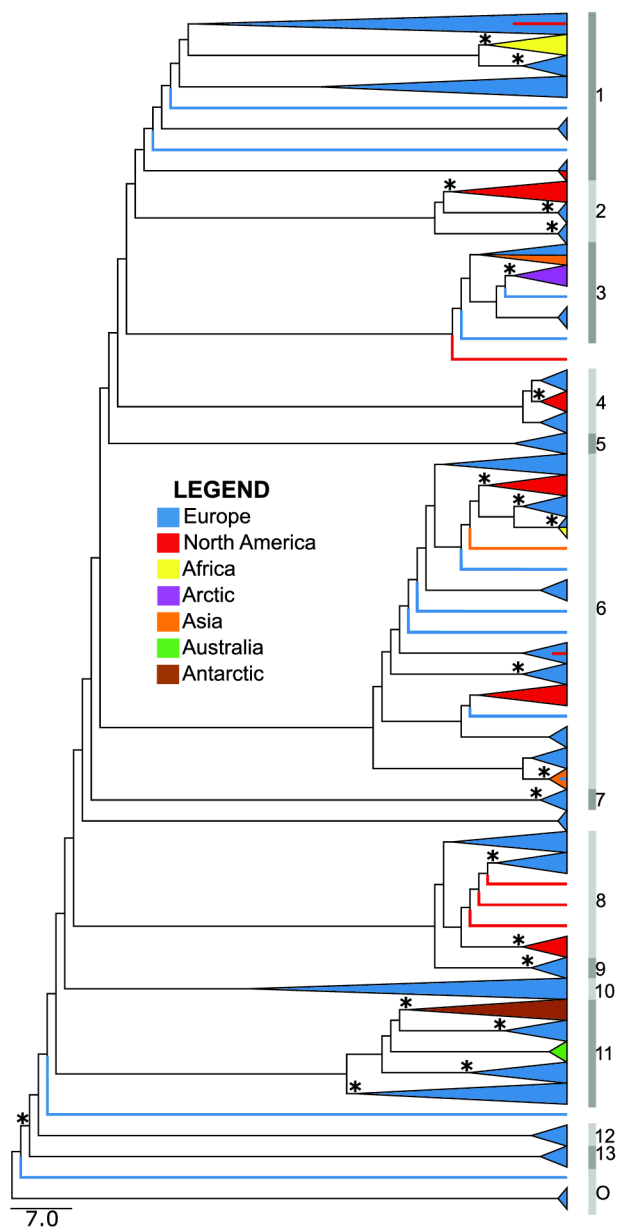
Fig. 2. Collapsed maximum likelihood phylogenetic tree based on 16S rRNA and 16S-23S ITS sequences of *Microcoleus* spp. with coloured clades indicating geographical distribution areas of strains. Two-coloured clades include strains with different geographic regions of origin. Coloured line within a subclade represents a single strain. Designation of clades to lineages (1–13) is shown. The outgroup is indicated as O. See Supplementary fig. S1 for a fully detailed phylogenetic tree. Asterisks represent the ultrafast bootstrap values ≥ 99. The scale bar indicates substitutions per site.

were strains from soil and puddles that clustered together. Moreover, a strain isolated from concrete grouped with strains isolated from puddles and soil. Lineage 2 included two subclades with soil isolates and one with isolates from puddle and concrete. All strains from lineages 3, 5, 7 and 9 originated from soil, whereas lineage 13 included strains only from puddles. A single puddle isolate was found among strains isolated from soil within lineage 4. Lineage 6 contained several subclades from soils and a subclade isolated from a puddle. Moreover, four puddle strains
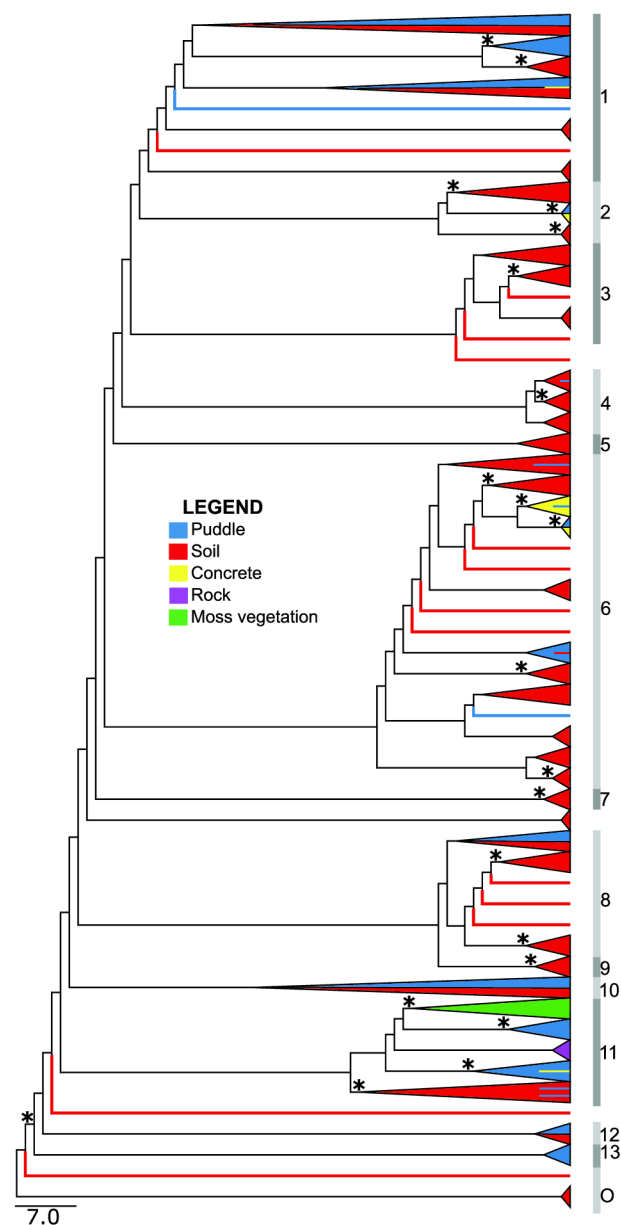

Fig. 3. Collapsed maximum likelihood phylogenetic tree based on 16S rRNA and 16S-23S ITS sequences of *Microcoleus* spp. with coloured clades indicating habitats of strains. Two-coloured clades include strains with different habitats of origins. Designation of clades to lineages (1–13) is shown. Coloured line within a subclade represents a single strain. The outgroup is indicated as O. See Supplementary fig. S1 for a fully detailed phylogenetic tree. Asterisks represent the ultrafast bootstrap values ≥ 99. The scale bar indicates substitutions per site.

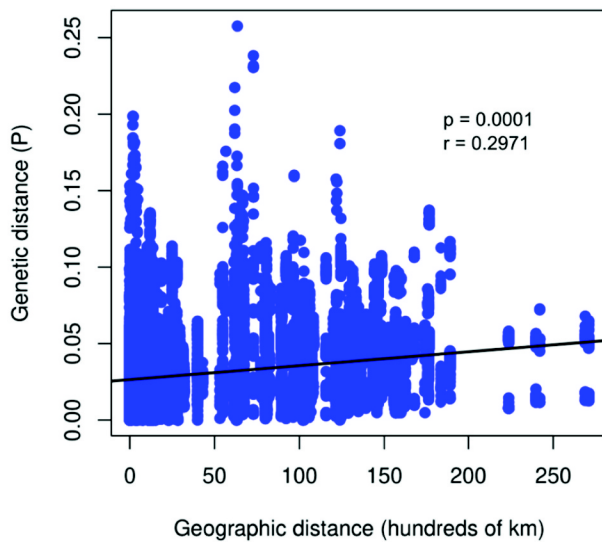evidence that geographic distance plays an important role in the evolution of *Microcoleus* spp. lineages.

An ancestral reconstruction of *Microcoleus* habitats indicates that the ancestor of the lineages most likely originated in soils (Supplementary fig. S3). The clustering of *Microcoleus* spp. strains mostly followed their habitat preference within the respective lineages as well (Fig. 3). Within lineage 1 were several well-defined subclades including soil and puddle strains. However, few exceptions from the observed pattern

Fig. 4. Scatter plot illustrating the relationship between genetic distance (P) and geographic distances (hundreds of km) between all *Microcoleus* spp. isolates. Mantel test result indicates a significant correlation between distances (r = 0.2971, p = 0.0001).

clustered with soil and concrete habitats. Lineage 8, 10 and 12 contained strains from soils and puddles. Lineage 11 encompassed strains isolated from all habitats. Isolates from soil, puddles and moss vegetation formed highly supported and diversified subclades, whereas a strain isolated from concrete and two from puddles did not appear to have followed the pattern of clustering according to habitat preference.

Phylogenetic signal

We tested the presence of phylogenetic signal in climatic niche space of *Microcoleus* spp. to investigate whether climatic preferences (temperature and precipitation) are well predicted by the phylogenetic relatedness. We observed a strong phylogenetic signal in bioclimatic variables after a removal of auto-correlated variables (Supplementary table S5). According to Pagel's λ, Moran's I, and Abouheif's C_{mean} measurements, 11 bioclimatic variables exhibited a statistically significant (p ≤ 0.001) phylogenetic signal (Supplementary table S5).

A Mantel test was employed to examine the effect of the climate on the *Microcoleus* spp. diversity. The distance between contemporary climatic conditions calculated with the euclidean method and genetic distance showed a significant correlation (r = 0.2941, p = 0.0001). This relationship stresses the importance of the climate on the divergence of *Microcoleus* spp. lineages.

Discussion

Detecting microbial biogeographic patterns and discovering the mechanisms driving them are crucial for understanding the microbial diversity, distribution and evolution (Hanson *et al.*, 2012). In this study, we examined the congruence between phylogenetic relationships, geography and the environment (climate and habitat) among *Microcoleus* spp. populations using 16S rRNA and 16S–23S ITS genes. We found that there is considerable genetic diversity within *Microcoleus* spp. clusters at both local and global scales. In addition, we note a significant influence of geographic and environmental heterogeneity in the distribution and diversification of *Microcoleus* spp. lineages.

Microcoleus is one of the dominant biological soil crust cyanobacteria (e.g. Boyer *et al.*, 2002; Gundlapally & Garcia-Pichel, 2006; Strunecký *et al.*, 2013). Of the thirteen lineages within *Microcoleus* spp. that we observed, we suspect that some of them might represent novel cyanobacterial species (Supplementary fig. S1). We noted high diversity among identified lineages. For instance, lineage 1 had 149 strains, which suggests it is more abundant than other lineages in the investigated environments or it is easier to isolate monoclonal culture (Supplementary table S3).

In general, the evolution of different prokaryotic lineages from an initial population is possible due to the high adaptability of these organisms, promiscuous gene exchange, reduced dispersal and restricted gene flow (Whitaker *et al.*, 2003; Cadillo-Quiroz *et al.*, 2012; Dvořák *et al.*, 2015). In most studies, a single sampling site is usually represented by a single isolate (hence, one sequence). Thus, it would only be possible to detect a single lineage of *Microcoleus* spp. Population-level sampling is necessary to discover multiple lineages at a site. Employing whole-genome sequencing, for example, Chase *et al.* (2017) and Hunt *et al.* (2008) observed such a pattern in soil *Curtobacterium* and marine bacterium *Vibrio*, respectively. Studies investigating that pattern are still rare in cyanobacteria. Nevertheless, Pietrasiak *et al.* (2014) isolated two different species of cyanobacterium *Symplocastrum* from the same soil crust locality using 16S rRNA and 16S–23S ITS. Our study is consistent with the aforementioned findings and here we note the coexistence of several *Microcoleus* spp. lineages at the same site (Supplementary table S4). Close to one third of samples contained only one *Microcoleus* spp. lineage, while the rest had two, three or four. Additionally, our data illustrate the co-occurrence of distantly related lineages from different populations within the same locality, whilst more related ones were far apart. As evidenced in *Prochlorococcus* by Kashtan *et al.* (2014), the coexistence of genetically distinct lineages could represent one of the characteristic traits of free-living prokaryotes.

Phylogenetic inference and ancestral area reconstruction analysis revealed the phylogeographic structure within the respective lineages of *Microcoleus* spp. (Fig. 2). An ancestral reconstruction of *Microcoleus* spp. geographic origin revealed that ancestors of our isolates most likely originated in Europe (Supplementary fig. S2). However, this finding could be an artefact of the over-representation of European strains in our collection. Although most lineages were diversified clusters of isolates within their respective lineages, a few strains did not follow this pattern (Fig. 2). Such a relationship between isolates suggests a certain probability of potential gene flow among strains from Europe and North America on one side, and Europe, Asia and Africa on the other. The Mantel test results showed that geographic isolation can affect speciation in *Microcoleus* spp. (Fig. 4); as physical distance becomes smaller, strains of *Microcoleus* tend to be more genetically similar. Therefore, our results do not agree with Baas-Becking's (1934) tenet that everything is everywhere, nor with studies of Fenchel (2003) and Finlay (2002) due to the influence of the isolation by distance on the diversification of *Microcoleus* spp. A similar pattern was documented in the archaeon *Sulfolobus islandicus* (Reno *et al.*, 2009). Thus, allopatry, i.e. speciation by geographically isolated populations, represents an important contributing factor in the speciation of *Microcoleus*, yet it may be of a temporary duration (Bahl *et al.*, 2011; Dvořák *et al.*, 2012). Some of the possible dispersal pathways of filamentous cyanobacteria between continents include the atmosphere (e.g. Sharma & Singh, 2010), animals (e.g. Moore, 1985), and human factors (reviewed in Curren & Leong, 2020). Nevertheless, we are unable to reconstruct the intensity of the gene flow between lineages in this study.

The climate has a significant effect on the diversification, distribution and composition of cyanobacterial assemblages in biological soil crusts (e.g. Büdel *et al.*, 2009; Bahl *et al.*, 2011; Garcia-Pichel *et al.*, 2013). Amid environmental variables that could explain the genetic diversity of cyanobacteria are temperature, precipitation, composition of soil and crust types. Ribeiro *et al.* (2020) demonstrated that contemporary and past climatic conditions significantly affect the genetic diversity of global *Microcystis aeruginosa* and *Raphidiopsis raciborskii* populations. Moreover, *Microcoleus*-dominated cyanobacterial communities have been shown to be affected by temperature and precipitation (Muñoz-Martin *et al.*, 2019). Our study is consistent with these findings as bioclimate variables significantly correlated with the *Microcoleus* spp. global genetic diversity. Thus, it showed that the global population structure of *Microcoleus* spp. is also affected by isolation by the environment.

Whilst phylogenetic signal is commonly investigated in macroorganisms (see Lososet *et al.,* 2008) and algae (e.g. Škaloud & Rindi, 2013; Narwani *et al.,* 2015), it remains understudied in prokaryotes. A strong phylogenetic signal was recently detected in some morphological, ecological and physiological traits in cyanobacteria (Uyeda *et al.*, 2016). Three independent phylogenetic signal measurements in this study supported the existence of a high phylogenetic signal in some ecological traits (Supplementary table S5). This trend suggests non-independent evolution between climate and phylogeny (i.e. closely related *Microcoleus* spp. strains tended to be more similar in their temperature and precipitation preferences than the more distantly related ones).

Additionally, we show that the strains also cluster by habitat within the respective lineages (Fig. 3). *Microcoleus* was initially described from soils (Gomont, 1892). Recently, strains of *Microcoleus* have been found in other habitats as well – freshwater epipelon, puddles, rocks and concrete (Dvořák *et al.*, 2012; Hašler *et al.*, 2012; this study). The ASR of habitats revealed that the majority of our isolates originated from the soil (Supplementary fig. S3). However, its oversampling in this study may affect the observed pattern.

Elevated temperature, shifts in precipitation frequencies and anthropogenic influences are altering the structure of cyanobacterial communities, their abundances, growth rates and distribution (Flombaum *et al.*, 2013; Steven *et al.*, 2015; Fernandes *et al.*, 2018). Our results suggest that diversification of *Microcoleus* spp. lineages might be driven by temperature and precipitation. Thus, climate change may alter the observed patterns of *Microcoleus* spp. diversity.

Our study showed biogeographic patterns of *Microcoleus* spp. populations and the relationship between geography, environment and genetic diversity. Isolation by distance and isolation by environment affected the distribution and the diversification of *Microcoleus* spp. strains. Consequently, at least thirteen distinct lineages were found that could be new cyanobacterial species with very similar morphologies. As climate is a driver of the evolution in *Microcoleus* spp., climate change may affect its distribution and diversity.

Acknowledgements

Disclosure statement

Funding

Supplementary information

The following supplementary material is accessible via the Supplementary Content tab on the article's online page at https://doi.org/10.1080/09670262.2021.2007420

Supplementary table S1: List of the *Microcoleus* spp. strains, GPS locations, habitats, accession numbers of 16S and 16S-23S ITS sequences.

Supplementary table S2: Extracted values of bioclimatic variables (downloaded from WorldClimv2.1) in R with the package 'raster' for *Microcoleus* spp. strains in this study.

Supplementary table S3: A number of strains in each of *Microcoleus* spp. lineages identified in this study

Supplementary table S4: A number of lineages found in environmental samples of *Microcoleus* spp. and the list of countries or regions where samples were collected.

Supplementary table S5: Summary of the phylogenetic signal measurement analysis for 19 bioclimatic variables (with Pagel's λ, Abouheif's C_{mean}, and Moran's I) used to detect the presence of the phylogenetic signal in the climatic nice space of *Microcoleus* spp. Highlighted bioclimatic variables were not auto-correlated.

Supplementary fig. S1. Phylogenetic tree inferred from the maximum-likelihood (ML) analysis based on 16S rRNA and 16S-23S ITS sequences. Asterisks at the nodes indicate maximum likelihood bootstrap support (≥ 0.99). Based on the monophyletic criterion, strains were associated with thirteen lineages (1-13). The outgroup is indicated as O. Strains forming singleton nodes are indicated as #. Scale bar indicates substitutions per site.

Supplementary fig. S2. Summary of the ancestral area reconstruction analysis based on Bayesian Binary MCMC (BBM) model in *Microcoleus* spp. The ancestral areas with the highest likelihood are represented within node pies. Colour code corresponds to the following regions: (A) Europe, (B) North America, (C) Africa, (D) Asia, (E) Arctic, (F) Antarctic, (G) Australia. The black asterisk represents other ancestral areas. Additional colours within the colour code correspond to the multiple region areas (AD).

Supplementary fig. S3. Summary of the ancestral character state reconstruction analysis based on Bayesian Binary MCMC (BBM) model in *Microcoleus* spp. The ancestral habitats with the highest likelihood are represented within node pies. Colour code corresponds to the following habitats: (A) puddle, (B) soil, (C) concrete, (D) moss vegetation, (E) rocks. The black asterisk represents other ancestral habitats. Additional colours within the colour code correspond to the multiple character states (AB, AE, BC).

Author contributions

A. Stanojković: culture isolation and maintenance, sampling, experimental analysis, statistical analysis, drafting and editing manuscript; S. Skoupý: isolation and maintenance of the cultures, experimental analysis, editing manuscript; P. Hašler: isolation and maintenance of the cultures, editing manuscript; A. Poulíčková: editing manuscript and review process; P. Dvořák: conceptualization, sampling, editing manuscript and review process.

Data availability statement

ORCID

Aleksandar Stanojković   http://orcid.org/0000-0003-3252-2068
Svatopluk Skoupý   http://orcid.org/0000-0001-8906-1544
Petr Hašler   http://orcid.org/0000-0003-4658-1114
Aloisie Poulíčková   http://orcid.org/0000-0001-9733-942X
Petr Dvořák   http://orcid.org/0000-0003-0101-9566

References

Abouheif, E. (1999). A method for testing the assumption of phylogenetic independence in comparative data. *Evolutionary Ecology Research*, **1**: 895–909.

Aguilar, M., Fiore-Donno, A.M., Lado, C. & Cavalier-Smith, T. (2014). Using environmental niche models to test the 'everything is everywhere' hypothesis for *Badhamia*. *ISME Journal*, **8**: 737–745.

Baas-Becking, L.G.M. (1934). *Geobiologie of Inleiding tot de Milieukunde*. W. P. Van Stockum & Zoon, The Hague.

Bahl, J., Lau, M.C.Y., Smith, G.J.D., Vijaykrishna, D., Cary, S.C., Lacap, D.C., Lee, C.K., Papke, R.T., Warren-Rhodes, K.A., Wong, F.K.Y., McKay, C.P. & Pointing, S.B. (2011). Ancient origins determine global biogeography of hot and cold desert cyanobacteria. *Nature Communications*, **2**: 1–6.

Bates, S.T., Clemente, J.C., Flores, G.E., Walters, W.A., Parfrey, L.W., Knight, R. & Fierer, N. (2013). Global biogeography of highly diverse protistan communities in soil. *ISME Journal*, **7**: 652–659.

Blomberg, S.P., Garland Jr., T. & Ives, A.R. (2003). Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution*, **57**: 717–745.

Boyer, S.L., Johansen, J.R., Flechtner, V.R. & Howard, G.L. (2002). Phylogeny and genetic variance in terrestrial *Microcoleus* (Cyanophyceae) species based on sequence analysis of the 16S rRNA gene and associated 16S–23S ITS region. *Journal of Phycology*, **38**: 1222–1235.

Büdel, B., Darienko, T., Deutschewitz, K., Dojani, S., Friedl, T., Mohr, K.I., Salisch, M., Reisser, W. & Weber, B. (2009). Southern African biological soil crusts are ubiquitous and highly diverse in drylands, being restricted by rainfall frequency. *Microbial Ecology*, **57**: 229–247.

Cadillo-Quiroz, H., Didelot, X., Held, N.L., Herrera, A., Darling, A., Reno, M.L., Krause, D.J. & Whitaker, R.J. (2012). Patterns of gene flow define species of thermophilic Archaea. *PLoS Biology*, **10**: e1001265.

Capella-Gutiérrez, S., Silla-Martínez, J.M. & Gabaldón, T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25**: 1972–1973.

Chase, A.B., Karaoz, U., Brodie, E.L., Gomez-Lunar, Z., Martiny, A.C. & Martiny, J.B.H. (2017). Microdiversity of an abundant terrestrial bacterium encompasses extensive variation in ecologically relevant traits. *MBio*, **8**: e01809–17.

Crisp, M.D. & Cook, L.G. (2012). Phylogenetic niche conservatism: what are the underlying evolutionary and ecological causes?. *New Phytologist*, **196**: 681–694.

Curren, E. & Leong, S.C.Y. (2020). Natural and anthropogenic dispersal of cyanobacteria: a review. *Hydrobiologia*, **847**: 2801–2822.

Dvořák, P., Hašler, P. & Poulíčková, A. (2012). Phylogeography of the *Microcoleus vaginatus* (cyanobacteria) from three continents – a spatial and temporal characterization. *PLoS ONE*, **7**: e40153.

Dvořák, P., Poulíčková, A., Hašler, P., Belli, M., Casamatta, D.A. & Papini, A. (2015). Species concepts and speciation factors in cyanobacteria, with connection to the problems of diversity and classification. *Biodiversity and Conservation*, **24**: 739–757.

Edgar, R.C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**: 1792–1797.

Fenchel, T. (2003). Biogeography for bacteria. *Science*, **301**: 925–926.

Fernandes, V.M.C., Machado de Lima, N.M., Roush, D., Rudgers, J., Collins, S.L. & Garcia-Pichel, F. (2018). Exposure to predicted precipitation patterns decreases population size and alters community structure of cyanobacteria in biological soil crusts from the Chihuahuan Desert. *Environmental Microbiology*, **20**: 259–269.

Fick, S.E. & Hijmans, R.J. (2017). WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, **37**: 4302–4315.

Fierer, N. (2008). Microbial biogeography: patterns in microbial diversity across space and time. In *Accessing Uncultivated Microorganisms: From the Environment to Organisms and Genomes and Back* (Zengler, K., editor), 9–115. American Society of Microbiology, Washington, DC.

Finlay, B.J. (2002). Global dispersal of free-living microbial eukaryote species. *Science*, **296**: 1061–1063.

Flombaum, P., Gallegos, J.L., Gordillo, R.A., Rincón, J., Zabala, L.L., Jiao, N., Karl, D.M., Li, W.K.W., Lomas, M.W., Veneziano, D., Vera, C.S., Vrugt, J.A. & Martiny, A.C. (2013). Present and future global distributions of the marine cyanobacteria *Prochlorococcus* and *Synechococcus*. *Proceedings of the National Academy of Sciences USA*, **110**: 9824–9829.

Foissner, W. (2006). Biogeography and dispersal of microorganisms: a review emphasizing protists. *Acta Protozoologica*, **45**: 111–136.

Fontaneto, D. & Brodie, J. (2011). Why biogeography of microorganisms? In *Biogeography of Microscopic Organisms: Is Everything Small Everywhere?* (Fontaneto, D., editor), 3–10. Cambridge University Press, Cambridge.

Garcia-Pichel, F. & Wojciechowski, M.F. (2009). The evolution of a capacity to build supra-cellular ropes enabled filamentous cyanobacteria to colonize highly erodible substrates. *PLoS ONE*, **4**: e7801.

Garcia-Pichel, F., Prufert-Bebout, L. & Muyzer, G. (1996). Phenotypic and phylogenetic analyses show *Microcoleus chthonoplastes* to be a cosmopolitan cyanobacterium. *Applied and Environmental Microbiology*, **62**: 3284–3291.

Garcia-Pichel, F., Loza, V., Marusenko, Y., Mateo, P. & Potrafka, R.M. (2013). Temperature drives the continental-scale distribution of key microbes in topsoil communities. *Science*, **340**: 1574–1577.

Gomont, M. (1892). Monographie des Oscillariées (Nostocacées homocystées). *Annales des Sciences Naturelles, Botanique, Série*, **7**: 263–368.

González-Rocha, G., Muñoz-Cartes, G., Canales-Aguirre, C.B., Lima, C.A., Domínguez-Yévenes, M., Bello-Toledo, H. & Hernández, C.E. (2017). Diversity structure of culturable bacteria isolated from the Fildes Peninsula (King George Island, Antarctica): a phylogenetic analysis perspective. *PLoS ONE*, **12**: e0179390.

Gundlapally, S.R. & Garcia-Pichel, F. (2006). The community and phylogenetic diversity of biological soil crusts in the Colorado Plateau studied by molecular fingerprinting and intensive cultivation. *Microbial Ecology*, **52**: 345–357.

Hanson, C.A., Fuhrman, J.A., Horner-Devine, M.C. & Martiny, J.B.H. (2012). Beyond biogeographic patterns: processes shaping the microbial landscape. *Nature Reviews Microbiology*, **10**: 497–506.

Hašler, P., Dvořák, P., Johansen, J.R., Kitner, M., Ondřej, V. & Poulíčková, A. (2012). Morphological and molecular study of epipelic filamentous genera *Phormidium*, *Microcoleus* and *Geitlerinema* (Oscillatoriales, Cyanophyta/Cyanobacteria). *Fottea*, **12**: 341–356.

Hijmans, R.J. (2020). Geographic Data Analysis and Modeling [R package raster version 3.3-13]. Retrieved from https://CRAN.R-project.org/package=raster.

Hijmans, R.J., Williams, E., Vennes, C. & Hijmans M. (2017). Package 'geosphere'. Retrieved from https://CRAN.R-project.org/package=geosphere.

Hoang, D.T., Chernomor, O., Von Haeseler, A., Minh, B.Q. & Vinh, L.S. (2018). UFBoot2: improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution*, **35**: 518–522.

Hunt, D.E., David, L.A., Gevers, D., Preheim, S.P., Alm, E.J. & Polz, M.F. (2008). Resource partitioning and sympatric differentiation among closely related bacterioplankton. *Science*, **320**: 1081–1085.

Jombart, T., Balloux, F. & Dray, S. (2010). Adephylo: new tools for investigating the phylogenetic signal in biological traits. *Bioinformatics*, **26**: 1907–1909.

Kashtan, N., Roggensack, S.E., Rodrigue, S., Thompson, J.W., Biller, S.J., Coe, A., Ding, H., Marttinen, P., Malmstrom, R.R., Stocker, R., Follows, M.J., Stepanauskas, R. & Chisholm, S.W. (2014). Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. *Science*, **344**: 416–420.

Keck, F., Rimet, F., Bouchez, A. & Franc, A. (2016). Phylosignal: an R package to measure, test, and explore the phylogenetic signal. *Ecology and Evolution*, **6**: 2774–2780.

Komárek, J. & Anagnostidis, K. (2005). Cyanoprokaryota 2. Teil: Oscillatoriales. In *Süsswasserflora von Mitteleuropa* (Büdel, B., Gärdner, G., Krienitz, L. & Schagerl, M., editors), 759. Elsevier, Munich.

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, **28**: 1–26. Retrieved from https://CRAN.R-project.org/package=caret.

Kumar, S., Stecher, G. & Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution*, **33**: 1870–1874.

Larkin, A.A., Blinebry, S.K., Howes, C., Lin, Y., Loftus, S.E., Schmaus, C.A., Zinser, E.R. & Johnson, Z.I. (2016). Niche partitioning and biogeography of high light adapted *Prochlorococcus* across taxonomic ranks in the North Pacific. *ISME Journal*, **10**: 1555–1567.

Larsson, A. (2014). AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*, **30**: 3276–3278.

Leliaert, F., Verbruggen, H., Vanormelingen, P., Steen, F., López-Bautista, J.M., Zuccarello, G.C. & De Clerck, O. (2014). DNA-based species delimitation in algae. *European Journal of Phycology*, **49**: 179–196.

Losos, J.B. (2008). Phylogenetic niche conservatism, phylogenetic signal and the relationship between phylogenetic relatedness and ecological similarity among species. *Ecology Letters*, **11**: 995–1003.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, **27**: 209–220.

Miller, S.R., Castenholz, R.W. & Pedersen, D. (2007). Phylogeography of the thermophilic cyanobacterium *Mastigocladus laminosus*. *Applied and Environmental Microbiology*, **73**: 4751–4759.

Moore, J.G. (1985). Structure and eruptive mechanisms at Surtsey Volcano, Iceland. *Geological Magazine*, **122**: 649–661.

Moran, P.A.P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, **37**: 17–23.

Muñoz-Martín, M.Á., Becerra-Absalón, I., Perona, E., Fernández-Valbuena, L., Garcia-Pichel, F. & Mateo, P. (2019). Cyanobacterial biocrust diversity in Mediterranean ecosystems along a latitudinal and climatic gradient. *New Phytologist*, **221**: 123–141.

Narwani, A., Alexandrou, M.A., Herrin, J., Vouaux, A., Zhou, C., Oakley, T.H. & Cardinale, B.J. (2015). Common ancestry is a poor predictor of competitive traits in freshwater green algae. *PLoS ONE*, **10**: e0137085.

Nguyen, L.T., Schmidt, H.A., Von Haeseler, A. & Minh, B.Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, **32**: 268–274.

Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., Mcglinn, D. & Wagner, H. (2016). Vegan: Community ecology package. Retrieved from https://CRAN.R-project.org/package=vegan.

Pagel, M. (1999). Inferring the historical patterns of biological evolution. *Nature*, **401**: 877–884.

Pietrasiak, N., Mühlsteinová, R., Siegesmund, M.A. & Johansen, J.R. (2014). Phylogenetic placement of *Symplocastrum* (Phormidiaceae, Cyanophyceae) with a new combination *S. californicum* and two new species: *S. flechtnerae* and *S. torsivum*. *Phycologia*, **53**: 529–541.

Reno, M.L., Held, N.L., Fields, C.J., Burke, P.V. & Whitaker, R.J. (2009). Biogeography of the *Sulfolobus islandicus* pan-genome. *Proceedings of the National Academy of Sciences USA*, **106**: 8605–8610.

Revell, L.J. (2012). Phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, **3**: 217–223.

Ribeiro, K.F., Ferrero, A.P., Duarte, L., Turchetto-Zolet, A.C. & Crossetti, L.O. (2020). Comparative phylogeography of two free-living cosmopolitan cyanobacteria: insights on biogeographic and latitudinal distribution. *Journal of Biogeography*, **47**: 1106–1118.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**: 461–464.

Sharma, N.K. & Singh, S. (2010). Differential aerosolization of algal and cyanobacterial particles in the atmosphere. *Indian Journal of Microbiology*, **50**: 468–473.

Siegesmund, M.A., Johansen, J.R., Karsten, U. & Friedl, T. (2008). *Coleofasciculus* gen. nov. (Cyanobacteria): morphological and molecular criteria for revision of the genus *Microcoleus* Gomont. *Journal of Phycology*, **44**: 1572–1585.

Škaloud, P. & Rindi, F. (2013). Ecological differentiation of cryptic species within an asexual protist morphospecies: a case study of filamentous green alga *Klebsormidium* (Streptophyta). *Journal of Eukaryotic Microbiology*, **60**: 350–362.

Staub, R. (1961). Ernährungsphysiologisch-autökologische Untersuchungen an der planktischen Blaualge *Oscillatoria rubescens* DC. *Schweizerische Zeitschrift für Hydrologie*, **23**: 82–198.

Steven, B., Kuske, C.R., Gallegos-Graves, L.V., Reed, S.C. & Belnap, J. (2015). Climate change and physical disturbance manipulations result in distinct biological soil crust communities. *Applied and Environmental Microbiology*, **81**: 7448–7459.

Strunecký, O., Komárek, J., Johansen, J., Lukešová, A. & Elster, J. (2013). Molecular and morphological criteria for revision of the genus *Microcoleus* (Oscillatoriales, Cyanobacteria). *Journal of Phycology*, **49**: 1167–1180.

Tamura, K. & Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution*, **10**: 512–526.

Uyeda, J.C., Harmon, L.J. & Blank, C.E. (2016). A comprehensive study of cyanobacterial morphological and ecological evolutionary dynamics through deep geologic time. *PLoS ONE*, **11**: e0162539.

Van Gremberghe, I., Leliaert, F., Mergeay, J., Vanormelingen, P., Van der Gucht, K., Debeer, A.E., Lacerot, G., Meester, L.D. & Vyverman, W. (2011). Lack of phylogeographic structure in the freshwater cyanobacterium *Microcystis aeruginosa* suggests global dispersal. *PLoS ONE*, **6**: e19561.

Whitaker, R.J., Grogan, D.W. & Taylor, J.W. (2003). Geographic barriers isolate endemic populations of hyperthermophilic archaea. *Science*, **301**: 976–978.

Yu, Y., Harris, A.J., Blair, C. & He, X. (2015). RASP (Reconstruct Ancestral State in Phylogenies): a tool for historical biogeography. *Molecular Phylogenetics and Evolution*, **87**: 46–49.

Yu, Y., Blair, C. & He, X. (2020). RASP 4: ancestral state reconstruction tool for multiple genes and characters. *Molecular Biology and Evolution*, **37**: 604–606.

Paper II

Check for updates

High genomic differentiation and limited gene flow indicate recent cryptic speciation within the genus *Laspinema* (cyanobacteria)

Aleksandar Stanojković[1], Svatopluk Skoupý[1],
Pavel Škaloud[2] and Petr Dvořák[1]*

[1]Department of Botany, Faculty of Science, Palacký University Olomouc, Olomouc, Czechia,
[2]Department of Botany, Faculty of Science, Charles University in Prague, Prague, Czechia

The sympatric occurrence of closely related lineages displaying conserved morphological and ecological traits is often characteristic of free-living microbes. Gene flow, recombination, selection, and mutations govern the genetic variability between these cryptic lineages and drive their differentiation. However, sequencing conservative molecular markers (e.g., 16S rRNA) coupled with insufficient population-level sampling hindered the study of intra-species genetic diversity and speciation in cyanobacteria. We used phylogenomics and a population genomic approach to investigate the extent of local genomic diversity and the mechanisms underlying sympatric speciation of *Laspinema thermale*. We found two cryptic lineages of *Laspinema*. The lineages were highly genetically diverse, with recombination occurring more frequently within than between them. That suggests the existence of a barrier to gene flow, which further maintains divergence. Genomic regions of high population differentiation harbored genes associated with possible adaptations to high/ low light conditions and stress stimuli, although with a weak diversifying selection. Overall, the diversification of *Laspinema* species might have been affected by both genomic and ecological processes.

KEYWORDS

cryptic species, cyanobacteria, gene flow, phylogenomics, recombination, sympatric speciation

Introduction

Comparative and population microbial genomics increased our understanding of genome evolution and genetic diversity among bacterial species (Koonin et al., 2021). Although cyanobacteria are a morphologically highly diverse group, many genera and species lack distinct phenotypic and ecological characteristics that would aid their delineation – cryptic species and genera (Casamatta et al., 2003; Dvořák et al., 2015). Whole-genome sequencing on a population level can overcome this problem by capturing fine genetic differences between cryptic species (Dvořák et al., 2017). Traditional markers such as 16S rRNA and 16S-23S ITS region are not variable enough to capture the diversity

among closely related species. Population genomic studies on cyanobacteria (e.g., *Microcystis*, Pérez-Carrascal et al., 2019; *Prochlorococcus*, Kashtan et al., 2014), human pathogens (e.g., *Lactobacillus salivarius*, Harris et al., 2017), and free-living soil bacteria (e.g., *Curtobacterium*, Chase et al., 2019) revealed that genetic differences among closely related species could actually be huge, despite not being exhibited by the phenotype.

By introducing new genes or allele variations between prokaryotic lineages, the divergence leads to genomic differentiation. However, the gene flow may still be present in sympatry (Polz et al., 2013). In addition to the gene flow, rearrangements and mutations can generate the genetic variability within closely related locally coexisting bacteria and promote their differentiation (Simmons et al., 2008; Reno et al., 2009; Barraclough et al., 2012; Shapiro et al., 2012). Wiedenbeck and Cohan (2011) noted that the evolution of bacteria is affected by ecological niches. Therefore, distinct genetic groups (species) that occupy different ecological niches arose in sympatry. However, cyanobacterial studies often have a limited number of closely related strains from the same locality. Comparing genome diversity on a population level is vital to fill the gap in understanding the mechanisms driving and maintaining the divergence in coexisting strains.

Gene flow in the form of homologous and non-homologous recombination is widespread across bacterial phyla, enabling genetic exchange within and between distinct groups (Bobay and Ochman, 2017; Sheinman et al., 2021). Although this mechanism of gene exchange frequently occurs between microbes, it is not simple to detect. Significant challenges to identifying recombination events are, for instance, overall sampling effort, the number of informative sites between sequences, and the evolutionary relationships among strains (Martin et al., 2011). Strains that are more distantly related have lower signals of recombination since recombination decays with sequence divergence (Bobay and Ochman, 2017).

Gene flow occurs regardless of evolutionary relatedness and differs among microorganisms (Bobay, 2020). The recombination has a lower impact on genetic diversity in free-living terrestrial bacteria, i.e., they are mostly clonal populations, whilst it has a higher impact on others, e.g., pathogenic (Vos and Didelot, 2009). Interestingly, recombination rates among species of the same genus can be substantially different. Such patterns were observed within *Microcystis* (Pérez-Carrascal et al., 2019) and *Prochlorococcus* (Coleman and Chisholm, 2010). Moreover, lower recombination rates and higher phylogenetic distance between strains can be responsible for maintaining cohesive genetic groups within closely related microbial lineages by affecting fine ecological differences between them. The limitations to recombination are associated with barriers like sequence divergence (Cordero et al., 2012), bacterial (in) competence (Jeltsch, 2003; Carrolo et al., 2009), selection, and ecological structuring (Shapiro and Polz, 2014); thus, enabling the species divergence. The presence of almost complete or partial (limited to some genomic segments) barriers to gene

exchange in a sympatric setting led to the divergence among strains of, e.g., *Sulfolobus* (Cadillo-Quiroz et al., 2012), *Vibrio* (Shapiro et al., 2012), and *Myxococcus* (Wielgoss et al., 2016). Although Pérez-Carrascal et al. (2019) showed that gene flow and recombination might be major drivers of speciation in cyanobacteria, the impact of recombination in most of them remains largely unexplored.

Terrestrial habitats harbor a tremendous diversity of bacteria (Grundmann, 2004). In this study, we investigated *Laspinema thermale*, which was first found in thermal springs (Heidari et al., 2018) and thermal mud (Duval et al., 2020). *Laspinema* is a recently described cyanobacterial genus with the type species *L. thermale*. Currently, the genus *Laspinema* encompasses several species *L. thermale*, *L. etoshii*, *L. lumbricale*, and previously misidentified *Laspinema* species - "*Oscillatoria*" *acuminata* (Zimba et al., 2020). However, due to the undersampling of this genus, the evolutionary history of *Laspinema* species is still not entirely resolved. Here, based on morphological features and phylogenetic analyses, we will refer to our strains as "*Laspinema* sp."

We aim to investigate patterns of genetic variability within *Laspinema* sp. lineages during speciation and the processes which affect it. We reconstruct the evolutionary history of isolated strains and other cyanobacteria. We next examine pangenome and fine-scale genomic diversity among coexisting *Laspinema* sp. strains in ongoing divergence with gene flow. Finally, we investigate genomic signatures of local selection and homologous recombination (HR) in diverging lineages.

Materials and methods

Sample collection, whole-genome sequencing, and assembly

Two samples (D2 and D3) were collected from different sides of a puddle with a sterile spatula and placed directly in a sterile plastic bag in September 2018 in Olomouc, Czech Republic (49.57459 N, 17.281988 E). The puddle is a concave water body 5–10 cm in depth, and the samples were centimeters apart. A portion of each sample was placed in 10 ml capped tubes with liquid Zehnder medium (Staub, 1961). Part of the grown biomass was then transferred to Petri dishes on solid agar Zehnder medium (1.5%) and unialgal cultures were isolated following Hašler et al. (2012). By isolating a single filament, we obtained four clonal cultures from D2 and five from the D3 sample. Altogether we had eight *Laspinema* sp. strains and one outgroup strain, *Ancylothrix* sp. Their morphology was assessed under 1,000× magnification using ZEISS Primo Star (Oberkochen, Germany) light microscope. Strains were identified following taxonomic system *sensu* Komárek and Anagnostidis (2005). The strains are maintained in culture collection at the Department of Botany, Palacký University in Olomouc. All cultures were grown in 10 ml capped tubes with liquid Zehnder medium, maintained at

$22 \pm 1°C$ and illuminated with an average photon flux density of $20\,\mu\text{mol}$ photons m^{-2} s^{-1} under regime 12h light/12h dark.

To confirm that our strains belonged to *Laspinema* sp., we isolated genomic DNA, amplified and purified partial 16S rRNA and 16S-23S ITS region according to Stanojković et al. (2022). The PCR products were sequenced by the Sanger sequencing method at Macrogen Europe, Inc. (Amsterdam, the Netherlands).[1] Sequences were identified using the BLAST nucleotide search.[2] Each 16S rRNA and 16S-23S ITS sequence represented one clonal culture. GeneBank accession numbers of 16S rRNA and 16S-23S ITS sequences are in Supplementary Table S1A.

We used 100mg of fresh biomass and extracted genomic DNA using UltraClean Microbial DNA Isolation Kit (MOBIO, Carlsbad, CA, United States), following the manufacturers' recommendations. DNA quality and concentration were assessed using ethidium bromide-stained 1.5% agarose gel and NanoDrop 1,000 (Thermo Fisher Scientific, Wilmington, DE, United States), respectively. The DNA fragments' size was assessed by Agilent 5,400 fragment analyzer system (Agilent Technologies, Santa Clara, CA, United States). Sequencing libraries were prepared using Nextera XT DNA Library Prep Kit (Illumina, San Diego, CA, United States). Samples were sequenced commercially on Illumina NovaSeq 6,000 platform (Novogene, United Kingdom) as a paired-end $2 \times 150\,\text{bp}$ layout. Nine whole-genome sequences were submitted to the NCBI database under the BioProject PRJNA849373.

Short raw reads were filtered and trimmed using Trimmomatic v0.39 (Bolger et al., 2014) with the following parameters ILLUMINACLIP:2:30:10, LEADING:3, TRAILING:3, SLIDINGWINDOW:4:15, and MINLEN:50. Genomes were assembled with SPAdes genome assembler v3.13.1 (Bankevich et al., 2012). Produced scaffolds were clustered into different bins by MaxBin v2.2.4 (Wu et al., 2016), where a single bin represented one genome. The completeness and contamination levels were assessed with CheckM v1.1.9 software (Parks et al., 2015). Annotations were performed using prokka v1.14.5 (Seemann, 2014).[3]

Short read mapping and variant calling

Raw Illumina reads with a quality score of less than three and trimmed lengths of less than 50 bases were filtered and trimmed using Trimmomatic v0.39. Trimmed and filtered reads were mapped to the reference genome of "*Oscillatoria*" *acuminata* PCC 6304 (CP003607) using Burrows-Wheeler Alignment mem v0.7.17 (Li and Durbin, 2009). Duplicated reads were marked using Picard Toolkit v2.25.1 (Broad Institute).[4] The mapped reads were sorted and indexed by SAMtools (Li et al., 2009). Variants for individual genomes without the outgroup were called using HaplotypeCaller with-ERC BP_RESOLUTION parameter within

GATK v4.2.0.0 (McKenna et al., 2010). CombineGVCFs and GenotypeGVCFs modules in GATK were used to merge variant call format (vcf) files and perform genotype calls. Single nucleotide polymorphisms (SNPs) were filtered using hard-filtering parameters following GATK's best practices pipeline and indels were removed. Additionally, we used VCFtools v0.1.16 (Danecek et al., 2011) to filter out sites with minimum allele frequency less than 0.1, mean depth values less than 40 and greater than 200, and genotype quality smaller than 30.

Phylogenetic and phylogenomic analyses

The most similar 16S rRNA sequences were identified using BLAST nucleotide search, and in Supplementary Tables S1B–J are all hits with $\geq$95% sequence identities. Additionally, reference sequences of *Oscillatoria*, *Phormidium*, and other cyanobacteria were added. *Gloeobacter violaceus* (NR074282) was used as an outgroup. Multiple sequence alignment was performed by the Muscle algorithm v3.8.425 (Edgar, 2004) in AliView (Larsson, 2014). The maximum likelihood (ML) phylogeny was performed with IQ-TREE v1.6.1 (Nguyen et al., 2015) using the GTR+I+G substitution model. Ultrafast bootstrapping with 2000 replicates was used (Hoang et al., 2018).

We used Orthofinder v2.3.1 (Emms and Kelly, 2019) with default parameters to infer multiple sequence alignment and investigate evolutionary histories among cyanobacteria and *Laspinema* sp. strains. Altogether, we used 133 whole-genome sequences. Next, we performed ML phylogenetic reconstruction in IQ-TREE with the best model selected by ModelFinder (Kalyaanamoorthy et al., 2017) – LG+F+I+G4 and ultrafast bootstrap with 2000 replicates. A list of single-copy orthologues used for the phylogenomic inference can be found in Supplementary Table S2A.

Eight whole-genome sequences of *Laspinema* sp. and a reference genome "*Oscillatoria*" *acuminata* PCC 6304 were used to infer whole-genome phylogeny. We used *Ancylothrix* sp. (strain D3o) as the outgroup. Thus, our dataset encompassed ten whole-genome sequences. OrthoFinder with the option-T iqtree was employed to identify single-copy orthologues (Supplementary Table S2B), infer unrooted gene trees, and generate multiple sequence alignment. Species trees were built using three different approaches. First, the ML tree was inferred in IQ-TREE based on the LG+I+G4 substitution model from multiple sequence alignment. The tree topology was tested using ultrafast bootstrapping with 2000 replicates. Second, a coalescent-based analysis was performed with ASTRAL-III (Mirarab and Warnow, 2015; Zhang et al., 2018) using gene trees inferred from individual single-copy orthologue alignments. For the third approach, we used the variant call dataset with the only variable sites obtained from GATK. The vcf file was converted to fasta format with vcf2phylip script[5] and then we used fasta file for the ML phylogeny inference in IQ-TREE. As the most

appropriate model was selected TVM + F + ASC by ModelFinder within IQ-TREE, and 2,000 replicates were used.

The proportions of dissimilarities (p-distance) between 16S rRNA sequences were calculated in MEGA11 (Tamura et al., 2021). To assess the genomic similarity between nine genomes (outgroup omitted), average nucleotide identity (ANI) was calculated with FastANI v1.33.[6] Both 16S rRNA and ANI matrices are in Supplementary Table S3. ANI matrix was visualized with 'gplots' v3.1.1 (Warnes et al., 2016) in R (version 4.1.2; R Core Team, 2021).

We used the software Saguaro v0.1 (Zamani et al., 2013) to investigate if genomic regions support different local phylogenetic relationships among aligned sequences of isolated *Laspinema* sp. strains. It detects boundaries between genomic segments that have different phylogenetic relationships and assigns "cacti" to them. Consensus genome sequences were created from the vcf file using VCFtools and whole-genome alignment was made. Saguaro was run with 20 iterations and it identified 21 cacti. We used the Saguaro2Phylip function to convert cacti into distance matrices suitable for phylogenetic analysis in Phylip v3.696 (Felsenstein, 2005) with the neighbor function. Boundaries between genomic segments with different cacti assigned to them were visualized with paint_chromosomes ruby script[7] and the trees using the program FigTree v1.4.4 (Rambaut, 2012).

Population genomic analyses

We investigated the pangenome of isolated *Laspinema* sp. strains and *"Oscillatoria" acuminata* PCC 6304 to reveal their genomic diversity. Genome annotations were used as an input to Roary v3.13.0 (Page et al., 2015) and pangenome was visualized using roary_plots python script.[8] Genes were clustered into core and flexible or accessory (shell and cloud) genomes. Core genes are found in all nine genomes, shell genes are present in 2–8 genomes, while cloud genes are found in a single genome.

For analyses sensitive to population definitions, we define population D2 consisting of D2a, D2b, and D2c; and the D3 population with D3a, D3b, D3c, and D3d strains, following detected recombination between them and consistent monophyly of the two clades across phylogenies (see results and discussion). The outlier strain D2d was included as an outgroup for recombination analyses but was excluded from the rest of the analyses requiring population-level data.

To explore the differences in functionality of flexible (cloud) genes among strains of D2 and D3 lineages, we performed gene ontology (GO) enrichment analysis. Sequences of flexible genes were annotated with pannzer2 webserver (Törönen et al., 2018),

and GO classes were predicted per strain. Enriched GO terms were searched using g:Profiler (Raudvere et al., 2019) using a diatom *Thalassiosira pseudonana* CCMP1335, due to the lack of species closely related to cyanobacteria in the database. GO terms with Benjamini-Hochberg FDR value of $p < 0.05$ were regarded as significant.

Intra-population diversity was assessed with nucleotide diversity (π). Inter-population genetic differentiation (F_{ST}) and divergence (D_{XY}) were estimated between *Laspinema* sp. strains (D2 and D3 lineages). We used package 'PopGenome' v2.7.5 (Pfeifer et al., 2014) in R (version 4.1.2) in 50 kb sliding windows with 12.5 kb step. One window yielded a high F_{ST} value with a few SNP sites and low read coverage and was excluded from the analysis (window coordinates: 2,837,501–2,887,501). The outputs were visualized using package ggplot2 v3.3.5 (Wickham, 2016) in R. To estimate the differences between lineages for nucleotide diversity, we performed Mann–Whitney U statistical test in R. To investigate the linkage disequilibrium (LD) over the whole genome, we calculated pairwise SNP correlation using plink v190 (Purcell et al., 2007) on a thinned D2 and D3 vcf files. Thinning was performed in VCFtools by keeping one SNP every 250 positions.

The genomes were scanned for HR using SplitsTree v4.18.1 (Huson and Bryant, 2006) and Gubbins v3.1.3 (Croucher et al., 2015). Phi test in SplitsTree calculates pairwise homoplasy index in genome alignment and reports phi statistics with assessed significances. Gubbins was run according to the software manual, with –first-tree-builder rapidnj and –tree-builder raxmlng parameters. The resulting phylogenetic tree and detected recombination sites were visualized with Phandango (Hadfield et al., 2018). We also calculated the ratio of recombination to mutation rate (ρ/θ) and the ratio of imported SNPs through recombination relative to substitution (*r/m*) between and within both populations.

Genome-wide scans for positive selection

Genomic regions with an elevated F_{ST} value (loci within 0.99 percentile) and suppressed recombination were extracted to investigate the genomic regions responsible for the differentiation between populations D2 and D3. Potential gene functions were obtained from the UniProt protein database.[9] Extracted nucleotide and protein sequences of genes were aligned with muscle algorithm, and codon alignments were made using perl script pal2nal (Suyama et al., 2006). On individual codon alignments of genes found in the regions of elevated F_{ST}, we performed the McDonald-Kreitman (MK) test in DnaSP v6.12.03 (Rozas et al., 2017) and calculated d_N/d_S ratios using SNAP v2.1.1 (Korber, 2000).

---

6   https://github.com/ParBLiSS/FastANI

7   https://github.com/IsmailM/

8   https://github.com/sanger-pathogens/Roary/

---

9   https://www.uniprot.org

Results

Evolutionary relationships

Phylogenetic reconstruction of 16S rRNA using ML placed our isolates in the family Laspinemaceae, in the same clade as *"Oscillatoria" acuminata* PCC 6304, *L. etoshii*, *L. lumbricale*, *Phormidium pseudopriestleyi*, and *Perforafilum tunelii* (Supplementary Figure S1). According to the phylogenetic inference, all our sequences had the highest similarity to *L. thermale*. Short branches between our *Laspinema* sp. isolates revealed very little divergence among them. P-distance analysis of 16S rRNA confirmed that all *Laspinema* sp. strains had a high sequence similarity of 99.4–100% between each other and 98.8–99.1% similarity with *"Oscillatoria" acuminata* PCC 6304 (Supplementary Table S3A).

As these conservative markers often mask the real diversity among species, we sequenced genomes of our *Laspinema* strains. *De novo* assembly yielded genomes of length 6.5–7.4 Mb with 43.5–48% GC content. Genomes were 99–100% complete with <2% contamination. All genome features are presented in Supplementary Table S4. We reconstructed the phylogenetic relationships between eight *Laspinema* sp. strains and 125 cyanobacteria using multiple sequence alignment with 165,686 amino acid sites (Supplementary Figure S2). *Laspinema* sp. isolates clustered together in a monophyletic clade with *"Oscillatoria" acuminata* PCC 6304 and *Phormidium pseudopriestleyi* FRX01 as sister species (Supplementary Figure S2).

Then, we reconstructed the evolutionary histories among our *Laspinema* sp. strains, the outgroup (*Ancylothrix* sp.), and *"Oscillatoria" acuminata* PCC 6304. The ML species tree was inferred from the multiple sequence alignment with 892,706 amino acid sites (Figure 1A). We then compared species trees inferred by three different approaches (aforementioned ML tree from 2,595 single-copy orthologues, ML tree from SNPs, and the ASTRAL tree from a set of unrooted gene trees under the multi-species coalescent model) to investigate the evolutionary histories among *Laspinema* sp. strains. All the analyses supported the differentiation of D3 and D2 strains in two clades, where D2d clustered with the D3 clade but with low bootstrap support of <60 (Figures 1A,B; Supplementary Figure S3). However, there was a discordance in the phylogenetic position of *"Oscillatoria" acuminata* PCC 6304. The ASTRAL analysis recovered it as a sister to the clade D2, while in the ML analysis, it appeared more distantly related, with low bootstrap support (<60).

The analyses showed that isolates D2a and D2b shared 97.6% ANI, while D3a, D3b, D3c, and D3d, shared 97.9–98.4% sequence identity across their genomes (Figure 2; Supplementary Table S3B). All *Laspinema* isolates had <91% similarity to *"Oscillatoria" acuminata* PCC 6304. Strain D2c had 93.9% genome similarity to D2a and D2b strains and <91% genome similarity to the D3 population, whereas the D2d shared the lowest sequence identity with all other strains (<90% ANI). In other words, following the 95% ANI threshold for species delineation (Goris et al., 2007), our

dataset consisted of five species – D2, D3, D2d, D3c, and *"Oscillatoria" acuminata* PCC 6304.

We used the Saguaro program to infer the distribution of the phylogenetic tree topologies over the genome. The program identified 1,237 segments and 21 cacti (Figure 1C). The length of genomic segments and cacti assigned to them are in Supplementary Table S5, and the percentage distribution of every cactus is in Supplementary Figure S4. Ten detected cacti covering 91% of the genome featured the expected differentiation of isolates into two clades, D2 and D3. The rest of identified cacti supported different local topologies and covered 6.88% of the genome (Figure 1C; Supplementary Figure S4). Cacti not assigned to any region covered 2.13% of the genome. The most common cacti (cactus4–27.22% and cactus0–24.44% of the genome) were scattered across the genome, and they represented topologies featuring the expected clade differentiation (Supplementary Figure S4).

The strain D2d likely represents a separate lineage due to its unresolved phylogenetic position, low sequence identity, pangenome content, and the lack of recombination events it shared with other *Laspinema* sp. isolates. As we lacked closely related strains to the D2d, we excluded it from the rest of the analyses based on pre-defined populations.

Laspinema sp. pan-genome

As a result of pangenome analysis, Roary clustered all coding DNA sequences (CDS) in the core and flexible genome (shell and cloud). The core genome included genes present in all nine strains; the shell had genes present in 2–8 strains, while the cloud genome had genes present in a single strain. Altogether, *"Oscillatoria" acuminata* PCC 6304 and eight *Laspinema* sp. strains had a pangenome of 21,048 genes (Table 1; Figure 3). The core genome had 1,497 CDS–7.11% of the pangenome and the flexible (shell) had 6,831 CDS–32.45% of the pangenome (Table 1). Moreover, the flexible genome (cloud) accounted for 60.43% of the pangenome with 12,720 singleton genes. In all isolates, the core genome accounted for 26.28–29.16% of the whole genome (Supplementary Table S4).

We then performed GO enrichment analysis on genes present only in the flexible genome (cloud) of D2 and D3 populations. Classification by molecular function showed that both populations had a number of genes in charge of catalytic activity (GO:0003824), organic cyclic compound binding (GO:0097159), and ion binding (GO:0043167). Under the catalytic activity category, there were genes responsible for oxidoreductase activity (GO:0016491), transferase activity (GO:0016772), and kinase activity (GO:0016301), etc. Classification by biological process included genes related to the metabolic processes, e.g., organic substance metabolic process (GO:0071704) and nitrogen compound metabolic process (GO:0006807). According to cellular component classification, there were genes regulating cellular anatomical entity (GO:0110165) and integral component of membrane (GO:0016021). The D3 population had GO terms associated with carbohydrate binding (GO:0030246), metal cluster binding (GO:0051540), and iron–sulfur cluster binding

FIGURE 1
Phylogenetic relationships and genealogical discordance in the chromosome of *Laspinema*. **(A)** Maximum likelihood (ML) species tree based on 2,595 single-copy orthologues. Asterisks represent ultrafast bootstrap support ≥99. The scale bar indicates substitutions per site. **(B)** ASTRAL species tree based on individual gene trees of the 2,595 single-copy orthologues. Asterisks represent coalescent units of 1. The scale bar indicates substitutions per site. **(C)** Saguaro plot illustrating the distribution of local phylogenetic topologies over the chromosome of *Laspinema*. Genomic regions supporting the differentiation of isolates into two clades (D2 and D3) are shown in dark grey, and genomic regions not supporting it are shown in red. The letter 'p' marks the location of the plasmid, and it is shown in orange. Genomic regions in white were not assigned to any of the detected cacti. See Supplementary Figure S4 for further details on local topologies inferred by Saguaro with the percentage of the genome cacti covered.

(GO:0051536) that were not found in the D2. Contrarily, the D2 population had more unique GO terms. They were genes responsible for different catalytic activities, e.g., ATP-dependent activity (GO:0140657) or helicase activity (GO:0004386), and metabolic processes, e.g., biosynthetic process (GO:0009058), protein metabolic process (GO:0019538), etc. A list of all GO terms, their functions, adjusted *p*-values associated with them, and highlighted differing GO terms of populations are presented in Supplementary Table S6.

Population differentiation

We used four metrics (F_{ST}, π, D_{XY}, Tajima's D) to estimate genetic diversity and population differentiation among *Laspinema* sp. isolates. The mean nucleotide diversity was 0.0179 for D2 and 0.0059 for the D3 population (Supplementary Figure S5). The nucleotide diversity of D3 was significantly lower than D2 (Mann–Whitney test, $p < 2.2e$-16). The mean absolute divergence at the

interspecific level was 0.039 and the mean fixation index was 0.69 (Figure 4). In genomic regions of elevated F_{ST} value (99th percentile) and suppressed recombination, we found 26 annotated genes. Seven genes were associated with different metabolic processes, three with DNA or RNA processing, five with adaptation to low/high light conditions, and 11 with adaptation to stress stimuli (Supplementary Table S7). Most of the loci were under weak selection (Supplementary Table S8).

Then we assessed the impact of HR on *Laspinema* sp. genome diversity. Phi statistics (the pairwise homoplasy index) supported that HR was present within *Laspinema* sp. isolates ($p < 0.0001$). The mean r^2 value was 0.702 for D2 and 0.413 for the D3 population indicating a medium linkage between loci over the genomes, consistent with the higher HR among D3 strains. Identified recombination fragments detected by Gubbins are presented in Supplementary Figure S6. We also calculated the ρ/θ and *r/m* ratios between and within D2 and D3 populations. Averaged per-branch ratios of ρ/θ and *r/m* were low for D2 (ρ/θ = 0.016; *r/m* = 0.822), whereas they were intermediate in D3

FIGURE 2
Pairwise genome-wide average nucleotide identity (ANI) heatmap showing 8 *Laspinema* sp. strains along with the reference strain *"Oscillatoria" acuminata* (marked PCC 6304). Red indicates a high ANI and blue indicates a lower ANI value. The scale bar represents color codes of ANI values, showing sequence identity percentage. See Supplementary Table S3B for the exact ANI values. The heatmap was created with heatmap.2 function within 'gplots' (Warnes et al., 2016) package in R.

TABLE 1 The proportion of protein-coding genes in the core, shell, and cloud of eight *Laspinema* sp. strains and *"Oscillatoria" acuminata* PCC6304 pangenome.

	Strain number	Gene number
Core genes	9	1,497
Shell genes	2–8	6,831
Cloud genes	1	12,720
Total genes		21,048

($\rho/\theta = 0.023$; $r/m = 1.368$). Moreover, low recombination rate was detected between populations ($\rho/\theta = 0.011$; $r/m = 0.608$). A summary of r/m values between different prokaryotic species is shown in Table 2. The fraction of the genome affected by HR was between 3.27–12.98% for the D2 and 9.05–11.67% for the D3 population. The mean fraction of the genome subjected to HR was 10.06% (Supplementary Table S9). Gubbins detected almost no HR of D2d with other *Laspinema* sp. strains, with 0.23% of its genome being subject to HR.

Discussion

We used the population genomic approach to gain deeper insight into the evolutionary processes of closely related free-living cyanobacteria within the genus *Laspinema*. We found two new cyanobacterial lineages that are highly genetically divergent, despite having almost identical 16S rRNA sequences and being isolated from the same locality. The pangenome analysis revealed that gene content within the flexible genome differs between them and might be responsible for the local adaptation. Moreover, the lineages exhibit low recombination rates between them. Finally,

the two lineages had suppressed recombination in genes associated with the adaptation to low/high light conditions and stress stimuli, suggesting their tentative role as niche-adaptive genes. Our results indicate that the two lineages are in the process of divergence, which might be driven by both ecological and genetic separation.

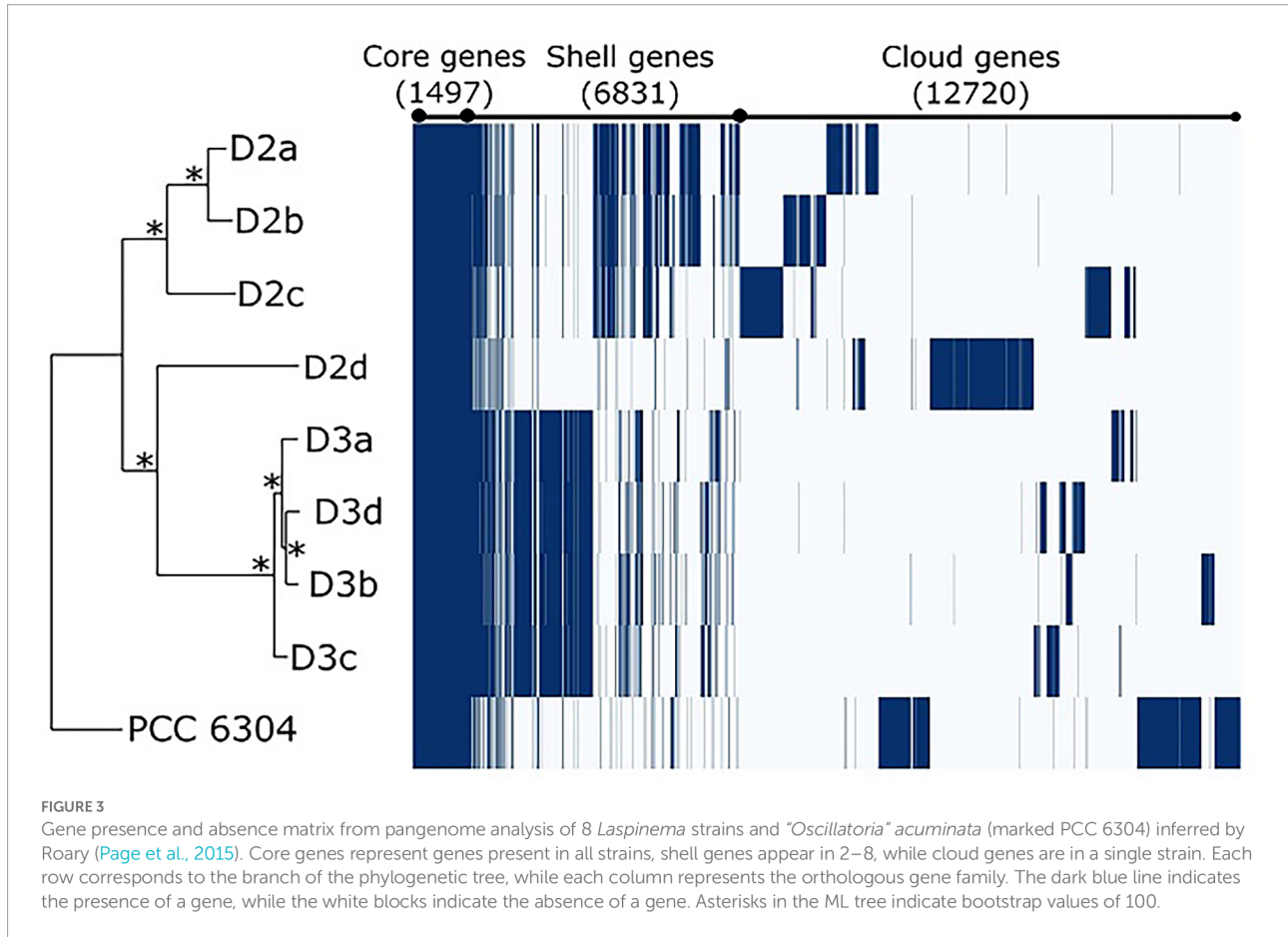## Thin line between cryptic lineages and species within *Laspinema thermale*

Many microbial taxa consist of morphologically almost indistinguishable but genetically differentiated lineages. These cryptic lineages are widespread across prokaryotes imposing significant issues on their taxonomic resolution (Dvořák et al., 2014; Walk, 2015). This issue arises due to the inconsistent application of species concepts, low phenotypic diversity among strains, and the limited sampling effort of many prokaryotic taxa (Ward et al., 2008). Some authors had already employed ecological and genetic data to overcome cryptic diversity and describe morphologically similar or identical closely related cyanobacteria (e.g., *Symplocastrum*, Pietrasiak et al., 2014). Moreover, screening the flexible genome for intrapopulation gene content and function variation could reveal if distinct lineages are associated with distinct ecological niches (Shapiro and Polz, 2014). However, detecting such differences in gene content is hindered by the presence of many genes with still unknown functions (Rodriguez-Valera and Ussery, 2012). Other authors suspected that lineages might represent different species if they formed distinct genetic clusters occupying different ecological niches with limited gene flow between them (*Vibrio*; Hunt et al., 2008; *Sulfolobus*; Cadillo-Quiroz et al., 2012; *Microcystis*; Pérez-Carrascal et al., 2019). We observed similar patterns of lineage differentiation in *Laspinema*. We suspect the divergence of sympatric D2 and D3 lineages was initiated by ecological heterogeneity, and the prevalence of population-specific recombination maintains them as distinct genetic clusters. That would be consistent with the biological species concept and dominant gene flow within rather than between lineages (Mayr, 1942; Shapiro and Polz, 2015). Both lineages have low morphological diversity (data not shown) and almost complete barriers to gene flow (discussed below), but they are still in the intermediate, so-called "grey zone" of speciation (Roux et al., 2016; Kollár et al., 2022). The D2 and D3 are on separate evolutionary pathways, likely to further diverge in time and space.

Nevertheless, additional investigations of the exact ecological and morphological differences as well as (dis)continuity of gene flow between the cryptic species are required to characterize them fully.

Whole-genome sequencing facilitates the taxonomic resolution of cyanobacteria

Reconstructing the evolutionary histories and understanding the extent of cyanobacterial diversity are hindered by insufficient

FIGURE 3
Gene presence and absence matrix from pangenome analysis of 8 *Laspinema* strains and *"Oscillatoria" acuminata* (marked PCC 6304) inferred by Roary (Page et al., 2015). Core genes represent genes present in all strains, shell genes appear in 2−8, while cloud genes are in a single strain. Each row corresponds to the branch of the phylogenetic tree, while each column represents the orthologous gene family. The dark blue line indicates the presence of a gene, while the white blocks indicate the absence of a gene. Asterisks in the ML tree indicate bootstrap values of 100.
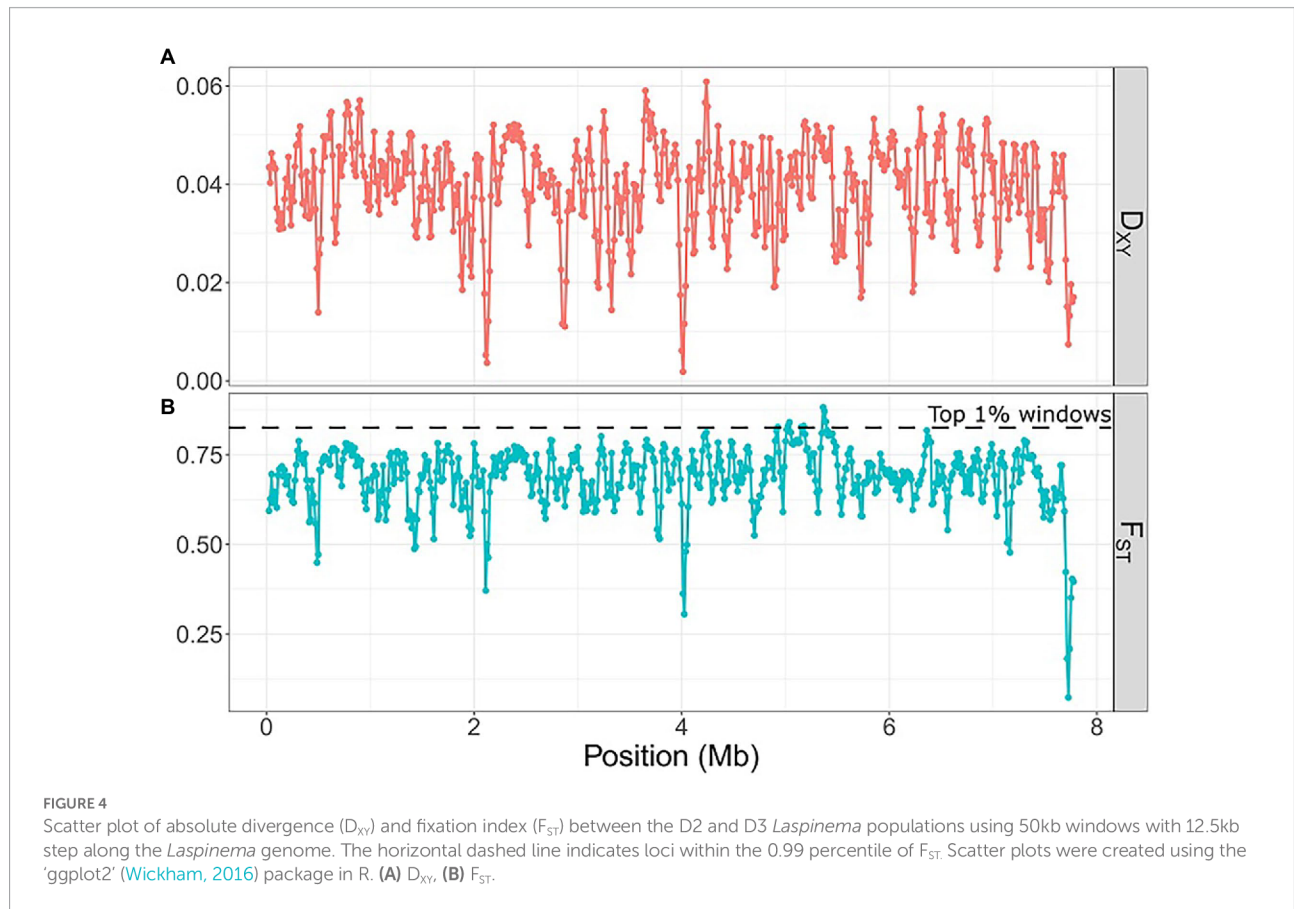
population-level sampling (Pietrasiak et al., 2019). The sequencing of conservative molecular markers has been a cornerstone for the characterization of prokaryotic organisms. Hence, a much larger genetic diversity among closely related microbes has been overlooked (Jaspers and Overmann, 2004; Yarza et al., 2014). Here, the ML phylogeny of 16S rRNA and p-distance similarity matrix (16S rRNA) showed that all *Laspinema* sp. isolates belong to one species – *L. thermale* (Supplementary Figure S1; Supplementary Table S3A). The strains exhibited short branches and clustered together, which suggests a low level of divergence among them (<0.55% divergence in 16S rRNA).

Nevertheless, employing whole-genome sequencing of closely related coexisting strains allowed us to detect the separation of *Laspinema* sp. into two highly supported (bootstrap value 100) well-differentiated subclades – D2 and D3 (Figures 1A,B; Supplementary Figures S3, S4). The discordant position of *"Oscillatoria" acuminata* PCC 6304 in the species trees indicates a potential presence of the gene tree conflict due to the incomplete lineage sorting (Figure 1B). Besides, the strain D2d had an unresolved position in the species tree with low support (Figure 1A). These results support the existence of at least four young species within *Laspinema* and highlight that more extensive population-level sampling is required to resolve phylogenetic relationships between them.

Previous studies suggested a 95–96% ANI cutoff for species delineation (e.g., Goris et al., 2007; Yarza et al., 2014). According to the 95% threshold, the ANI analysis yielded five independent lineages – D2, D3, *"Oscillatoria" acuminata* PCC 6304, D2c, and D2d (Figure 2). However, the characterization of new species based on arbitrary sequence thresholds can be ambiguous as they disregard evolutionary causes of divergence (see Bobay, 2020). For instance, mutation rates, genetic exchange, and selection pressures can significantly differ among microbes (Vos and Didelot, 2009; Bobay, 2020). Thus, the 95% ANI threshold for species delineation is likely to be lower in *Laspinema*, and we consider D2c to be a part of the lineage – D2, which was consistently monophyletic with the D2a and D2b strains in phylogenetic analyses. As previously highlighted, we suspect the D2d strain might be a new species. Still, due to the lack of population data, we excluded it from analyses that needed pre-defined populations.

Conflicts in phylogenetic relationships among organisms arise from incongruent evolutionary histories between individual genes or different parts of the genome (Galtier and Daubin, 2008). The extent of phylogenetic incongruence in microbes has rarely been researched genome-wide (including coding and non-coding regions), especially in cyanobacteria. We used Saguaro to detect and visualize conflicting topologies over the whole chromosome (Figure 1C; Supplementary Figure S4). Although ML and ASTRAL species trees revealed high support for the phylogenetic

FIGURE 4
Scatter plot of absolute divergence (D_{XY}) and fixation index (F_{ST}) between the D2 and D3 *Laspinema* populations using 50kb windows with 12.5kb step along the *Laspinema* genome. The horizontal dashed line indicates loci within the 0.99 percentile of F_{ST}. Scatter plots were created using the 'ggplot2' (Wickham, 2016) package in R. **(A)** D_{XY}, **(B)** F_{ST}.

TABLE 2 The relative effect of recombination and mutation (*r/m*) for different microbial species based on whole genomes or multilocus sequence analysis from previous studies.

Species	Phylum	*r/m*	Reference
Laspinema sp. (D3)	Cyanobacteria	1.368	This study
Laspinema sp. (D2)	Cyanobacteria	0.822	This study
Laspinema sp. (D3&D2)	Cyanobacteria	0.608	This study
Streptococcus pneumoniae	Firmicutes	0.0–45.4	Chaguza et al. (2016)
Listeria monocytogenes	Firmicutes	0.01–0.5	Zamudio et al. (2020)
Vibrio parahaemolyticus	γ-proteobacteria	1.5–24.3	Martinez-Urtaza et al. (2017)
*Microcystis aeruginosa**	Cyanobacteria	1.5–13.9	Pérez-Carrascal et al. (2019)
*Sulfolobus islandicus**	Thermoprotei (Archaea)	3.8	Held et al. (2010)

Asterisks mark *r/m* ratios obtained with the program other than Gubbins.

relationships between our *Laspinema* strains, we noted that 6.88% of the genome is discordant with the species tree topology. Those regions do not support distinct monophyletic clades of D2 and D3, and they could be affected by incomplete lineage sorting or gene flow (Galtier and Daubin, 2008). Genomic segments

supporting the differentiation of *Laspinema* sp. isolates into two monophyletic clades comprise 91% of the genome (10 cacti; Figure 1C; Supplementary Figure S4) and indicate that they are likely separate species. A similar example of scattered congruent and incongruent topologies across the genome has been found in animals – *Heliconius* butterflies (Martin et al., 2013).

Sympatric, cryptic Laspinema lineages D2 and D3 are substantially genetically differentiated (Figures 2, 3), but they lack distinct morphological characters (data not shown). Whole-genome sequencing facilitated the taxonomic resolution of cyanobacteria and we demonstrated the existence of two likely new species within *L. thermale*, despite cryptic diversity.

Extensive diversity of *Laspinema* sp. pangenome

The pangenome analyses revealed intraspecies genetic diversity and further supported that strains D2d and "*Oscillatoria*" *acuminata* PCC 6304 as well as clades D2 and D3 represent co-occurring genetically distinct lineages in *Laspinema* sp. Each lineage had its own block of genes within the flexible genome (cloud+shell), i.e., the gene pool of *Laspinema* sp. is extensive (Figure 3). These gene blocks accounted for ~60% of the whole genome, a common proportion observed in free-living microbes (McInerney et al., 2017).

The pangenome compositions differed between the D2 and D3, which is explained by strains sharing more similar blocks of flexible genes within respective lineages (Figure 3). Given that their flexible genome (cloud) was enriched with genes involved in respiratory chains, regulation of responses to the environment as well as transport of substrates and their metabolic processing, it is likely that one fraction of them might be advantageous and involved in niche adaptation (Rodriguez-Valera and Ussery, 2012; Shapiro et al., 2012). Moreover, the functional differences of flexible genes may contribute to the better substrate utilization of one lineage over another (Supplementary Table S6). For instance, the D3 lineage may be better adapted for the uptake of carbohydrates and limiting micronutrients (e.g., iron), and the D2 may synthesize organic compounds that would improve its response to various environmental stimuli (Rodriguez-Valera et al., 2016). However, there is still a high percentage of genes with unknown functions in the *Laspinema* pangenome (76.9%), which might have arisen due to the ecological divergence between the lineages.

Our study aligns with previous observations that conservative markers are insufficient to capture the true extent of microbial intraspecies diversity (Richter and Rosselló-Móra, 2009; Zhaxybayeva et al., 2009). Moreover, applying various population genomics approaches to characterize new species offers a better understanding of the fine microbial diversity than using different thresholds of sequence similarities.

High genomic divergence and recombination suppression of light and stress adaptation genes in D2 and D3

Genome-wide F_{ST} and absolute divergence (D_{XY}) that exceeded nucleotide divergence both suggest high genetic differentiation between D2 and D3 lineages. The majority of the genome exhibits high values of F_{ST} (mean F_{ST} of 0.69; Figure 4), which could indicate the presence of diversifying selection among many loci scattered across the genome. Additionally, it highlights low gene transfer between the lineages. We investigated potential signatures of the selection in genomic segments with a high level of divergence (99th percentile of F_{ST} values), low nucleotide diversity, and suppressed recombination. Tentative adaptive genes that were localized in these regions regulate physiological processes associated with the exposure to high/low light conditions, sugar and protein metabolism, and the response to various environmental stresses, e.g., nitrogen, phosphate, or magnesium deprivation (Supplementary Table S7). Experimental and genomic studies of closely related marine cyanobacteria *Prochlorococcus* and *Synechococcus* identified some genes that could be associated with adaptations to ecological niches (Rocap et al., 2003; Stuart et al., 2013; Kashtan et al., 2014). Detected genes were involved in, e.g., nutrient uptake (nitrogen, phosphorus, iron), light acclimation, regulatory functions, and response to various environmental stimuli. Similar patterns were discovered in other microbes like *Vibrio* (e.g., stress-response genes; Shapiro et al., 2012) and *Curtobacterium* (carbohydrases; Chase et al., 2019).

We detected a weak signal of diversifying selection on potential adaptive genes in *Laspinema* using the MK test (Supplementary Table S8), which suggests that loci evolve neutrally. Consistent with observations in *Sulfolobus* (Cadillo-Quiroz et al., 2012), maintenance of multiple adaptive alleles might be beneficial for the coexistence of *Laspinema* lineages and the introduction of small differences in their fitness. These fine differences are linked to a wide range of niches that can be really small in soil ecosystems – microniches (Vos et al., 2013). *L. thermale* is well-adapted to diverse soil habitats - from radioactive thermal springs (Heidari et al., 2018) with extreme temperatures and contamination to puddles, which frequently undergo substantial environmental changes (e.g., drying, wetting, nutrient limitation, exposure to UV light). Genes regulating phototaxis (cheW, CheY, cheB) and response to various environmental stimuli (rcsC, rssB) exhibited a strong interspecific structuring, i.e., they were different genes associated with the same function (Supplementary Table S8). This could indicate the separation of ecological niches between D2 and D3, which might have initiated the process of ecological differentiation in *Laspinema* lineages (Shapiro et al., 2012). Hence, one lineage may grow better in unfavorable conditions or burrow deeper in the soil due to the excess light or lack of moisture (Mager and Thomas, 2011). Nevertheless, the functions of niche-specific genes require future physiological investigations to confirm whether these genes are directly involved in adaptive processes.

Recent speciation of cryptic *Laspinema* sp. lineages

Many microbial species are in different stages of speciation, which has been proposed to be driven by the adaptation to various ecological niches (Polz et al., 2013; Shapiro and Polz, 2015). *Laspinema* exhibits the overall genomic pattern of high genetic divergence and neutral evolution in highly differentiated genomic regions. Besides, a trend of decreased HR and high LD suggests the clonal nature of our *Laspinema* strains. The observed recombination rates concur with previous estimates for some cyanobacteria (e.g., *Microcoleus*) and free-living prokaryotes (Vos and Didelot, 2009; González-Torres et al., 2019), suggesting that *Laspinema* strains engage in little recombination compared to mutation (Table 2). Following the *r/m* ratio, which is the relative impact of recombination and mutation on lineage diversification, population-specific recombination is stronger in the D3 ($r/m = 1.368$) than in the D2 lineage ($r/m = 0.822$). In spite of occurring less frequently than mutation (D3$_{\rho/\theta} = 0.023$), recombination events introduced almost 1.5 as many substitutions as mutations. That highlights the importance of HR over mutation in shaping the genetic diversity of the D3 lineage. However, lower recombination rates among D2 strains could be an artifact of fewer individuals in that clade. Such a pattern of reduced gene flow between lineages ($r/m = 0.608$) indicates the existence of a recombination barrier which may contribute to the maintenance of strains in cohesive genetic groups and a further spread of tentative adaptive genes for niche specialization (Fraser et al., 2007).

Given all of the patterns of genetic differentiation among D2 and D3, following Shapiro and Polz's (2015) bacterial speciation model, we estimate that *Laspinema* lineages might be at speciation stage 3 or 4. Although we detected weak diversifying selection in tentative adaptive genes, established genomic isolation with reduced gene flow between lineages could suggest that *Laspinema* is in ongoing ecological differentiation, which may reach completion.

Notably, a low number of individuals in both lineages might have impacted our observations. Thus, we avoided analyses requiring larger population sampling and pre-defined populations. Another limitation is the lack of population genetic data for isolated strain D2d, which is likely a separate species. However, due to the random sampling and the high diversity of terrestrial microbial communities, finding strains closely related to that specific one would be extremely challenging. Forthcoming physiological and modeling studies are crucial for testing the adaptive differences between two lineages and discovering genes of unknown function, which harbor significant evidence for unexplored ecological niches in soil.

Overall, we provide a deeper insight into the genetic diversity that underlies the divergence of cyanobacterium *Laspinema*. We have highlighted the importance of using the population genetics approach over traditional markers in studying cyanobacterial diversity. We identified two cryptic genetically differentiated *Laspinema* lineages that could represent new cyanobacterial species. A suite of potential adaptive alleles associated with specialization to niches in the soil already emerged in the two lineages coexisting in a sympatric setting, although with a weak signature of diversifying selection. Lineage divergence might be driven by genomic and ecological processes and is further maintained by the limited gene flow between them. The origin of barriers to gene flow among *Laspinema* strains remains to be investigated.

Data availability statement

All 16S rRNA and 16S–23S ITS sequences have been deposited in GenBank under accession numbers ON814838-ON814846 and ON814847-ON814854, respectively. Biosample identification and accession numbers are available in Table S4. Relevant codes used in the analyses are available at https://github.com/AleksandarStan/Laspinema. The vcf file and multiple sequence alignments (for whole-genome and 16S rRNA phylogeny and extracted from the vcf file) can be found on figshare (https://doi.org/10.6084/m9.figshare.20116118).

Author contributions

AS, SS, and PD designed the research and analyzed the data. AS wrote the paper and conducted the research. All authors contributed to the article and approved the submitted version.

Funding

Conflict of interest

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.977454/full#supplementary-material

References

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19, 455–477. doi: 10.1089/cmb.2012.0021

Barraclough, T. G., Balbi, K. J., and Ellis, R. J. (2012). Evolving concepts of bacterial species. *Evol. Biol.* 39, 148–157. doi: 10.1007/s11692-012-9181-8

Bobay, L. M. (2020). "The prokaryotic species concept and challenges," in *The Pangenome*. eds. H. Tettelin and D. Medini (Cham: Springer), 21–49.

Bobay, L. M., and Ochman, H. (2017). Biological species are universal across life's domains. *Genome Biol. Evol.* 9, 491–501. doi: 10.1093/gbe/evx026

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Cadillo-Quiroz, H., Didelot, X., Held, N. L., Herrera, A., Darling, A., Reno, M. L., et al. (2012). Patterns of gene flow define species of thermophilic archaea. *PLoS Biol.* 10:e1001265. doi: 10.1371/journal.pbio.1001265

Carrolo, M., Pinto, F. R., Melo-Cristino, J., and Ramirez, M. (2009). Pherotypes are driving genetic differentiation within *Streptococcus pneumoniae. BMC Microbiol.* 9, 1–10. doi: 10.1186/1471-2180-9-191

Casamatta, D. A., Vis, M. L., and Sheath, R. G. (2003). Cryptic species in cyanobacterial systematics: a case study of *Phormidium retzii* (Oscillatoriales) using

RAPD molecular markers and 16S rDNA sequence data. *Aquat. Bot.* 77, 295–309. doi: 10.1016/j.aquabot.2003.08.005

Chaguza, C., Andam, C. P., Harris, S. R., Cornick, J. E., Yang, M., Bricio-Moreno, L., et al. (2016). Recombination in *Streptococcus pneumoniae* lineages increase with carriage duration and size of the polysaccharide capsule. *MBio* 7, e01053–e01016. doi: 10.1128/mBio.01053-16

Chase, A. B., Arevalo, P., Brodie, E. L., Polz, M. F., Karaoz, U., and Martiny, J. B. (2019). Maintenance of sympatric and allopatric populations in free-living terrestrial bacteria. *MBio* 10, e02361–e02319. doi: 10.1128/mBio.02361-19

Coleman, M. L., and Chisholm, S. W. (2010). Ecosystem-specific selection pressures revealed through comparative population genomics. *Proc. Natl. Acad. Sci. U. S. A.* 107, 18634–18639. doi: 10.1073/pnas.1009480107

Cordero, O. X., Ventouras, L. A., DeLong, E. F., and Polz, M. F. (2012). Public good dynamics drive evolution of iron acquisition strategies in natural bacterioplankton populations. *Proc. Natl. Acad. Sci. U. S. A.* 109, 20059–20064. doi: 10.1073/pnas.1213344109

Croucher, N. J., Page, A. J., Connor, T. R., Delaney, A. J., Keane, J. A., Bentley, S. D., et al. (2015). Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* 43:e15. doi: 10.1093/nar/gku1196

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330

Duval, C., Hamlaoui, S., Piquet, B., Toutirais, G., Yepremian, C., Reinhart, A., et al. (2020). Characterization of cyanobacteria isolated from thermal muds of Balaruc-les-Bains (France) and description of a new genus and species *Pseudochroococcus couteii*. bioRxiv. doi: 10.1101/2020.12.12.422513

Dvořák, P., Casamatta, D. A., Hašler, P., Jahodářová, E., Norwich, A. R., and Poulíčková, A. (2017). "Diversity of the cyanobacteria," in *Modern Topics in the Phototrophic Prokaryotes*. ed. P. C. Hallenbeck (Cham: Springer), 3–46.

Dvořák, P., Casamatta, D. A., Poulíčková, A., Hašler, P., Ondřej, V., and Sanges, R. (2014). *Synechococcus*: 3 billion years of global dominance. *Mol. Ecol.* 23, 5538–5551. doi: 10.1111/mec.12948

Dvořák, P., Poulíčková, A., Hašler, P., Belli, M., Casamatta, D. A., and Papini, A. (2015). Species concepts and speciation factors in cyanobacteria, with connection to the problems of diversity and classification. *Biodivers. Conserv.* 24, 739–757. doi: 10.1007/s10531-015-0888-6

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Emms, D. M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20, 238–214. doi: 10.1186/s13059-019-1832-y

Felsenstein, J. (2005). *PHYLIP: Phylogeny Inference Package, Version 3.6.* Seattle, WA: University of Washington.

Fraser, C., Hanage, W. P., and Spratt, B. G. (2007). Recombination and the nature of bacterial speciation. *Science* 315, 476–480. doi: 10.1126/science.1127573

Galtier, N., and Daubin, V. (2008). Dealing with incongruence in phylogenomic analyses. *Philos. Trans. R. Soc. B Biol. Sci.* 363, 4023–4029. doi: 10.1098/rstb.2008.0144

González-Torres, P., Rodríguez-Mateos, F., Antón, J., and Gabaldón, T. (2019). Impact of homologous recombination on the evolution of prokaryotic core genomes. *MBio* 10, e02494–e02418. doi: 10.1128/mBio.02494-18

Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P., and Tiedje, J. M. (2007). DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *Int. J. Syst. Evol. Microbiol.* 57, 81–91. doi: 10.1099/ijs.0.64483-0

Grundmann, G. L. (2004). Spatial scales of soil bacterial diversity – the size of a clone. *FEMS Microbiol. Ecol.* 48, 119–127. doi: 10.1016/j.femsec.2004.01.010

Hadfield, J., Croucher, N. J., Goater, R. J., Abudahab, K., Aanensen, D. M., and Harris, S. R. (2018). Phandango: an interactive viewer for bacterial population genomics. *Bioinformatics* 34, 292–293. doi: 10.1093/bioinformatics/btx610

Harris, H. M. B., Bourin, M. J. B., Claesson, M. J., and O'Toole, P. W. (2017). Phylogenomics and comparative genomics of *lactobacillus salivarius*, a mammalian gut commensal. *Microb. Genom.* 3:e000115. doi: 10.1099/mgen.0.000115

Hašler, P., Dvořák, P., Johansen, J. R., Kitner, M., Ondřej, V., and Poulíčková, A. (2012). Morphological and molecular study of epipelic filamentous genera *Phormidium, microcoleus* and *Geitlerinema* (Oscillatoriales, Cyanophyta/cyanobacteria). *Fottea* 12, 341–356. doi: 10.5507/fot.2012.024

Heidari, F., Hauer, T., Zima, J. R. H., and Riahi, H. (2018). New simple trichal cyanobacterial taxa isolated from radioactive thermal springs. *Fottea* 18, 137–149. doi: 10.5507/fot.2017.024

Held, N. L., Herrera, A., Cadillo-Quiroz, H., and Whitaker, R. J. (2010). CRISPR associated diversity within a population of *Sulfolobus islandicus*. *PLoS One* 5:e12988. doi: 10.1371/journal.pone.0012988

Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q., and Vinh, L. S. (2018). UFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* 35, 518–522. doi: 10.1093/molbev/msx281

Hunt, D. E., David, L. A., Gevers, D., Preheim, S. P., Alm, E. J., and Polz, M. F. (2008). Resource partitioning and sympatric differentiation among closely related bacterioplankton. *Science* 320, 1081–1085. doi: 10.1126/science.1157890

Huson, D. H., and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* 23, 254–267. doi: 10.1093/molbev/msj030

Jaspers, E., and Overmann, J. (2004). Ecological significance of microdiversity: identical 16S rRNA gene sequences can be found in bacteria with highly divergent genomes and ecophysiologies. *Appl. Environ. Microbiol.* 70, 4831–4839. doi: 10.1128/AEM.70.8.4831-4839.2004

Jeltsch, A. (2003). Maintenance of species identity and controlling speciation of bacteria: a new function for restriction/modification systems? *Gene* 317, 13–16. doi: 10.1016/S0378-1119(03)00652-8

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14, 587–589. doi: 10.1038/nmeth.4285

Kashtan, N., Roggensack, S. E., Rodrigue, S., Thompson, J. W., Biller, S. J., Coe, A., et al. (2014). Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. *Science* 344, 416–420. doi: 10.1126/science.1248575

Kollár, J., Poulíčková, A., and Dvořák, P. (2022). On the relativity of species, or the probabilistic solution to the species problem. *Mol. Ecol.* 31, 411–418. doi: 10.1111/mec.16218

Komárek, J., and Anagnostidis, K. (2005). "Cyanoprokaryota 2. Teil: oscillatoriales," in *Süsswasserflora von Mitteleuropa*. eds. B. Büdel, G. Gärdner, L. Krienitz and M. Schagerl (München: Elsevier), 759.

Koonin, E. V., Makarova, K. S., and Wolf, Y. I. (2021). Evolution of microbial genomics: conceptual shifts over a quarter century. *Trends Microbiol.* 29, 582–592. doi: 10.1016/j.tim.2021.01.005

Korber, B. (2000). "HIV signature and sequence variation analysis," in *Computational Analysis of HIV Molecular Sequences*. eds. A. G. Rodrigo and G. H. Learn (Dordrecht: Kluwer Academic Publishers), 55–72.

Larsson, A. (2014). AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* 30, 3276–3278. doi: 10.1093/bioinformatics/btu531

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Mager, D. M., and Thomas, A. D. (2011). Extracellular polysaccharides from cyanobacterial soil crusts: a review of their role in dryland soil processes. *J. Arid Environ.* 75, 91–97. doi: 10.1016/j.jaridenv.2010.10.001

Martin, S. H., Dasmahapatra, K. K., Nadeau, N. J., Salazar, C., Walters, J. R., Simpson, F., et al. (2013). Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Res.* 23, 1817–1828. doi: 10.1101/gr.159426.113

Martin, D. P., Lemey, P., and Posada, D. (2011). Analysing recombination in nucleotide sequences. *Mol. Ecol. Resour.* 11, 943–955. doi: 10.1111/j.1755-0998.2011.03026.x

Martinez-Urtaza, J., Van Aerle, R., Abanto, M., Haendiges, J., Myers, R. A., Trinanes, J., et al. (2017). Genomic variation and evolution of *Vibrio parahaemolyticus* ST36 over the course of a transcontinental epidemic expansion. *MBio* 8, e01425–e01417. doi: 10.1128/mBio.01425-17

Mayr, E. (1942). *Systematics and the Origin of Species.* New York, NY: Columbia University Press.

McInerney, J. O., McNally, A., and O'connell, M. J. (2017). Why prokaryotes have pangenomes. *Nat. Microbiol.* 2, 1–5. doi: 10.1038/nmicrobiol.2017.40

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110

Mirarab, S., and Warnow, T. (2015). ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* 31, i44–i52. doi: 10.1093/bioinformatics/btv234

Nguyen, L. T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., et al. (2015). Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 31, 3691–3693. doi: 10.1093/bioinformatics/btv421

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., and Tyson, G. W. (2015). Check M: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* 25, 1043–1055. doi: 10.1101/gr.186072.114

Pérez-Carrascal, O. M., Terrat, Y., Giani, A., Fortin, N., Greer, C. W., Tromas, N., et al. (2019). Coherence of *Microcystis* species revealed through population genomics. *ISME J.* 13, 2887–2900. doi: 10.1038/s41396-019-0481-1

Pfeifer, B., Wittelsbürger, U., Ramos-Onsins, S. E., and Lercher, M. J. (2014). PopGenome: an efficient Swiss army knife for population genomic analyses in R. *Mol. Biol. Evol.* 31, 1929–1936. doi: 10.1093/molbev/msu136

Pietrasiak, N., Mühlsteinová, R., Siegesmund, M. A., and Johansen, J. R. (2014). Phylogenetic placement of *Symplocastrum* (Phormidiaceae, Cyanophyceae) with a new combination *S. californicum* and two new species: *S. flechtnerae* and *S. torsivum.* *Phycologia* 53, 529–541. doi: 10.2216/14-029.1

Pietrasiak, N., Osorio-Santos, K., Shalygin, S., Martin, M. P., and Johansen, J. R. (2019). When is a lineage a species? A case study in *Myxacorys* gen. Nov. (Synechococcales: cyanobacteria) with the description of two new species from the Americas. *J. Phycol.* 55, 976–996. doi: 10.1111/jpy.12897

Polz, M. F., Alm, E. J., and Hanage, W. P. (2013). Horizontal gene transfer and the evolution of bacterial and archaeal population structure. *Trends Genet.* 29, 170–175. doi: 10.1016/j.tig.2012.12.006

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

R Core Team (2021). *R: A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing.

Rambaut, A. (2012). FigTree, version 1.4.4, pp. Computer Program Distributed by the Author. Available at: http://tree.bio.ed.ac.uk/software/figtree/ (Accessed January 10, 2022).

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., et al. (2019). G:profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* 47, W191–W198. doi: 10.1093/nar/gkz369

Reno, M. L., Held, N. L., Fields, C. J., Burke, P. V., and Whitaker, R. J. (2009). Biogeography of the *Sulfolobus islandicus* pan-genome. *Proc. Natl. Acad. Sci. U. S. A.* 106, 8605–8610. doi: 10.1073/pnas.0808945106

Richter, M., and Rosselló-Móra, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. *Proc. Natl. Acad. Sci. U. S. A.* 106, 19126–19131. doi: 10.1073/pnas.0906412106

Rocap, G., Larimer, F. W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N. A., et al. (2003). Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* 424, 1042–1047. doi: 10.1038/nature01947

Rodriguez-Valera, F., Martin-Cuadrado, A. B., and López-Pérez, M. (2016). Flexible genomic islands as drivers of genome evolution. *Curr. Opin. Microbiol.* 31, 154–160. doi: 10.1016/j.mib.2016.03.014

Rodriguez-Valera, F., and Ussery, D. W. (2012). Is the pan-genome also a pan-selectome? *F1000Research* 1:16. doi: 10.12688/f1000research.1-16.v1

Roux, C., Fraisse, C., Romiguier, J., Anciaux, Y., Galtier, N., and Bierne, N. (2016). Shedding light on the grey zone of speciation along a continuum of genomic divergence. *PLoS Biol.* 14:e2000234. doi: 10.1371/journal.pbio.2000234

Rozas, J., Ferrer-Mata, A., Sánchez-DelBarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S. E., et al. (2017). DnaSP 6: DNA sequence polymorphism analysis of large datasets. *Mol. Biol. Evol.* 34, 3299–3302. doi: 10.1093/molbev/msx248

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068–2069. doi: 10.1093/bioinformatics/btu153

Shapiro, B. J., Friedman, J., Cordero, O. X., Preheim, S. P., Timberlake, S. C., Szabó, G., et al. (2012). Population genomics of early events in the ecological differentiation of bacteria. *Science* 336, 48–51. doi: 10.1126/science.1218198

Shapiro, B. J., and Polz, M. F. (2014). Ordering microbial diversity into ecologically and genetically cohesive units. *Trends Microbiol.* 22, 235–247. doi: 10.1016/j.tim.2014.02.006

Shapiro, B. J., and Polz, M. F. (2015). Microbial speciation. *Cold Spring Harb. Perspect. Biol.* 7:a018143. doi: 10.1101/cshperspect.a018143

Sheinman, M., Arkhipova, K., Arndt, P. F., Dutilh, B. E., Hermsen, R., and Massip, F. (2021). Identical sequences found in distant genomes reveal frequent horizontal transfer across the bacterial domain. *elife* 10:e62719. doi: 10.7554/eLife.62719

Simmons, S. L., DiBartolo, G., Denef, V. J., Goltsman, D. S. A., Thelen, M. P., and Banfield, J. F. (2008). Population genomic analysis of strain variation in *Leptospirillum* group II bacteria involved in acid mine drainage formation. *PLoS Biol.* 6:e177. doi: 10.1371/journal.pbio.0060177

Stanojković, A., Skoupý, S., Hašler, P., Poulíčková, A., and Dvořák, P. (2022). Geography and climate drive the distribution and diversification of the cosmopolitan cyanobacterium *microcoleus* (Oscillatoriales, cyanobacteria). *Eur. J. Phycol.* 1–10, 1–10. doi: 10.1080/09670262.2021.2007420

Staub, R. (1961). Ernährungsphysiologisch-autökologische Untersuchungen an der planktischen Blaualge *Oscillatoria rubescens* DC. *Schweiz. Z. Hydrol.* 23, 82–198.

Stuart, R. K., Brahamsha, B., Busby, K., and Palenik, B. (2013). Genomic island genes in a coastal marine *Synechococcus* strain confer enhanced tolerance to copper and oxidative stress. *ISME J.* 7, 1139–1149. doi: 10.1038/ismej.2012.175

Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* 34, W609–W612. doi: 10.1093/nar/gkl315

Tamura, K., Stecher, G., and Kumar, S. (2021). MEGA11: molecular evolutionary genetics analysis version 11. *Mol. Biol. Evol.* 38, 3022–3027. doi: 10.1093/molbev/msab120

Törönen, P., Medlar, A., and Holm, L. (2018). PANNZER2: a rapid functional annotation web server. *Nucleic Acids Res.* 46, W84–W88. doi: 10.1093/nar/gky350

Vos, M., and Didelot, X. (2009). A comparison of homologous recombination rates in bacteria and archaea. *ISME J.* 3, 199–208. doi: 10.1038/ismej.2008.93

Vos, M., Wolf, A. B., Jennings, S. J., and Kowalchuk, G. A. (2013). Micro-scale determinants of bacterial diversity in soil. *FEMS Microbiol. Rev.* 37, 936–954. doi: 10.1111/1574-6976.12023

Walk, S. T. (2015). The "cryptic" *Escherichia. EcoSal Plus* 6. doi: 10.1128/ecosalplus.ESP-0002-2015

Ward, D. M., Cohan, F. M., Bhaya, D., Heidelberg, J. F., Kühl, M., and Grossman, A. (2008). Genomics, environmental genomics and the issue of microbial species. *Heredity* 100, 207–219. doi: 10.1038/sj.hdy.6801011

Warnes, M. G. R., Bolker, B., Bonebakker, L., Gentleman, R., and Huber, W. (2016). Package 'gplots'. *Various R Programming Tools for Plotting data.*

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis.* New York: Springer-Verlag.

Wiedenbeck, J., and Cohan, F. M. (2011). Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. *FEMS Microbiol. Rev.* 35, 957–976. doi: 10.1111/j.1574-6976.2011.00292.x

Wielgoss, S., Didelot, X., Chaudhuri, R. R., Liu, X., Weedall, G. D., Velicer, G. J., et al. (2016). A barrier to homologous recombination between sympatric strains of the cooperative soil bacterium *Myxococcus xanthus. ISME J.* 10, 2468–2477. doi: 10.1038/ismej.2016.34

Wu, Y. W., Simmons, B. A., and Singer, S. W. (2016). Max bin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* 32, 605–607. doi: 10.1093/bioinformatics/btv638

Yarza, P., Yilmaz, P., Pruesse, E., Glöckner, F. O., Ludwig, W., Schleifer, K. H., et al. (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat. Rev. Microbiol.* 12, 635–645. doi: 10.1038/nrmicro3330

Zamani, N., Russell, P., Lantz, H., Hoeppner, M. P., Meadows, J. R., Vijay, N., et al. (2013). Unsupervised genome-wide recognition of local relationship patterns. *BMC Genomics* 14, 1–11. doi: 10.1186/1471-2164-14-347

Zamudio, R., Haigh, R. D., Ralph, J. D., De Ste Croix, M., Tasara, T., Zurfluh, K., et al. (2020). Lineage-specific evolution and gene flow in *listeria monocytogenes* are independent of bacteriophages. *Environ. Microbiol.* 22, 5058–5072. doi: 10.1111/1462-2920.15111

Zhang, C., Rabiee, M., Sayyari, E., and Mirarab, S. (2018). ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinform.* 19, 153. doi: 10.1186/s12859-018-2129-y

Zhaxybayeva, O., Doolittle, W. F., Papke, R. T., and Gogarten, J. P. (2009). Intertwined evolutionary histories of marine *Synechococcus* and *Prochlorococcus marinus. Genome Biol. Evol.* 1, 325–339. doi: 10.1093/gbe/evp032

Zimba, P. V., Shalygin, S., Huang, I. S., Momčilović, M., and Abdulla, H. (2020). A new boring toxin producer – Perforafilum tunnelli gen. & sp. nov.(Oscillatoriales, Cyanobacteria) isolated from Laguna Madre, Texas, USA. *Phycologia* 60, 10–24. doi: 10.1080/00318884.2020.1808389

Paper III

The global speciation continuum of cyanobacterium *Microcoleus*

Aleksandar Stanojković, Svatopluk Skoupý & Petr Dvořák*

Palacký University Olomouc, Faculty of Sciences, Department of Botany, Šlechtitelů 27, 78371 Olomouc, Czech Republic

Running title: The global speciation continuum

Correspondence: Petr Dvořák**,** Palacký University Olomouc, Faculty of Sciences, Department of Botany, Šlechtitelů 27, 78371 Olomouc, Czech Republic; e-mail: p.dvorak@upol.cz

Abstract:

Speciation is a continuous process driven by genetic, geographic, and ecological barriers to gene flow. Yet, we are only beginning to comprehend the relative importance of mechanisms driving them at different points of microbial population differentiation. We examined the divergence patterns of 201 *Microcoleus vaginatus* strains and explored the diversification patterns of 291 *Microcoleus* genomes (8 from herbarium specimens and 82 genomes from the GenBank), shedding light on the emergence of a global continuum of at least 13 cyanobacterial species over time. We find that species exhibit varying degrees of divergence and gene flow, reflecting their relative positions along the speciation continuum. Genetic divergence and selection are widespread across the genome, which act in synergy with geography and environmental factors to drive species differentiation in *Microcoleus*. We show species diversification started 29.6 million years ago during Eocene/Oligocene aridification with significant shifts in population sizes during the Pleistocene climate oscillations. Our genome scan analyses revealed putative genes associated with stress response and biosynthetic processes in the regions of exceptional differentiation, suggesting their importance in cyanobacterial adaptations to dominate heterogeneous soil systems. These results provide new insights into the notion of microbial species and argue that the emergence of a global continuum of *Microcoleus* species is a complex process driven by the coaction of recombination, selection, geography, and environmental forces acting at different stages of speciation.

1. Introduction

A complex interplay of evolutionary forces generates and governs the differentiation within and among species leading to species emergence (Shapiro, 2018; Koonin et al., 2021). The population and comparative genomics breakthrough have greatly facilitated disentangling evolutionary forces shaping the speciation and key genes involved in environmental adaptation (Nosil & Feder, 2012). However, understanding of evolutionary processes driving the differentiation and the emergence of new species in prokaryotes is still hindered by the ambiguous species concept for microbes (Roselló-Mora & Amann, 2001), their extensive genetic and phenotypic diversity (Whitaker & Banfield, 2006) and highly heterogeneous ecological niches that they can occupy (Vos et al., 2013).

The debate on delineating microbial species and which species concept to apply is still contentious, as traditional morphological and reproductive criteria often do not apply to prokaryotes (Achtman & Wagner, 2008). Despite their prevalent asexual nature, microbes exhibit high levels of genetic diversity, and gene flow acts at varying frequencies within and well beyond species boundaries, thus blurring them (Bobay & Ochman, 2017; Fraser et al., 2007). The operational species delimitation based on ANI value (average nucleotide identity, i.e., genome similarity) was introduced as the solution for the fuzzy species boundaries (Konstantinidis & Tiedje, 2005). However, the single threshold may not capture variable speciation rates and the whole complexity of the speciation among the bacterial lineages (Shapiro, 2018). The need for a more nuanced framework in discerning microbial species is further exacerbated by the high cryptic diversity, in which multiple distinct lineages might be hidden within a single phenotype (Dvořák et al., 2015). Nevertheless, several studies have already shown that evolutionary forces contributing to the divergence, like genetic and ecological isolation as well as the selection, could aid us in delineating closely related microbes (*Vibrio*; Shapiro et al., 2012; *Sulfolobus*, Cadillo-Quiroz et al., 2012; *Microcystis*, Pérez-Carrascal et al., 2019; *Laspinema*, Stanojković et al., 2022b).

Mutations and gene flow mediated by homology-(in)dependent mechanisms are motors introducing genetic changes to microbial populations, and the interaction between selection, recombination, and genetic drift dictate the magnitude of differentiation among them (Reno et al., 2009; Shapiro et al., 2012). Depending on the coaction of these forces, species can be at various stages of speciation across space and time, from undergoing continuous gene flow to its complete

cessation (Shapiro & Polz, 2015; Roux et al., 2016; Kollár et al., 2022). The former perspective on speciation centering on the extent of gene flow between species, which varies along a continuum, has been renowned (Coyne & Orr, 1989), and it challenges the traditional view of speciation as a discrete and irreversible event (Coyne & Orr, 2004; Nosil et al., 2017). An increasing number of genomic studies extended this idea by focusing on genetic differentiation between closely related incipient species to characterize them at various stages of genetic and ecological divergence along the continuum in eukaryotes (*Heliconius* butterflies, Martin et al., 2013; *Anopheles* mosquitoes, Turner & Hahn, 2010; *Helianthus* sunflowers, Renaut et al., 2014) and prokaryotes (*Vibrio*, Shapiro et al., 2012; *Laspinema*, Stanojković et al., 2022b). Although the concept of speciation as a continuum provides a pragmatic, unifying framework for analyzing speciation (Stankowski & Ravinet, 2021), there is a notable inclination toward eukaryotes, with limited attention given to a highly diverse group of prokaryotes, particularly cyanobacteria.

In the early stages of speciation, microbial populations may exhibit subtle genomic hallmarks of elevated differentiation localized in small genomic regions under strong divergent selection and impervious to recombination, often referred to as 'islands of speciation' (Wu, 2001; Feder et al., 2012; Shapiro & Polz, 2015). The emergence of these regions can be propelled by adaptation to various environmental factors such as light and stress (e.g., *Prochlorococcus*, Kashtan et al., 2014; Coleman et al., 2006; *Laspinema*, Stanojković et al., 2022b) or nutrient uptake (e.g., *Vibrio*, Shapiro et al., 2012). As speciation progresses, genomic landscapes of divergence may become more distinct and fixed, establishing genome-wide barriers and, ultimately, genetically and ecologically distinct species (Nosil, 2012). Widespread genetic differentiation was reported in prokaryotes *Sulfolobus* (Cadillo-Quiroz et al., 2012) and *Laspinema* (Stanojković et al., 2022b) as well as in eukaryotes like mosquitoes (Lawniczak et al., 2010) and fruit flies (Egan et al., 2015). Whether the speciation initiates from 'islands' or arises across the whole genome from the beginning remains controversial, with authors reporting contrasting patterns among different taxa (Pennisi, 2014). Hence, understanding the genomic architecture of species at different speciation stages and mechanisms driving the adaptation and barriers to gene flow are crucial for elucidating the evolutionary dynamics of microbes in response to changing environments.

Cyanobacteria have been dominating terrestrial ecosystems as primary producers for billions of years. Today, the most prominent and widely distributed is *Microcoleus*, mat-forming filamentous

3

cyanobacteria inhabiting the most diverse aerophytic and benthic habitats (Garcia-Pichel & Wojciechowski, 2009; Hašler et al., 2012). Particularly successful and abundant cyanobacterium is *Microcoleus vaginatus,* occurring in biological soil crusts of arid and semiarid ecosystems, which cover up to 40% of the land surface on Earth (Belnap & Lange, 2001). The Cenozoic rise to dominance of *M. vaginatus* 39.5 Ma years ago (Dvořák et al., 2012) undoubtedly affected dry ecosystems and was central to the evolution of biocrust communities (Couradeau et al., 2019). Thriving under unique ecosystem conditions (e.g., water scarcity, high climatic variability, low carbon storage capability), *M. vaginatus* plays a key role by being a pivot of soil stability, primary productivity, and carbon influx (Belnap & Gardner, 1993). Moreover, *Microcoleus*-dominated crusts improve soil fertility and moisture retention, molding microhabitats for other microbial and plant communities (Nelson et al., 2021). More than 3500 studies focused on *M. vaginatus* species, but it is actually just one of the many species along the entire continuum of species at varying stages of speciation (we queried the Google Scholar database; accessed 25.04.2023; the term used: *Microcoleus vaginatus*). In our previous study, using a global dataset of almost 500 strains based on 16S rRNA and 16S-23S ITS markers, we noticed that *M. vaginatus* embodies vast genetic diversity and consists of at least 12 potential species, which diversified due to the influence of geographical and ecological separation (Stanojković et al., 2022a).

Here, we sought to investigate the continuum of *Microcoleus* populations varying in differentiation, i.e., the speciation continuum, and possible genetic determinants connected with the capability of *Microcoleus* to dominate dryland ecosystems worldwide. We sequenced 201 whole genomes of multiple closely related strains in each of the 12 lineages. We additionally sequenced eight whole genome sequences of *Microcoleus* herbarium specimens spanning from 1851 and 1938 and added them to our dataset, enhancing its diversity and exclusivity. We examined the genetic population structure and whether geography and environment affect the differentiation within *Microcoleus*. We next explored genome-wide inter- and intraspecific diversity to provide a better understanding of the potential cause of the divergence within *Microcoleus*. Finally, we investigate the signatures of local selection and gene flow over the whole genome during adaptive divergence. Comprehending microbial speciation provides the ecological context for the present-day dominance of cyanobacteria in drylands and may aid in disentangling adaptive evolutionary mechanisms that contributed to the global success of *Microcoleus*.

2. Material and methods

2.1. Sample collection, whole-genome sequencing, and genome assemblies

Sample collection and culture conditions have been described in Stanojković et al. (2022a). From 57 environmental samples, we selected one to ten strains, and altogether, the genomic DNA of 202 isolates was extracted. Approximately 100mg of fresh biomass was used, and genomic DNA was isolated with UltraClean Microbial DNA Isolation Kit (MOBIO, Carlsbad, CA, USA). Nextera XT DNA Library Preparation Kit (Illumina San Diego, CA, USA) was used for marking genome sequence libraries. Isolates were commercially sequenced on Illumina NovaSeq 6000 platform (Novogene, United Kingdom) at the 150 bp x 2 paired-end mode. Adaptors and low-quality reads were filtered and trimmed using Trimmomatic v0.39 (Bolger et al., 2014), and quality reads were assembled into genomes using SPAdes v3.13.1 (Bankevich et al., 2012). Genome contamination was removed using binning of the resulting scaffolds with MaxBin v2.2.4 (Wu et al., 2016), where one bin represented one genome. Genome completion and contamination were assessed via CheckM v1.1.9 (Parks et al., 2015).

Eight *Microcoleus vaginatus* samples collected from North America (3) and Europe (5) between 1851 and 1938 were obtained from the herbarium of the Natural History Museum (London, United Kingdom). Samples were extracted following the protocol of Kistler (2012) and placed in a sterile Eppendorf tube for the subsequent analyses. Following the established protocol by Meyer & Kircher (2010), DNA libraries for sequencing were prepared using NEBNext DNA Sample Prep Master Mix Set 2 and Illumina-specific adapters. The libraries were sequenced on an Illumina NovaSeq 6000 instrument (Novogene, United Kingdom) to generate 150 bp paired-end reads per sample. Filtering the reads, assembly, binning, and quality assessment was carried out as previously described.

Altogether, we constructed three datasets for different analyses. The dataset I included 202 sequenced genomes in this study was used for phylogenomics (with the outgroup, strain M2_D5). The outgroup was omitted for variant calling, delimitation, pangenome, recombination, and population genetic analyses. Dataset II included sequenced genomes from dataset I with additional eight herbarium specimens sequenced in this study and 81 *Microcoleus* genomes obtained from the GenBank database (accession numbers in Supplementary Table 1); wholly, 291 genomes originating from all continents except for South America (Figure 1A). This dataset was used for

ancestral area reconstruction, phylogenomics, phylogenetic signal, and Mantels' correlation tests. The final dataset III encompassed 208 genomes, out of which 42 were selected from dataset I and 166 representatives from different cyanobacterial taxa acquired from the GenBank (accession numbers in Supplementary Table 1). Dataset III was used to investigate the phylogenetic positions of *Microcoleus* genomes sequenced here and other cyanobacterial taxa.

All the genome sequences were deposited in the GenBank database under the BioProject #.

2.2. Genome annotation, variant calling, and filtering

Automated functional annotation for each assembled genome was performed using prokka v1.14.5 package (Seemann, 2014). Filtered and trimmed Illumina sequencing reads were mapped to the reference genome of *Oscillatoria nigro-viridis* PCC 7112 (Accession no. GCA_000317475), and single nucleotide polymorphisms (SNPs) were detected by freebayes v1.3.2 (Garrison & Marth, 2012) with default parameters. We excluded indels and multiple nucleotide polymorphisms and kept only biallelic variants. Then, we used snippy v4.6 (https://github.com/tseeman/snippy) with default parameters to filter the low-confidence variants.

2.3. Pangenome diversity

We investigated the global genomic diversity of all *Microcoleus* strains sequenced in this study by characterizing the pangenome in Roary v3.13.0 (Page et al., 2015) with default parameters. As input to Roary, we used genome annotations in gff3 format. For the visualization, we used roary_plots python script (https://github.com/sanger-pathogens/Roary). To determine the gene cooccurrence (associations) and avoidance (dissociations), we computed networks using Coinfinder v1.0.7 (Whelan et al., 2020). We used a presence/absence matrix as the input with default parameters and a threshold of 0.05 by Bonferroni corrected exact binominal test.

2.4. Phylogenomic analysis

The species tree of *Microcoleus* isolates sequenced in this study (dataset I) was inferred following three different approaches. We used Orthofinder v2.3.1 (Emms & Kelly, 2019) with default parameters to identify single-copy orthologues, infer unrooted gene trees and acquire multiple sequence alignment (MSA). The first species tree was the maximum-likelihood (ML) tree inferred in IQ-TREE v1.6.1 (Nguyen et al., 2015) based on the best model selected by ModelFinder

(Kalyaanamoorthy et al., 2017) – JTT+F+I+G4 and ultrafast bootstrap with 2000 replicates. *Microcoleus* sp. was used as an outgroup (M2_D5). The second species tree was constructed using unrooted gene trees with a coalescent-based analysis in ASTRAL-III (Mirarab & Warnow, 2015; Zhang et al., 2018). Individual ML gene trees were inferred from the single-copy alignments in IQTREE based on the model identified for each alignment separately by ModelTest (Posada & Crandall, 1998). The third species tree was inferred from extracted consensus sequences as fasta files from SNPs detected by snippy and constructed the whole-genome alignment with *Oscillatoria nigro-viridis* PCC 7122 set as the reference (outgroup omitted). Subsequently, it was used for the ML inference in IQ-TREE based on the best model TVM+F+ASC+G4 and 2000 ultrafast bootstrap replications.

Orthofinder was used to obtain the MSA for the isolates from dataset II, and the ML tree was inferred with IQ-TREE. The ModelFinder module was employed to identify JTT+F+I+G4 as the best-fitting model for the phylogenetic reconstruction. The topology was tested with 2000 replicates.

Dataset III was used to reconstruct evolutionary relationships between selected *Microcoleus* isolates and other cyanobacteria. *Gloeobacter violaceus* PCC 7421 (Accession no. GCA_000011385) was used as an outgroup. We used Orthofinder to infer the MSA, which was then used as input for IQ-TREE and ML reconstruction. The best-fitting model selected by ModelFinder was LG+I+G and branch supports were computed using ultrafast bootstrapping with 2000 replicates.

2.5. Designating isolates to genetic clusters

We applied four methods to elucidate the population structure of *Microcoleus* isolates (SNPs from dataset I) and assign them to genetically distinct clusters. First, we used a hierarchical Bayesian clustering algorithm implemented in fastBAPS (https://github.com/gtonkinhill/fastbaps). Fastbaps clusters were determined using functions optimised.symmetric prior (optimized), and baps prior (unoptimized). Additionally, we used the snapclust function (Beugin et al., 2018), which uses the fast distance-based clustering implemented in the package adegenet v2.1.5 (Jombart, 2008; Jombart et al., 2018) in R (version 4.1.3; R Core Team, 2021). To identify the optimal number of clusters within the dataset, we used Akaike Information Criterion (AIC) and the function snapclust.choose.k. Lastly, the pairwise average nucleotide identity (ANI) was calculated with

fastANI v1.33 (https://github.com/ParBLiSS/FastANI) with a minimum fraction of the genome shared 0.1 and 1.5kb fragment length.

2.6. Inference of time-calibrated phylogeny, biogeographic history, and diversification rates

A time-calibrated phylogeny was inferred in BEAST (Bayesian Evolutionary Analysis Sampling Trees) v1.10.4 (Drummond & Rambaut, 2007) to estimate the age of splits of *Microcoleus* species. The analysis was performed on dataset I using the GTR+I+G model. The dating was performed using a 16S rRNA mutation rate estimated by Dvořák et al. (2012). The 16S rRNA sequences were aligned with mafft v7.453 (Katoh et al., 2002; Katoh & Standley, 2013), and the XML file for the BEAST was prepared in BEAUTi (Drummond & Rambaut, 2007). The MCMC chains were run for 10 million generations and sampled every 1000th generation. The 16S rRNA alignment was used only to estimate dating, while the tree topology was fixed to the tree topology based on the amino acid MSA of dataset I. The tree was dated using strict molecular clocks. After the MCMC run, the ESS values were evaluated in Tracer v1.7.1 (Rambaut et al., 2018), and they were all above 200. The TreeAnnotator (Helfrich et al., 2018) was used to produce the final tree with a burn-in of 25%.

Ancestral geographical area reconstruction was estimated in RASP v4.3 (Yu et al., 2015) using the ML reconstruction under the Bayesian binary model (BBM) with the fixed state frequency model (Jukes-Cantor) and nucleotide rate variation for 50000 generations. As input, we used ML phylogeny inferred from dataset II. We considered seven geographical areas of origin: Europe, North America, Africa, Asia, Australia, the Antarctic, and the Arctic. We allowed a single species could occupy a maximum of three geographic areas.

2.7. Phylogeography, phylogenetic signals, and Mantel

We investigated the relationship between genetic distance, geography, and different environmental variables using the Mantel test (Mantel, 1967) to explore drivers of genetic diversity and diversification in *Microcoleus*. The distance matrix (from dataset II) was constructed from the ANI values as 1-ANI and used as an input for obtaining mantel statistics. Bioclimatic variables were downloaded from the WorldClim v2.1 database (Fick & Hijmans, 2017) at 2.5 arc minutes resolution and extracted in R. Soil variables were downloaded from the ISRIC SoilGrids (www.isric.org), global UV-B radiation parameters from the glUV database (Beckmann et al.,

2014) and the global human appropriation of net primary production (HANPP) parameters from the data published by Haberl et al. (2007) and then extracted with QGis v3.22.8 software (www.qgis.org).

The spatial matrix was generated from the geographic coordinates of localities with the function distGEO from the R package geosphere v1.5-10 (Hijmans et al., 2017), and the environmental matrices were calculated with the euclidean distance method. The Mantel tests were performed in the vegan v2.5.6 (Oksanen et al., 2016) package in R with 9999 permutations. The significance of correlations was calculated with Pearson's r. We assessed the climatic niche preferences of the lineages, focusing on the variables exhibiting the highest correlations according to the Mantel test, particularly associated with precipitation, temperature, radiation, and soil properties. We then assessed the significance of differences between climatic niches by performing a Kruskal-Wallis nonparametric analysis of variance (Kruskal &Wallis, 1952), followed by Dunn's test for multiple comparisons (Dunn, 1964), along with Bonferroni correction, using the R package dunn.test v1.3.5 (Dinno, 2017).

Additionally, we measured two independent phylogenetic signal indices in R, Pagel's λ (Pagel, 1999) and Blomberg's K (Blomberg et al., 2003), using functions fitContinuous (Geiger v2.0.1; Harmon et al., 2015) and phylosignal (picante v1.8.2; Kembel et al., 2010) respectively. The significance of lambda values was performed with likelihood ratio tests.

2.8. Recombination analysis

The recent recombination within *Microcoleus* isolates (dataset I) was explored with Gubbins v3.1.3 (Croucher et al., 2015), and the whole-genome alignment generated for the SNP phylogeny was used as an input. It was run according to the software manual and using -first-tree-builder rapidnj and -tree-builder raxmlng parameters to build a recombination-free phylogeny with RaxML (Stamatakis, 2006). The outputs were visualized with Phandango v1.3.0 (Hadfield et al., 2018). Estimates r/m (ratio of the number of SNPs introduced through recombination relative to mutation), and ρ/θ (ratio of recombination relative to mutation rate) for each population and individual strains were derived from the Gubbins output. Owing to the many overlaps between recombination blocks, we counted all recombination events (both shared and unique) specific to one *Microcoleus* species (i.e., one block could have been counted more than once) to calculate the genome fractions subjected to recombination. Consequently, the sum of genome fractions

undergoing gene flow specific to one species may be larger than 100%. Calculated genome fractions were visualized as bubble and box plots using the R packages reshape2 v1.4.4 (Wickham, 2017) and ggplot2 v3.3.5 (Wickham, 2016). To test if there were any differences in recombination rates and the frequency of recombination among the populations, we used the Kruskal-Wallis test on all the estimated r/m and ρ/θ values. The test was then followed by Dunn's test with Bonferroni correction, performed in the R package dunn.test. We also tested whether the recombination parameters significantly differed between strains occupying different habitats (soil and puddle) with the Kruskal-Wallis test.

Additionally, we evaluated the relationship between the number of clustered regularly interspaced short palindromic repeats (CRISPR) per species with recombination parameters derived from the Gubbins output (r/m, ρ/θ, and the number of recombination blocks) to investigate how the presence of CRISPR spacer repeats affected recombination in *Microcoleus* (corr.test in R).

2.9. Historical demography

We explored the demographic history of each *Microcoleus* lineage by calculating neutrality statistics and inferring Bayesian skyline plots (BSPs). Tajima's D and Fu's F neutrality statistics were assessed in overlapping 10 kb sliding windows with a 2.5 kb step using the package PopGenome v2.7.5 (Pfeifer et al., 2014) in R. Significant departure of these statistics from 0 was estimated with a t-test.

BSP (Drummond et al., 2005) implemented in BEAST2 v2.7.4 (Bouckaert et al., 2019) was utilized to estimate changes in the effective population size of individual *Microcoleus* population across time. The input MSA was produced for each population separately. The XML files for each population were first generated with BEAUTi. The MCMC chains were run for 50 million generations and sampled every 10000th generation. The ML model was set to HKY, each tree root was restricted based on the node age of the divergence of each population (dataset I), and the dating was performed using strict clocks. The BSP was estimated using the Coalescent Bayesian Skyline model. The bPopSizes and bGroupSizes were set to 2 in population M1, in populations M2, M4, M8, and M9 to 5; in the rest of the populations, the value was set to 3. The ESS values were evaluated in Tracer v1.7.1 (they were all above 200; Rambaut et al., 2018), as well as the final BSP plots.

2.10. Estimation of genetic diversity and selection

We used three genetic parameters to estimate the intra- and interpopulation genetic differentiation of *Microcoleus* populations. We calculated nucleotide diversity (π), the fixation index (F_{ST}), and absolute divergence (D_{XY}) with the package PopGenome. All statistics were estimated in overlapping 50 kb sliding windows with a 12.5 kb step. The outputs were visualized using the R package ggplot2. Differences between nucleotide diversities per population were estimated with the Kruskal-Wallis test. For these analyses, we excluded population M13, which included a single strain (N3_A4).

Genomic regions of elevated genetic divergence among *Microcoleus* lineages were identified with both F_{ST} and D_{XY}. Windows within the 0.99 percentile of the highest F_{ST} and D_{XY} were selected, and only genomic regions identified by both measurements were considered responsible for the differentiation between the populations. Further, to infer the overall presence of positive selection, we selected the coding sequence (CDS) from all genes present in the genome annotation of *Oscillatoria nigro-viridis* PCC 7112. We performed the McDonald-Kreitman (MK) test on all samples with a Fisher's exact significance test in PopGenome. The MK test compares levels of polymorphism and divergence to detect deviations from neutrality regarding nonsynonymous substitutions while also controlling for gene-specific mutation rates. By calculating the neutrality index (NI), where an NI greater than 1 indicates an excess of silent divergence and an NI lower than 1 suggests an excess of nonsilent divergence (i.e., positive selection), the MK test is able to estimate whether a gene is under the effect of selection. For all the genes that were found to be significant and had an NI lower than 1, putative gene functions were obtained from the UniProt protein database (www.uniprot.org). The genes found within genomic regions of elevated divergence under the effect of positive selection were considered candidates associated with the differentiation between *Microcoleus* species.

3. Results

We obtained complete genome sequences for 202 *Microcoleus* isolates and eight *M. vaginatus* using cyanobacterial material preserved in herbaria. *De novo* assembly yielded 202 genomes of the final mean depth 65.05x, ranging in length from 6.39-9.64 Mb with 45.4-47.4% GC content. Genomes of herbarium specimens had a final mean depth of 28.8x and length between 6.90-8.28 Mb with 45.2-46.6% GC content. We identified a total of 1,038,522 SNPs in 201 *Microcoleus* individuals.

All features of individual genomes can be found in Supplementary Table 1.

3.1 Evolutionary history and diversification patterns of *Microcoleus* sp.

The evolutionary history between selected *Microcoleus* isolates and other cyanobacteria was reconstructed using the MSA with 129,335 amino acid sites (dataset III) to place our isolates in the broader cyanobacterial phylogeny. The inferred ML tree revealed that our strains clustered in a monophyletic clade, with the closest sister species being *Microcoleus* sp. M2_D5 and *Kamptonema animale* PCC 6506 (Supplementary Figure 1). Then we compared species trees inferred with three different approaches to explore the evolutionary relationships among our *Microcoleus* isolates (dataset I) – the ML tree from 2020 single-copy orthologues, the ML tree from SNPs, and the ASTRAL tree from a set of unrooted trees under the multispecies coalescent model (Supplementary Figures 2-4). There was a discordance in the phylogenetic positions of a few strains in the ASTRAL and the SNP tree compared to the ML tree from single-copy orthologues, likely due to incomplete lineage sorting. The topology of the species tree did not alter when recombination spots were removed (Supplementary Figure 5). Interestingly, most samples contained isolates from a single lineage (45), 11 samples contained isolates from two lineages, and one sample revealed *Microcoleus* isolates belonging to three distinct lineages (Supplementary Figure 2).

Both snapclust and Bayesian optimized clustering analyses recovered 13 clusters as the best-supported partitioning of *Microcoleus*, while unoptimized clustering yielded 18 clusters (Supplementary Figure 6). The ANI analyses showed that *Microcoleus* isolates shared 80.1-99.9% sequence identity across their genomes (Supplementary Figure 6, Supplementary Table 2), generating 37 clusters. Given that population genetic analyses are sensitive to population definition

12

and that (1) the Bayesian optimized method generated the optimal number of clusters following monophyly in the species trees, (2) the ANI thresholds do not reflect variable diversification rates of species, and it is often lower than the standard 95% threshold in microbes (Bobay et al., 2020; discussed below), we favor the relatively conservative clustering of 13 clades/lineages for our dataset, that represent putative species.

Ancestral geographical area reconstruction conducted to understand the spatial dispersal history and geographical origins of *Microcoleus* sp. suggested a European origin for most species (dataset II; Supplementary Figure 7). The basal clade includes African and North American strains, and it had a deep split from the common ancestor of all our *Microcoleus* isolates (Figure 1A). The phylogeographic patterns of our *Microcoleus* sp. isolates suggest that the species found in other geographical areas than Europe are descendants of more than one expansion out of Europe.
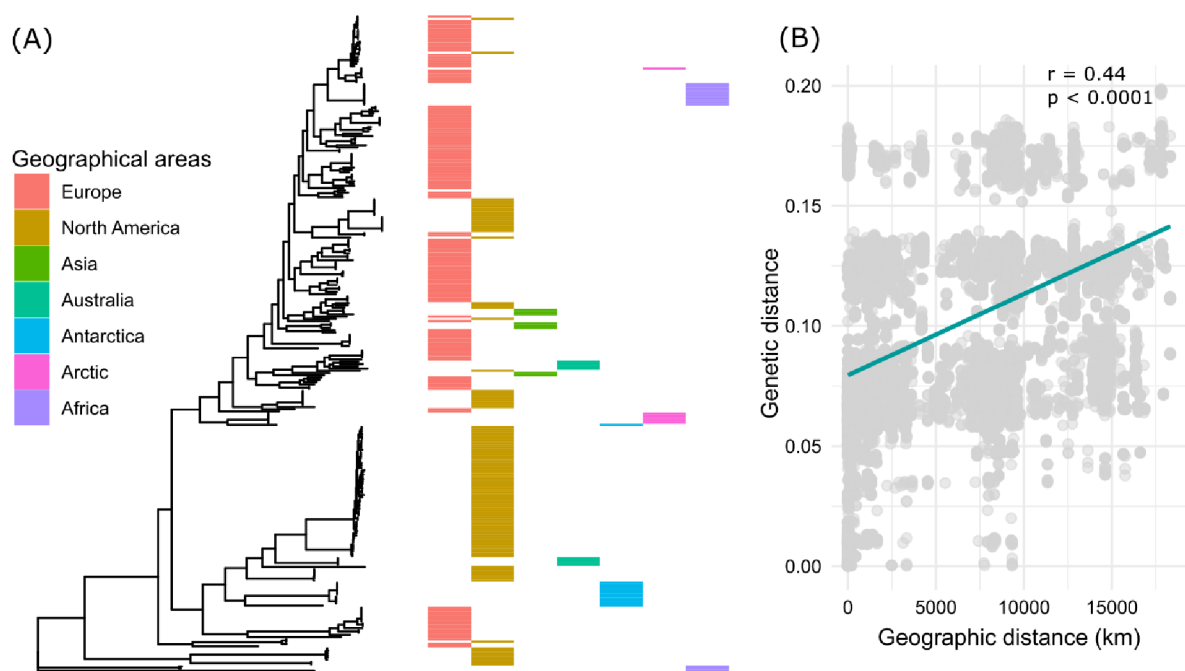
Figure 1. Phylogenomics of the global *Microcoleus* collection and the impact of isolation-by-distance. (A) Genome phylogeny of the *Microcoleus* isolates sequenced in this study with the genomes obtained from the GenBank (dataset II) and their distribution patterns. The strains are annotated with colors corresponding to the geographic areas of their isolation. (B) Scatter plot demonstrating a significant effect of isolation-by-distance on *Microcoleus* genetic diversity (r = 0.44, p < 0.0001).

Moreover, *Microcoleus* species displayed a distance-decay relationship (r = 0.44, p < 0.0001), indicating that dispersal was geographically restricted (Figure 1B).

The dating analysis of the 16S rRNA in BEAST was calibrated at an evolutionary rate of 0.001861 substitutions per site per million years and 95% highest probability density (0.000643-0.003079) (Dvořák et al., 2012). According to the dating, the divergence of our *Microcoleus* isolates commenced between Eocene and Oligocene, which was before 29.6 Ma (Figure 2A). The diversification of *Microcoleus* sp. resulted in at least 12 species that emerged during the middle and late Miocene and Pliocene Periods, which were between 4.7-13.7 Ma when the climate was generally warmer and wetter than it is currently (Figure 2A). One strain, N3_A4 formed a singleton node that was considered a separate lineage (M13); thus, it was excluded from population genetics analyses.

3.2 Genetic diversity and differentiation of the *Microcoleus* species

We analyzed the pangenome of *Microcoleus* sp. using Roary, which grouped the coding DNA sequences (CDS) into the core and flexible genome (shell and cloud). The core genome consisted of genes present in all 201 strains, while the soft core genes were present in $190 - 200$ strains. The shell contained genes present in $30 - 190$ strains, and the cloud had genes present in only $1 - 30$ strains. Overall, the pangenome contained 133,737 genes, with the core and soft genomes accounting for 0.54% of the pangenome and having 639 and 87 CDS, respectively (Supplementary Table 1). The shell part of the flexible genome comprised 6053 CDS, representing 4.52% of the pangenome, while the majority of the pangenome (94.9%) belonged to the cloud, which contained 126,958 CDS. Further, we searched for statistically significant associations and disassociations of genes within the pangenome and found the non-random cooccurrence of many genes belonging to the flexible genome. In particular, 10,945 metabolic genes (987 annotated) showed significant associations, while 5,458 genes (1,346 annotated) were significantly disassociating, suggesting the potential influence of selection on the flexible genome (Supplementary Tables 3A and 3B). The genes were involved in a range of functions, including transport, stress response, cell division, biosynthesis, toxin-antitoxin activity, and antiviral defense.
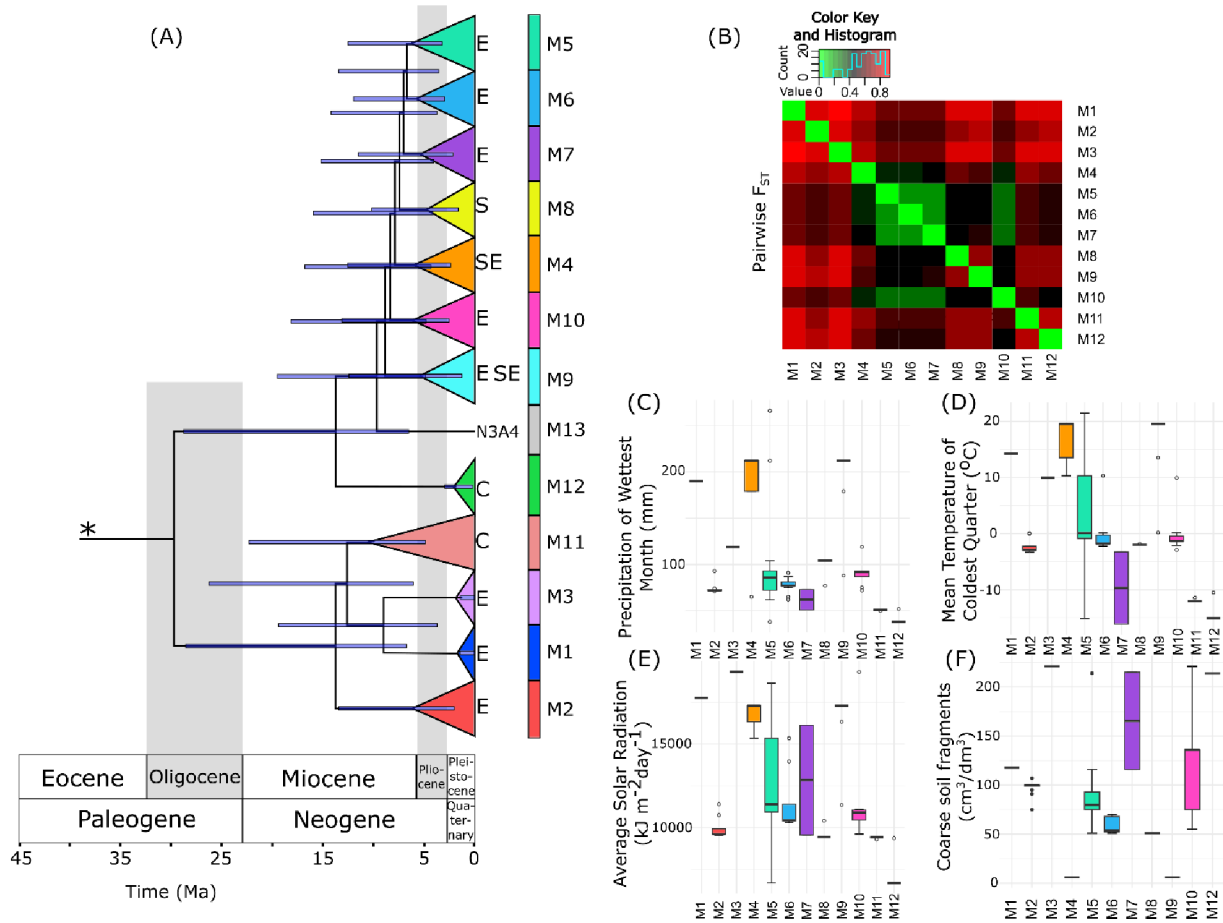
Figure 2. Phylogenetic relationships and divergence times of 13 *Microcoleus* lineages along with genetic and ecological differentiation patterns among them. (A) The dating of the divergence times among *Microcoleus* lineages based on 16S rRNA. Branches were collapsed, and the clades are color-coded according to the lineage designation. The blue bars represent 95% highest posterior density (HPD) values for each node. An asterisk denotes a node of the removed outgroup. The time axis is a million years ago (Ma) with chronological dating of geological intervals. The letters denote demographic events of species: E – population expansion, S – population shrinkage, C – constant effective population size (B) Heatmap for pairwise F_{ST} distances between lineages. The color spectrum from green to red illustrates genetic differentiation from low to high. (C, D, E, F) Differences in habitat preferences of lineages for (C) precipitation of wettest month, (D) mean temperature of coldest quarter and (E) average solar radiation, (F) volume of coarse soil fragments. Boxplot colors correspond to the lineages' color codes.

Using overlapping 50kb sliding windows spanning the genome, we evaluated the genetic diversity and divergence for all *Microcoleus* sp. species. Estimated mean nucleotide diversities (π) of lineages over the whole genome ranged from 0.000772-0.01067. The Kruskal-Wallis rank sum test suggested that π was significantly different between any two lineages (p < 0.0001). The mean

absolute divergence at the interspecific level spanned 0.011-0.017, whereas the mean fixation index ranged from 0.20-0.93 (Supplementary Table 4). The wide range of calculated F_{ST} values, from almost 0 to almost 1, implies that species pairs represent distinct stages across the divergence continuum of *Microcoleus* sp., from early to late stages of speciation (Figure 2B). Moreover, high interspecific genome-wide F_{ST} values detected between species indicate strong genetic differentiation, with the overall pattern of high heterogeneity along the genome and less evident peaks of elevated divergence in *Microcoleus* genomes (Supplementary Figure 8).

3.3. Ecological diversity and differentiation of *Microcoleus* lineages

The differences in climatic preferences of *Microcoleus* species in their habitats are shown in Figures 2C-F and in Supplementary Table 5 are significant differences between lineages identified by Dunn's test. The species M4 and M9 are adapted to habitats with high precipitation and temperature ranges and are significantly different from most other species (Supplementary Table 5). Further, species M1, M3, M4, and M9 occupy habitats with high average solar radiation, while species M3, M7, and M12 favor habitats with a larger volume of coarse fragments than the other species (Figures 2C-F). Conversely, species M7, M11, and M12 prefer dry, cold habitats with low radiation levels. In addition, species M4 and M9 opt for soils with finer particles than the other species. Overall, the species showed significant differentiation based on their ecological preferences.

To assess the role and identify potential environmental factors that could drive selective pressures, we performed tests for isolation by the environment (Supplementary Table 6). Firstly, all tested environmental variables showed a statistically significant phylogenetic signal, where Pagel's λ ranged from 0.95-099 (p < 0.0001) and Bloomerg's K had values below 0.001 (p = 0.001). These results indicate that the isolates' preferences for environmental niche space align well with their phylogenetic relatedness but might be evolutionarily constrained by other factors. Next, the Mantel test of genetic distances and environmental variables confirmed that contemporary climatic and habitat conditions govern the divergence of *Microcoleus* sp. species. The analysis showed that selective pressures were likely triggered by precipitation (annual range, r= 0.317, r = 0.001; the wettest month, r = 0.434, p = 0.0001), soil properties (organic carbon, r = 0.231, p = 0.0001; coarse fragments volume, r = 0.314, p = 0.0001), and human activities (net primary production remaining in the ecosystem after harvest, r = 0.548, p = 0.0001; total human appropriation on net primary

production, r = 0.515, p = 0.0001; nitrogen fertilizer levels (r = 0.237, p = 0.0001). Overall, these findings suggest that the genomic diversity and population structure of cosmopolitan *Microcoleus* species may be shaped by adaptive selection.

3.4. Demographic history

For neutrality statistics, Tajima's D values were significantly departing from 0 in all lineages, apart from M1 (t-test, p < 0.001, p = 0.47). Mean Tajima's D values over the whole genome varied from -1.35–1.73, while Fu's F ranged from -1.25–1.91. Statistically positive Tajima's D values were observed in four species (M3, M4, M7, and M10), which suggests that species might be under the balancing selection with signatures of population shrinkage. In contrast, all other species exhibited statistically negative Tajima's D values, which implies a recent population expansion and possible influence of the purifying selection on their genetic diversity. Similar patterns were observed with Fu's F (Supplementary Table 7).

Then, we reconstructed Bayesian skyline plots to estimate the effective population size change over time. Major oscillations in the estimated effective population size trajectory for all *Microcoleus* species occurred during the middle and late Pleistocene, around 200 to 500 kya (Supplementary Figure 9). During the Pliocene epoch, around 3-5 Ma, a gradual expansion in population size occurred in species M2, M5, M6, M7, and M10, while M1 and M3 expanded recently, 100-500 kya. M9 underwent the initial population expansion around 4.5 Ma, followed by shrinkage and expansion events during the Pleistocene, less than 500 kya. Species M11 and M12 had relatively stable population sizes, with small shrinkage events occurring 1.5 Ma and 12.5 kya before the present. A sudden population contraction happened less than 500 kya in species M8 and M4, but in M4, it was followed by a sudden expansion of the population as well (Figure 2A; Supplementary Figure 9).

3.5. Gene flow barriers between lineages

From the list of recombination events detected by Gubbins, we extracted all genomic regions that underwent recombination per strain. The mean genome fraction subjected to gene flow between strains of the same species varied from 10.9-53.1% (within the species), while the genome fraction affected by gene flow from outside the species varied from 1.5-29.04% (Figures 3A and 3B). The mean ρ/θ and r/m values estimated per species ranged from 0.012-0.041 and 0.38-2.0

17
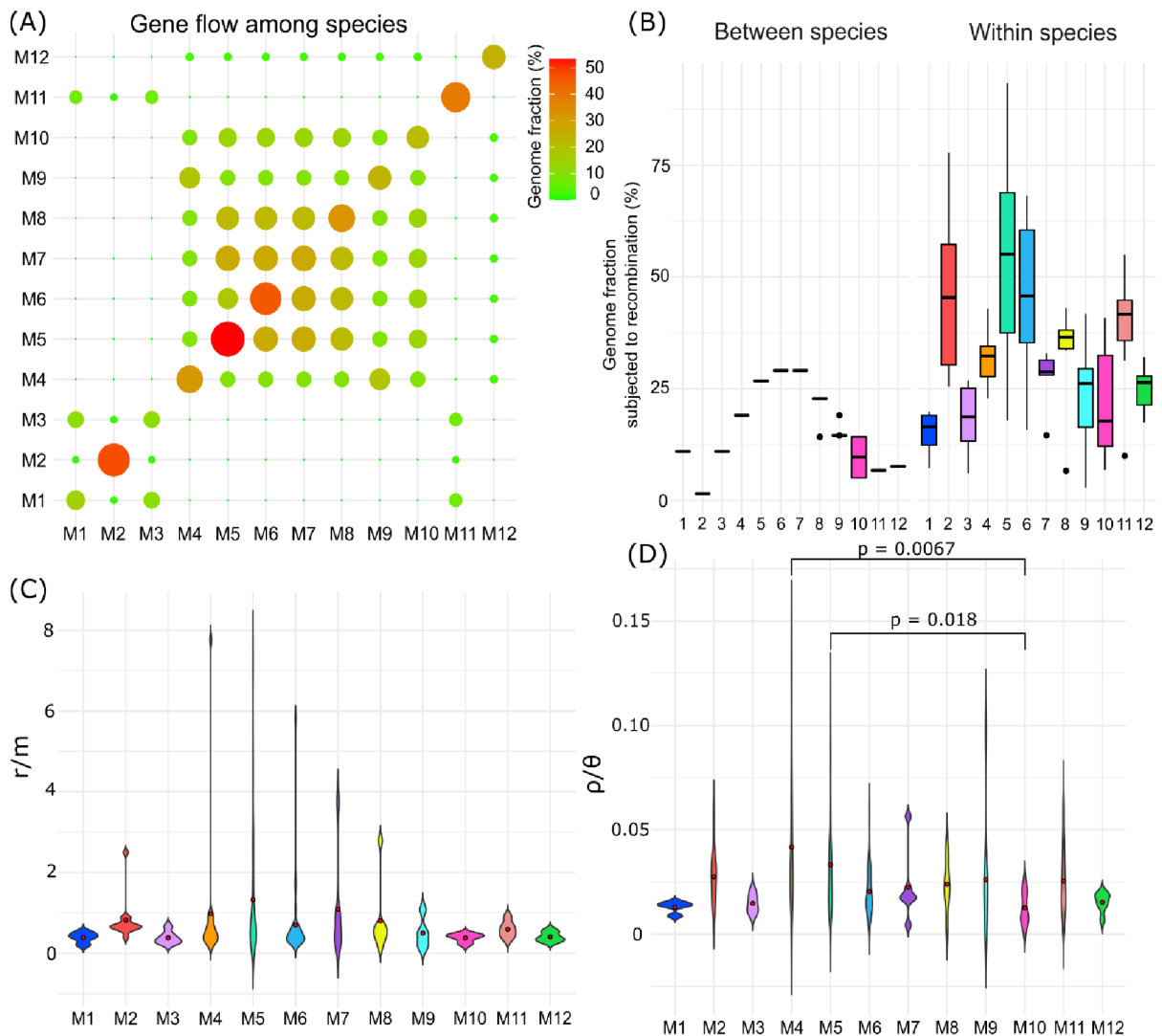
Figure 3. Patterns of gene flow within and between lineages along with violin plots of the per-strain recombination parameters. (A) Bubble plot illustrating the extent of gene flow among lineages based on the genome fraction subjected to recombination shared between a pair of lineages. The size and color of each bubble correspond to the level of gene flow, where the absence and limited gene flow between lineages are shown in green, while high levels are in red. (B) Boxplots indicating the precise variation of genome fraction subjected to recombination between and within lineages. (C) Violin plots illustrate differences in r/m ratios for each lineage. The highest outlier for lineage M5 (r/m = 18) was removed for clarity. (D) Violin plots of per-strain ρ/θ values for each lineage. Significant differences between lineages are indicated as p-values, determined by Kruskal-Wallis and Dunn's test with Bonferroni correction for multiple testing. The colors of violin plots correspond to the lineages' color codes from Figure 2. The red dot represents the average value of parameters.

(Supplementary Table 8). The distribution of ρ/θ and r/m estimates of isolated in each species is shown in Figures 3C and 3D. The Kruskal-Wallis rank sum test indicated that ρ/θ significantly

differed between any two species (p = 0.0002967), suggesting species must evolve differently regarding their recombination rates. The post hoc pairwise statistical testing using Dunn's test revealed that the recombination rates were significantly different between M4 and M10 (p adj. = 0.013) as well as M5 and M10 (p adj. = 0.036). This is indicative that recombination events happen more frequently in species M4 and M5 than in M10. Conversely, although the Kruskal-Wallis test indicated that r/m significantly differed between some species (p = 0.0082), Dunn's post hoc test showed no pairwise differences. These results imply that selection might not have had a chance yet to remove deleterious mutations introduced by recombination owing to lineages being very closely related and recently diverged. In addition, the Kruskal-Wallis test showed significant differences in the recombination parameters between strains occupying different habitats (ρ/θ, p = 0.001; r/m, p = 0.0009), suggesting the influence of habitat preference on the variability in HR levels (Supplementary Figures 10A and 10B).

We conducted correlation tests to explore the potential impact of CRISPR repeats on the recombination fluxes. Our findings revealed a significant positive correlation between the number of CRISPR repeats and the r/m ratio (Pearson's correlation coefficient = 0.157, p < 0.05), suggesting that species with more CRISPRs tend to introduce more genetic variation through recombination than mutations. On the other hand, a significant negative correlation between the number of CRISPR repeats and the frequency of recombination blocks/events (Pearson's correlation coefficient = -0.2, p < 0.005) indicates that species with fewer CRISPRs generally experience more frequent recombination events.

3.6. Genomic scan for divergence and candidate genes related to genetic divergence among *Microcoleus vaginatus*

Numerous genes are involved in the genetic divergence and ecological differentiation among *Microcoleus* species. Here, we focused on characterizing the functions of genes found in genomic regions of elevated F_{ST} and D_{XY} values (99th percentile). We found 58 annotated genes and 74 genes with a hypothetical function in four genomic regions. Out of all, 19 genes were found to be under positive selection (14 annotated; NI < 1), 13 genes were found to be under negative selection (10 annotated; NI > 1), seven genes were under both negative and positive selection (5 annotated) in different lineages, and the rest did not show the signature of selection (Table 1). Annotated genes with the signature of positive selection had functions associated with stress response (10),

chemotaxis (1), biosynthesis (6), cell wall formation (1), and transposition (2). Despite not being in the region of high genetic differentiation, other genes may play important roles in species divergence as the selection was widespread across the genome (a complete list of genes affected by selection per population can be found in Supplementary Table 9; Supplementary Figure 11).

Table 1. Genes within the 99th percentile of elevated F_{ST} and D_{XY} genomic regions identified in the McDonald-Kreitman test as significantly selected. Given are their gene names, a short functional description (from the UniProt), and a neutrality index (NI). The NI < 1 indicates positive selection, and NI > 1 indicates negative selection. Loci under positive and negative selection between pairwise lineage comparisons are marked with an asterisk.

	Gene ID	Gene name	Localization	Function	Neutrality Index (NI)
1	dxs	1-deoxy-D-xylulose-5-phosphate synthase	Core genome	Biosynthesis	N < 1 and N > 1
2	hepT	Heptaprenyl diphosphate synthase component 2		Toxin-antitoxin system	N < 1
3	srrA	Transcriptional regulatory protein		Transcription	N < 1
4	cheR	Chemotaxis protein methyltransferase	Flexible genome	Chemotaxis	N < 1 and N > 1
5	COQ3_4	O-methyltransferase		Biosynthesis	N > 1
6	glyS	Glycine-tRNA ligase subunit β		Biosynthesis	N < 1
7	hcnC	Hydrogen cyanide synthase		Biosynthesis/Oxidoreductase activity	N > 1
8	hemY	Protoporphyrinogen oxidase		Biosynthesis	N > 1
9	ISAcma16	IS4 family transposase ISAcma16		Transposition	N < 1
10	ISWen2	IS110 family transposase ISWen2		Transposition	N < 1
11	menD	2-succinyl-5-enolpyruvyl-6-hydroxy-3-cyclohexene-1-carboxylate synthase		Biosynthesis	N < 1
12	panE	2-dehydropantoate 2-reductase		Biosynthesis	N > 1 and 1 N < 1
13	pchA	Salicylate biosynthesis isochorismate synthase		Biosynthesis	N > 1

14	phrB	(6-4) Photolyase		DNA repair	N > 1
15	pknD_39	Serine/threonine-protein kinase PknD		Protein phosphorylation/Virulence	N < 1 and N > 1
16	ppsA_1	Phosphoenolpyruvate synthase		Phosphorylation	N > 1
17	rcp1_3	Response regulator rcp1		Stress response	N < 1
18	rcsC_52	Sensor histidine kinase RcsC		Stress response/Capsule biogenesis or degradation	N > 1
19	rcsC_74	Sensor histidine kinase RcsC		Stress response/Capsule biogenesis or degradation	N < 1
20	rcsC_75	Sensor histidine kinase RcsC		Stress response/Capsule biogenesis or degradation	N < 1
21	rcsC_76	Sensor histidine kinase RcsC		Stress response/Capsule biogenesis or degradation	N < 1
22	rcsC_78	Sensor histidine kinase RcsC		Stress response/Capsule biogenesis or degradation	N > 1
23	rcsC_93	Sensor histidine kinase RcsC		Stress response/Capsule biogenesis or degradation	N < 1
24	sasA_33	Histidine kinase		Stress response/Biological rhythms	N > 1
25	sasA_43	Adaptive-response sensory-kinase SasA		Stress response/Biological rhythms	N < 1
26	sasA_44	Adaptive-response sensory-kinase SasA		Stress response/Biological rhythms	N < 1
27	trmR_2	tRNA 5-hydroxyuridine methyltransferase		tRNA processing	N > 1

28	murF	UDP-N-acetylmuramoyl-tripeptide--D-alanyl-D-alanine ligase		Biosynthesis/Cell wall activities	N < 1, N > 1, Inf
29	hypothetical	Putative oxidoreductase			N < 1

4. Discussion

Microcoleus plays a pivotal role as ecosystem engineers in terrestrial environments, shaping the dynamics of local microniches, soil properties, and nutrient cycling, sustaining the life of many organisms. Here we finally shed light on the emergence of a continuum of *Microcoleus* species over time. This study employed a population genomics approach to unravel the evolutionary processes responsible for the differentiation of incipient, closely related *Microcoleus* species while considering ecological and geographic factors. Our analyses demonstrated that *Microcoleus* species are at varying stages of speciation with gene flow, which reflects the speciation continuum. We found that the diversification of *Microcoleus* species commenced between 4.7-13.7 Ma (middle Miocene to Pliocene), with major demographic oscillations occurring during the Pleistocene. A considerable portion of the pangenome comprised a non-randomly associated accessory genome, suggesting its substantial role in diversification. Accessory genes associated with stress response and biosynthesis processes, located in regions of elevated genetic divergence, exhibited signatures of positive selection, indicating their tentative role in adaptive evolution. Our findings suggest that *Microcoleus* species divergence may have involved a combined interplay of speciation with gene flow, ecological, and geographic factors.

4.1. Navigating the grey speciation zone: From populations to distinct species

Population and comparative genomics provide valuable resources to extend our knowledge of the potential contributors to the development of the continuum of barriers to gene flow (Stankowski & Ravinet, 2021). Comparisons of the genome-wide divergence patterns unveiled that some *Microcoleus* species are high and some slightly genetically differentiated, thereby capturing the continuum of genetic divergence (F_{ST} spanning 0.20-0.93; Supplementary Table 4). Although some species exhibited a high genetic differentiation, they retained the signature of admixture with others, with up to 70% of the genome subjected to recombination (Figures 3A and B). This

underlies that barriers to gene flow are at various stages of completion along the continuum of *Microcoleus* species. The permeability of barriers allows *Microcoleus* to diverge with gene flow gradually.

Common crypsis across microbial species, undersampling on a population level (apart from pathogenes), and still missing adequate species concepts complicate prokaryotic delineation (Ward et al., 2008). For a long time, the golden standard for microbial species delineation has been a combination of various (dis)similarity thresholds, such as genome-wide ANI and 16S rRNA (Konstantinidis & Tiedje, 2005; Richter & Rosselló-Móra, 2009). Although pragmatic, arbitrary thresholds do not capture the complexity of the evolutionary processes and dynamics of evolutionary changes over time (Hugenholtz et al., 2016). For instance, the genome-wide ANI of *Microcoleus* sp. (80.14 to 99.996%) suggests as many as 37 independent species; thus, we favored a more conservative approach, where 13 groups assigned by clustering algorithm corresponded to the monophyly of the species (Supplementary Figure 6). Moreover, when delineating microbial species, it is essential to take into consideration the evolutionary patterns of divergence, including gene flow, population structure, selection, and genetic and ecological differences among lineages, and to place them along the speciation continuum (Dvořák et al., 2023; Bolnick et al., 2023).

Previous evolutionary studies demonstrated that the initial stages of the divergence process are characterized by high levels of gene flow between individuals of a population with the selection of specific genomic regions that are limited to a local population (Shapiro et al., 2012; Cadillo-Quiroz et al., 2012; Chase et al., 2019). While these studies confirmed the presence of genomic regions containing genes potentially associated with adaptation to local ecological microniches, such as nutrient uptake and stress response genes, they were limited to 2 or a few incipient species comparisons. In contrast, we observed a continuum of 12 *Microcoleus* species at different stages of genetic divergence (Figures 2A and B). We found no apparent peaks of elevated differentiation (neither 'islands' nor 'continents') across the genome but rather a broad elevation of its diversity (Supplementary Figure 8). This suggests that divergence patterns in *Microcoleus* may have temporary character after the initial species split, as previously observed in some prokaryotes, e.g., *Laspinema* (Stanojković et al., 2022b) and *Sulfolobus* (Cadillo-Quiroz et al., 2012) and eukaryotes, e.g., *Heliconius* (Martin et al., 2013) or *Rhagoletis* (Michel et al., 2010). Our findings supported

these studies by showing that divergence is likely maintained in multiple genomic regions from the beginning instead of being localized in a few 'islands'.

The high ratios of r/m (mean 0.38-2.0; Figure 3C) between species indicate that recombination has been important for the introduction of SNPs in *Microcoleus* more than mutation, while low ρ/θ ratios (mean 0.012-0.041; Figure 3D) highlight a general pattern of the clonal nature of *Microcoleus* individuals. These estimates aligned with previous reports of recombination rates for cyanobacteria and free-living bacteria (Vos & Didelot, 2009). Recombination analyses further suggest the connectivity by gene flow among strains from different species, despite being separated by large spatial scales, which concurs with observations that *Microcoleus* commonly undergoes gene flow (Bouma-Gregson et al., 2022). In some species (M5 and M6), the genome fraction subjected to recombination exceeds 70%, underlining gene flow as an important driver of *Microcoleus* diversity (Figure 3B). Additionally, the genome fraction shared within species is much higher than that shared between them, denoting species-specific recombination and implying the existence of a recombination barrier which may contribute to the maintenance of strains in cohesive genetic groups (Figures 3A and B). Bearing a significant role in the speciation of *Microcoleus*, gene flow may introduce new, locally beneficial alleles, enabling local adaptation to diverse conditions in the local environments.

The overall genomic patterns of the high genome-wide divergence (Supplementary Figure 8), gene flow, and selection (discussed below) across the speciation continuum might explain how *Microcoleus* species emerged. Individual strains with low-fitness genotypes might have genetic modifiers that increase the recombination rates, which would be favored by selection when exposed to abrupt changes in environmental conditions like desiccation or nutrient limitation (Redfield, 1993; Hadany & Becker, 2003). The ability of some *Microcoleus* individuals to fine-tune their mutation and recombination rates, also called fitness-associated recombination (Hadany & Becker, 2003), might provide them an advantage in local microniches over others that lack it. Six of the 12 species studied had one to a few highly recombinant genotypes (n = 23, r/m > 2, Supplementary Table 8) that might oversee acquiring novel adaptive alleles from the environment. Further, the Kruskal-Wallis test showed notable discrepancies in the levels of recombination (r/m and ρ/θ) among strains originating from different habitats (Supplementary Figures 10A and B). For instance, strains found in soil often had higher HR levels than those found in puddles, which

could imply that soil environments are more unstable. A variety of prokaryotes and eukaryotes have been previously shown to undergo gene flow more often than mutation when exposed to stress (e.g., *Bacillus*, Jarmer et al., 2002; *E. coli*, Foster, 2005; *Aspergillus*, Schoustra et al., 2010), highlighting its importance of removing deleterious mutations by accelerating the rate of adaptation. However, the number of recombinant genotypes obtained from other *Microcoleus* species was limited. They might occur in nature but remain unsampled, so the mechanisms driving the speciation may become more complex and intertwined.

Comprehending the evolutionary patterns of closely related incipient species involves defining the stages along the speciation continuum, which can range from the emergence of adaptive variation within individuals that freely exchange genes in a population to complete barriers to gene flow between distinct species (Nosil & Feder, 2012; Shapiro & Polz, 2015). A scenario with prevalent population-specific recombination, i.e., dominant gene flow within the species, conforms to the biological species concept (Mayr, 1942). However, the biological species concept requires fully developed barriers to the gene flow between two species. Some *Microcoleus* species have completed barriers to gene flow, while others remain in an intermediate, grey zone (Figures 3A and B). As a result, assigning species to distinct units is not feasible until they have fully diverged as required by biological and any other species concepts (Kollár et al., 2022). A new approach to studying currently diverging species is necessary, and the universal probabilistic concept of evolutionary lineages has been developed for this purpose – UPCEL (Kollár et al., 2022). UPCEL places probabilities on the populations based on their divergence through gene flow, which can vary over time. In other words, putative species are not defined discretely, but their divergence is described using the probability at a given time. As a proxy of the divergence probability, we can use the genome fraction resistant to gene flow between the evolving species (Kollár et al., 2022).

Following Shapiro & Polz's (2015) and Kollár et al. (2022) speciation frameworks, we considered all the patterns of genetic differentiation, where F_{ST} roughly coincides with the extent of gene flow (Figure 2B, Supplementary Figures 8) and recombination (Figure 3) to place different *Microcoleus* species along the speciation continuum. We estimate that M1, M2, M3, M11, and M12 are on separate evolutionary trajectories, and they can be considered separate species based on the biological species concept (stages 4-5). Moreover, regarding UPCEL, they have a >93.24% probability of becoming fully separated species (between M1 and M3 is 89%). On the other hand,

M4, M9, and M10 remain in the grey zone (stages 2-3) with a probability of >85.7% of eventually reaching speciation completion. The remaining M5, M6, M7, and M8 have a probability of >73.2% becoming fully separated species, and they are likely in the early speciation stages (stages 1-2).

4.2. Historical demography of speciation

The initial radiation of *Microcoleus* species started during the transition between Eocene and Oligocene Periods (before 29.6 Ma, Figure 2A) when the climate was warmer and wetter than it is today. The split between two major *Microcoleus* clades coincides with the drop in global temperatures (Zachos et al., 2001) and the enhanced aridification period, which began at around 34 Ma and prevailed across the Eocene/Oligocene transition (Sun & Windley, 2015).

Molecular dating analysis showed that four *Microcoleus* species (M1-M3, M11, M12) diverged around 10-15 Ma in the Miocene, whereas seven (M4-M10) diverged almost at the same time in the late Miocene and Pliocene, around 5-10 Ma. Compared to macroorganisms that have relatively short divergence times (e.g., mosquitoes, 0.7-4.7 Ma, Mirabello & Conn, 2008; sunflowers, 1.5-2.2 Ma; Lee-Yaw et al., 2019), *Microcoleus* and other microorganisms have much longer divergence times (e.g., diatoms, 0.6-5 Ma, Casteleyn et al., 2010; protists, 8-25 Ma, Škaloud et al., 2019; cyanobacteria, 8-26 Ma, Dvořák et al., 2012; pathogens 20-21 Ma, den Bakker et al., 2010). The climate was warm during the middle and late Miocene, with increasing aridification (Eronen et al., 2012). This trend transpired during the Pliocene, which had a cooler climate and experienced high-magnitude shifts in temperature and precipitation (Zachos et al., 2001; Eronen et al., 2012). Long-distance dispersal could be one possible scenario of the mechanisms by which climatic fluctuations and aridity facilitated divergence within *Microcoleus*. The widespread occurrence of aridity might have resulted in the transportation of soil particles across local and regional distances by various processes like aeolian and fluvial, as well as the continuous growth on soil, via animal vectors or their combination, ultimately leading to the long-distance passive dispersal of *Microcoleus* (Elliott et al., 2019). Deserts, shrublands, and grasslands also expanded globally (Herbert et al., 2016), possibly providing vacant ecological niches suitable for cyanobacteria to colonize. Consequently, these processes might have enhanced the divergence of *Microcoleus* and facilitated speciation.

The species M7, M9, and M10 reached their maximum effective population size and population expansion during the warmer Period of the late Pliocene (around 2-5 Ma), whereas in most other species, these events occurred during the middle and late Pleistocene (around 200-500 kya). Repeated cycles of climate shifts during the Pliocene and Pleistocene (Hays et al., 1976; Huang et al., 2016) could have caused recent demographic oscillations in *Microcoleus* species by creating new habitats. These estimates concur with previous estimates that Pleistocene climate oscillations likely triggered speciation and diversification events of protists, plants, and mammals (Casteleyn et al., 2010; Levsen et al., 2012; Weir & Schluter, 2004). Hence, increased global aridity and high-magnitude climate variability could have heightened the divergence times, diversification rates, and demography changes between *Microcoleus* incipient species.

These population demography changes mostly concurred with Tajima's D and Fu's F values, except for the population shrinkage observed in lineages M3 and M10, which was not recovered by these statistics (Supplementary Table 7). Notably, inconsistency between these two approaches is common in population genetics studies. Neutrality statistics estimate the balance between genetic diversity and the frequency of rare alleles in a population and, thus, may be less precise compared to the BSP approach, which provides a more detailed estimation of population dynamics as well as genealogical relationships and coalescence processes (Grant, 2015).

4.3. Selection on stress response genes initiated adaptive evolution

Ecological differentiation is essential during the early stages of speciation (Feder et al., 2012). We studied the significance of the relationships between genetic distance and environmental factors (Mantel test and phylogenetic signal test) on a large global dataset. We demonstrated that precipitation, soil properties like nitrogen levels and volume of coarse fragments, and human-induced changes (net primary production of an ecosystem) as the drivers of genetic variation observed among *Microcoleus* species (Supplementary Table 6). In fact, the preference for climatic niche space has already emerged as the species tended to occupy niches with significantly different precipitation levels, soil properties, temperature, and light regimes (Figures 2C-F, Supplementary Table 5), providing additional support for isolation by the environment. A similar pattern was previously observed, where the availability of nutrients drove the abundance of physiologically

different *Microcoleus* strains, although the sampling design and the nature of the studied river system might have obscured the discovery of a much finer diversity (Bouma-Gregson et al., 2022).

Several *Microcoleus* species have developed mechanisms to withstand desiccation and thrive in particular habitats, such as humid and warm soils exposed to high radiation levels containing fine particles (M4, M9) or colder and drier environments by tolerating freezing temperatures or producing mucilaginous sheaths (M7, M11, and M12). Possible varied strategies of species to respond to shifts in environmental selection pressures might explain the patterns of variable HR rates observed in some *Microcoleus* individuals (Figure 3C). Differential HR rates could have been generated through temporal and spatial heterogeneities by favoring different beneficial alleles or modulating processes such as adaptations to a new environment (Dapper & Payseur, 2017). However, more data is needed to draw more robust conclusions. These observations underlie a different ability to respond to stress and nutrients between species, which could also contribute to ecological differentiation (Figures 2C-F, Supplementary Table 5).

Soil represents an extremely heterogenous mosaic environment, with varying textures, aeration, hydration, and nutrient levels even on a micrometer scale, thereby harboring a myriad of potential ecological microniches for cyanobacteria (Vos et al., 2013). Human-induced changes in the soil composition or dynamic shifts of hydration in drylands could disrupt the microniches and consequently elicit new selective pressures on populations. Differences in a huge accessory genome (94.9% of the pangenome; Supplementary Table 1) and a non-random association of hundreds of genes (Supplementary Table 3) might have contributed to fine variations in fitness between *Microcoleus* species (Polz et al., 2013). Moreover, the MK test supported the selection of loci scattered across the genome (Supplementary Table 9, Supplementary Figure 11), indicating that *Microcoleus* species are poised to respond quickly to sudden changes in heterogeneous environments. Although the preferred ecological niches of *Microcoleus* species are still unknown, this divergence pattern could be a signature and another evidence of ongoing ecological differentiation affecting the *Microcoleus* speciation continuum.

Interestingly, in the regions of elevated differentiation (outliers in the 99th percentile of Fst and Dxy), the majority of positively selected genes were part of the flexible genome. The MK test revealed that positively selected genes were functionally linked to stress response (Table 1), a characteristic that has been observed during the speciation of cyanobacteria (*Prochlorococcus*,

Kashtan et al., 2014; *Laspinema*, Stanojković et al., 2022b) and other prokaryotes (*Vibrio*, Shapiro et al., 2012). The fact that these genes likely initiated the process of divergence means that ecological adaptation or ecological barrier to gene flow has already been established between some of the *Microcoleus* species.

Drylands are characterized by an overall climatic water scarcity and often irregular shifts in extreme aridity and increased moisture with extremely scarce vegetation (Prăvălie, 2016). Investigating the putative function of the genes found in the region of elevated differentiation may provide an understanding of how *Microcoleus* prevails over these regions. The rcsC locus encodes a transmembrane sensor kinase, which plays a critical role in phosphorelay, a signal transduction mechanism associated with various responses to environmental stimuli, such as osmotic shock and capsule synthesis (Chen et al., 2001). The rcsC proteins are also known to serve as crucial virulence factors in pathogenic bacteria (Takeda et al., 2001), potentially contributing to host interactions. Further, genes regulating the biosynthesis of proteins (glyS), menaquinone (menD), or different cofactors (panE) contribute to the ability of bacteria to generate energy for essential processes like resource exploitation or protection against oxidative stresses (Table 1). Indeed, transcriptomic experiments on cyanobacteria showed the same gene function as our selection analysis. The desiccation and rehydration experiments confirmed that many genes linked to signal transduction (e.g., che locus), protein transports, and biosynthesis are expressed when exposed to stress (Murik et al., 2017; Rajeev et al., 2013). These findings indicate that adapting to xerotolerance and osmotic pressures in microniches of arid environments may involve precise genetic adjustments in a few key functional pathways.

4.4. Exploring the speciation continuum of *Microcoleus* sp.

Many mechanisms can evoke barriers to gene flow, a common feature of emerging species. Microbes were expected to have unlimited dispersal based on morphology, expressed by the famous tenet "everything is everywhere, but the environment selects" (Baas Becking, 1934; Finlay, 2002). Thus, the species were thought to evolve in sympatry. However, our results concur with previous large-scale studies on cyanobacteria *Microcoleus* (Dvořák et al., 2012; Stanojković et al., 2022a) and *Raphidiopsis* (Ribeiro et al., 2020) that found patterns of genetic differentiation consistent with the isolation by distance (r = 0.44, p = 0.0001; Figure 1B). Hence, the existence of

some geographical barriers on a large scale, i.e., allopatric differentiation, is suggested. Gene flow between *Microcoleus* species persisted across spatially distinct sites, such as between North America and Europe and other regions (Supplementary Figure 7). Notably, the dispersion of European strains to North America occurred multiple times within different species spanning a wide period, ranging from around 5.98 Ma in species M4, 5.43 Ma in M9, 1.3 Ma in M6 to 60 kya in species M5 (Figures 1A and 2A), while North American strains dispersed to Europe in one event 4.3 Ma in M10. Moreover, the split between European and Asian strains happened around 3 Ma, while Antarctic strains diverged from North American and Australian ones at around 12.57 Ma in the past. These findings suggest that barriers may have only intermittently played a role in the dispersal and geographical limitation of *Microcoleus* species (Bahl et al., 2011; Dvořák et al., 2012). Additionally, we observed multiple genetically distinct *Microcoleus* species coexisting in sympatry, which could mean that physical barriers can also emerge in centimeter-scaled niches (Kashtan et al., 2014; Wielgoss et al., 2016; Chase et al., 2019; Supplementary Table 1).

The barriers to gene flow might also arise due to sequence divergence between species (Cordero et al., 2012), the presence of CRISPR repeats, or the arrangement of ecological microniches (Shapiro & Polz, 2014). Genetic barriers to gene flow have already emerged between many *Microcoleus* species, demonstrated through preferential HR observed within rather than between species (Figures 3A and B). The mechanisms of the DNA uptake in filamentous cyanobacteria are not well understood. The natural transformation has only been recently achieved under laboratory conditions in *Microcoleus'* sister species, *Coleofasciculus* (Nies et al., 2020), and the number of CRISPR repeats in microbes has been linked to their resistance to phage predation and viral-mediated HGT (Held et al., 2010; Dvořák et al., 2017). The significant correlation between the number of CRISPR repeats and recombination rates of *Microcoleus* (Pearson's correlation coefficient = 0.157, $p < 0.05$) suggests that transduction may shape the observed recombination patterns and mechanistic barriers to gene flow, along with transformation (Held et al., 2010). Nevertheless, further investigation is needed to determine the origin and mechanisms of these barriers to gene flow as well as their significance to *Microcoleus* evolution.

Our study has certain shortcomings. Firstly, particular clades (M1, M3, M7, M12) analyzed in this study included a relatively small number of isolates ($\leq$ 6), which may have resulted in an underestimation of their diversity. More isolates from these clades should be included in future

genomic studies. Secondly, it is possible that other environmental factors, besides those considered in our study, may have also affected the divergence and selection pressures observed between strains. Thirdly, the correlations between variables and genetic diversity presented here do not necessarily imply the direction of a causal relationship (whether ecology and geography drive the barriers to gene flow or vice versa), so further physiological, transcriptomic, taxonomic, and epigenomic explorations will be performed to confirm the role of the putative adaptive genes identified here.

5. Conclusions

Microcoleus vaginatus has emerged to be among the most important microorganisms in topsoil communities, serving as the fundamental ecosystem engineer and an instigator of the carbon and nutrient cycles in drylands. From enhancing soil properties and affecting ecological niches at microscales, it played a crucial role in shaping the evolution of other microbial communities and plants for billions of years (Couradeau et al., 2019). However, until now, it remained poorly understood how *Microcoleus* diversified to have such an impact on dryland ecosystems, and this is closely tied to the still unclear relative importance of evolutionary forces on its speciation.

Overall, this study represents the largest speciation continuum observed among prokaryotes and the most comprehensive examination of genetic isolation and genomic differentiation locally and globally in free-living soil prokaryotes, particularly cyanobacteria. We identified 12 *Microcoleus* species at various stages of genetic differentiation, and they make the global continuum of *Microcoleus* species. They display an enormous genomic diversity, predominantly due to variation in gene content resulting from the frequent exchange of DNA with their local environment. Genome flexibility, along with patterns of gene flow within *Microcoleus* species, are the results of natural selection favoring the survival of organisms in fluctuating environmental conditions of drylands. The observed speciation patterns of *Microcoleus* presented in this study can be generalized with a model that predicts different stages of microbial speciation, considering various evolutionary forces and mechanisms (Figure 4). Collectively, our findings suggest that the speciation of microbial species entails divergence over the whole genome by periodic allopatric speciation, ecological speciation with a selection acting on loci in the accessory genome, and

extensive gene flow. Further studies on the speciation continuum in other species are needed to reveal how barriers to gene flow emerge and affect microbial speciation and diversification.
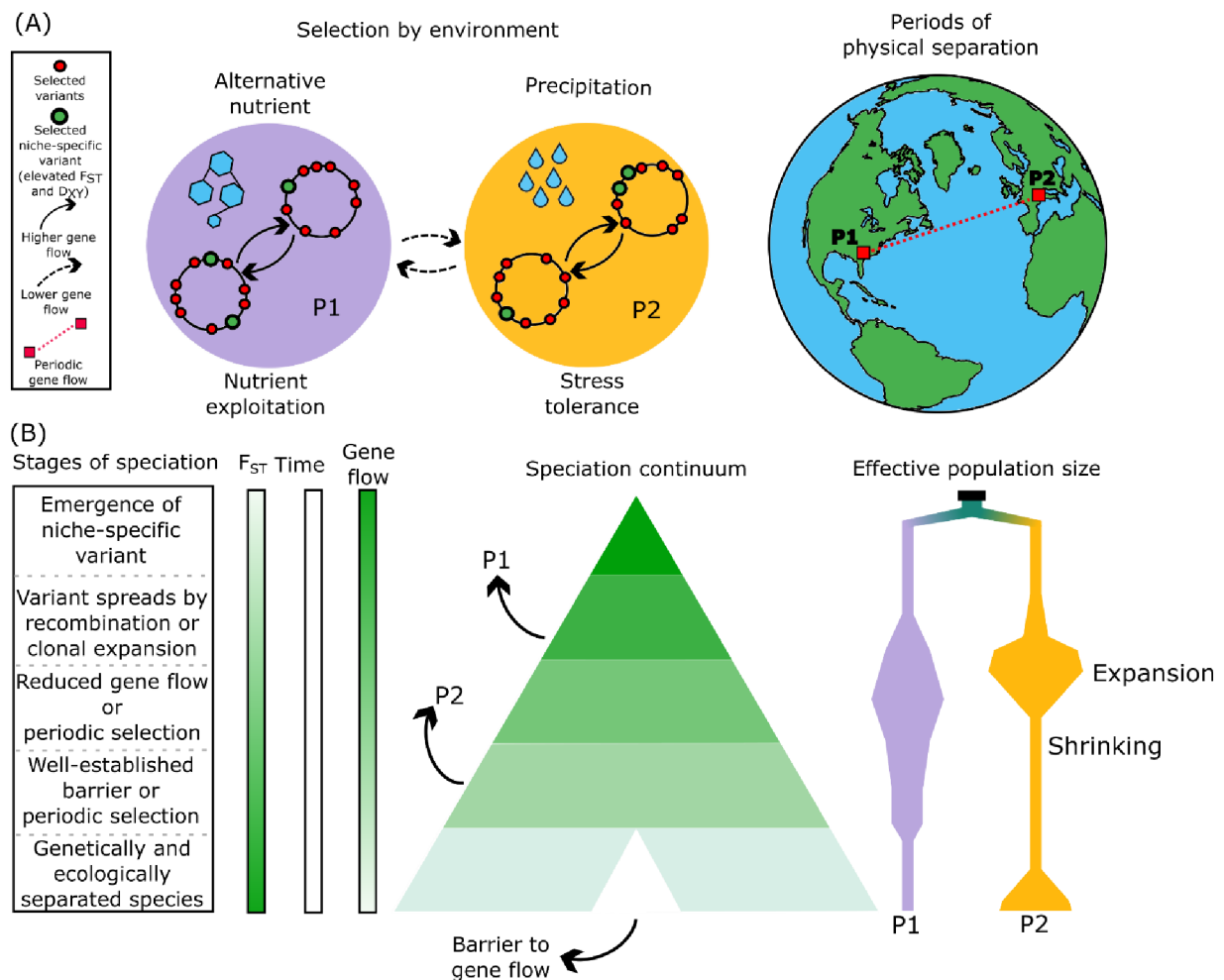
Figure 4. Speciation model in *Microcoleus*. (A) Ecological and geographic isolation as drivers of the speciation process. In this example, two *Microcoleus* populations (P1 and P2) have diverged for nutrient exploitation and stress tolerance. Selection acts on variants scattered across the genome (red circles) and also on variants within the regions of elevated F_{ST} and D_{XY} (green circles), which likely represent niche-specific adaptations. The higher intra-specific gene flow maintains them in cohesive ecological units. Novel ecological niches arose, and periodic physical separation on a large scale interchanges with periods of gene flow. Geographic and ecological barriers to gene flow cause the speciation event. (B) Divergence of two *Microcoleus* populations along the speciation continuum. Populations undergo demographic changes and diverge over time, with different mechanisms acting at different stages of speciation. Barriers to gene flow spread as the genomes become more diverged, eventually resulting in complete ecological and genetic isolation. We can position diverging populations along the speciation continuum by considering various evolutionary forces and

References

Achtman, M., & Wagner, M. (2008). Microbial diversity and the genetic nature of microbial species. *Nature Reviews Microbiology*, *6*(6), 431–440. https://doi.org/10.1038/nrmicro1872

Baas-Becking, L. G. M. (1934). Geobiologie of Inleiding tot de Milieukunde. W. P. Van Stockum & Zoon, The Hague.

Bahl, J., Lau, M. C. Y., Smith, G. J. D., Vijaykrishna, D., Cary, S. C., Lacap, D. C., Lee, C. K., Papke, R. T., Warren-Rhodes, K. A., Wong, F. K. Y., McKay, C. P., & Pointing, S. B. (2011). Ancient origins determine global biogeography of hot and cold desert cyanobacteria. *Nature Communications*, *2*(1), 163. https://doi.org/10.1038/ncomms1167

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*, *19*(5), 455–477. https://doi.org/10.1089/cmb.2012.0021

Beckmann, M., Václavík, T., Manceur, A. M., Šprtová, L., Von Wehrden, H., Welk, E., & Cord, A. F. (2014). glUV: A global UV-B radiation data set for macroecological studies. *Methods in Ecology and Evolution*, *5*(4), 372–383. https://doi.org/10.1111/2041-210X.12168

Belnap, J., & Gardner, J.S. (1993). Soil microstructure in soils of the Colorado Plateau: the role of the cyanobacterium Microcoleus vaginatus. *The Great Basin Naturalist*, *53*(1), 40-47.

Belnap, J., & Lange, O. L. (2001). Structure and functioning of biological soil crusts: a synthesis. In: J. Belnap, & O. L. Lange (Eds.), *Biological soil crusts: structure, function, and management* (471-479). Springer, Berlin, Heidelberg.

Beugin, M., Gayet, T., Pontier, D., Devillard, S., & Jombart, T. (2018). A fast likelihood solution to the genetic clustering problem. *Methods in Ecology and Evolution*, *9*(4), 1006–1016. https://doi.org/10.1111/2041-210X.12968

Blomberg, S. P., Garland, T., & Ives, A. R. (2003). Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution*, *57*(4), 717–745. https://doi.org/10.1111/j.0014-3820.2003.tb00285.x

Bobay, L. M. (2020). The prokaryotic species concept and challenges. In: H. Tettelin, & D. Medini (Eds.), *The Pangenome* (21-49). Springer, Cham.

Bobay, L.-M., & Ochman, H. (2017). Biological Species Are Universal across Life's Domains. *Genome Biology and Evolution*, *9*(3), 491–501. https://doi.org/10.1093/gbe/evx026

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bolnick, D. I., Hund, A. K., Nosil, P., Peng, F., Ravinet, M., Stankowski, S., Subramanian, S., Wolf, J. B. W., & Yukilevich, R. (2023). A multivariate view of the speciation continuum. *Evolution*, *77*(1), 318–328. https://doi.org/10.1093/evolut/qpac004

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., De Maio, N., Matschiner, M., Mendes, F. K., Müller, N. F., Ogilvie, H. A., Du Plessis, L., Popinga, A., Rambaut, A., Rasmussen, D., Siveroni, I., … Drummond, A. J. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS Computational Biology*, *15*(4), e1006650. https://doi.org/10.1371/journal.pcbi.1006650

Bouma-Gregson, K., Crits-Christoph, A., Olm, M. R., Power, M. E., & Banfield, J. F. (2022). *Microcoleus* (Cyanobacteria) form watershed-wide populations without strong gradients in population structure. *Molecular Ecology*, *31*(1), 86–103. https://doi.org/10.1111/mec.16208

Cadillo-Quiroz, H., Didelot, X., Held, N. L., Herrera, A., Darling, A., Reno, M. L., Krause, D. J., & Whitaker, R. J. (2012). Patterns of Gene Flow Define Species of Thermophilic Archaea. *PLoS Biology*, *10*(2), e1001265. https://doi.org/10.1371/journal.pbio.1001265

Casteleyn, G., Leliaert, F., Backeljau, T., Debeer, A.-E., Kotaki, Y., Rhodes, L., Lundholm, N., Sabbe, K., & Vyverman, W. (2010). Limits to gene flow in a cosmopolitan marine planktonic diatom. *Proceedings of the National Academy of Sciences*, *107*(29), 12952–12957. https://doi.org/10.1073/pnas.1001380107

Chase, A. B., Arevalo, P., Brodie, E. L., Polz, M. F., Karaoz, U., & Martiny, J. B. H. (2019). Maintenance of Sympatric and Allopatric Populations in Free-Living Terrestrial Bacteria. *MBio*, *10*(5), e02361-19. https://doi.org/10.1128/mBio.02361-19

Chen, M. H., Takeda, S. I., Yamada, H., Ishii, Y., Yamashino, T., & Mizuno, T. (2001). Characterization of the RcsC→YojN→RcsB phosphorelay signaling pathway involved in capsular synthesis in Escherichia coli. *Bioscience, biotechnology, and biochemistry*, *65*(10), 2364-2367. https://doi.org/10.1271/bbb.65.2364

Coleman, M. L., Sullivan, M. B., Martiny, A. C., Steglich, C., Barry, K., DeLong, E. F., & Chisholm, S. W. (2006). Genomic Islands and the Ecology and Evolution of *Prochlorococcus*. *Science*, *311*(5768), 1768–1770. https://doi.org/10.1126/science.1122050

Cordero, O. X., Ventouras, L.-A., DeLong, E. F., & Polz, M. F. (2012). Public good dynamics drive evolution of iron acquisition strategies in natural bacterioplankton populations. *Proceedings of the National Academy of Sciences*, *109*(49), 20059–20064. https://doi.org/10.1073/pnas.1213344109

Couradeau, E., Giraldo-Silva, A., De Martini, F., & Garcia-Pichel, F. (2019). Spatial segregation of the biological soil crust microbiome around its foundational cyanobacterium, Microcoleus vaginatus, and the formation of a nitrogen-fixing cyanosphere. *Microbiome*, *7*(1), 55. https://doi.org/10.1186/s40168-019-0661-2

Coyne, J. A., & Orr, H. A. (1989). Patterns of speciation in Drosophila. *Evolution*, *43*(2), 362–381. https://doi.org/10.1111/j.1558-5646.1989.tb04233.x

Coyne, J. A., & Orr, H. A. (2004). Speciation. Sinauer Associates.

Croucher, N. J., Page, A. J., Connor, T. R., Delaney, A. J., Keane, J. A., Bentley, S. D., Parkhill, J., & Harris, S. R. (2015). Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Research*, *43*(3), e15–e15. https://doi.org/10.1093/nar/gku1196

Dapper, A. L., & Payseur, B. A. (2017). Connecting theory and data to understand recombination rate evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1736), 20160469. https://doi.org/10.1098/rstb.2016.0469

Den Bakker, H. C., Cummings, C. A., Ferreira, V., Vatta, P., Orsi, R. H., Degoricija, L., Barker, M., Petrauskene, O., Furtado, M. R., & Wiedmann, M. (2010). Comparative genomics of the bacterial genus Listeria: Genome evolution is characterized by limited gene acquisition and limited gene loss. *BMC Genomics*, *11*(1), 688. https://doi.org/10.1186/1471-2164-11-688

Dinno, A. (2017). Package dunn.test. Retrieved from: https://CRAN.R-project.org/package=dunn.test

Drummond, A. J. (2005). Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences. *Molecular Biology and Evolution*, *22*(5), 1185–1192. https://doi.org/10.1093/molbev/msi103

Drummond, A. J., & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, *7*(1), 214. https://doi.org/10.1186/1471-2148-7-214

Dunn, O. J. (1964). Multiple Comparisons Using Rank Sums. *Technometrics*, *6*(3), 241–252. https://doi.org/10.1080/00401706.1964.10490181

Dvořák, P. (2017) Genome-wide analysis of cyanobacterial evolution: the example of Synechococcus. In: D. A., Los (Ed.), *Cyanobacteria: omics and manipulation* (35-54). Caister Academic Press, United Kingdom

Dvořák, P., Hašler, P., & Poulíčková, A. (2012). Phylogeography of the Microcoleus vaginatus (Cyanobacteria) from Three Continents – A Spatial and Temporal Characterization. *PLoS ONE*, *7*(6), e40153. https://doi.org/10.1371/journal.pone.0040153

Dvořák, P., Jahodářová, E., Stanojković, A., Skoupý, S., & Casamatta, D. A. (2023). Population genomics meets the taxonomy of cyanobacteria. *Algal Research*, *72*, 103128. https://doi.org/10.1016/j.algal.2023.103128

Dvořák, P., Poulíčková, A., Hašler, P., Belli, M., Casamatta, D. A., & Papini, A. (2015). Species concepts and speciation factors in cyanobacteria, with connection to the problems of diversity and classification. *Biodiversity and Conservation*, *24*(4), 739–757. https://doi.org/10.1007/s10531-015-0888-6

Egan, S. P., Ragland, G. J., Assour, L., Powell, T. H. Q., Hood, G. R., Emrich, S., Nosil, P., & Feder, J. L. (2015). Experimental evidence of genome-wide impact of ecological selection during early stages of speciation-with-gene-flow. *Ecology Letters*, *18*(8), 817–825. https://doi.org/10.1111/ele.12460

Elliott, D. R., Thomas, A. D., Strong, C. L., & Bullard, J. (2019). Surface Stability in Drylands Is Influenced by Dispersal Strategy of Soil Bacteria. *Journal of Geophysical Research: Biogeosciences*, *124*(11), 3403–3418. https://doi.org/10.1029/2018JG004932

Emms, D. M., & Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*(1), 238. https://doi.org/10.1186/s13059-019-1832-y

Eronen, J. T., Fortelius, M., Micheels, A., Portmann, F. T., Puolamaki, K., & Janis, C. M. (2012). Neogene aridification of the Northern Hemisphere. *Geology*, *40*(9), 823–826. https://doi.org/10.1130/G33147.1

Feder, J. L., Egan, S. P., & Nosil, P. (2012). The genomics of speciation-with-gene-flow. *Trends in Genetics*, *28*(7), 342–350. https://doi.org/10.1016/j.tig.2012.03.009

Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International journal of climatology*, *37*(12), 4302-4315. https://doi.org/10.1002/joc.5086

Finlay, B. J. (2002). Global Dispersal of Free-Living Microbial Eukaryote Species. *Science*, *296*(5570), 1061–1063. https://doi.org/10.1126/science.1070710

Foster, P. L. (2005). Stress responses and genetic variation in bacteria. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, *569*(1–2), 3–11. https://doi.org/10.1016/j.mrfmmm.2004.07.017

Fraser, C., Hanage, W. P., & Spratt, B. G. (2007). Recombination and the Nature of Bacterial Speciation. *Science*, *315*(5811), 476–480. https://doi.org/10.1126/science.1127573

Garcia-Pichel, F., & Wojciechowski, M. F. (2009). The Evolution of a Capacity to Build Supra-Cellular Ropes Enabled Filamentous Cyanobacteria to Colonize Highly Erodible Substrates. *PLoS ONE*, *4*(11), e7801. https://doi.org/10.1371/journal.pone.0007801

Garrison, E., & Marth, G. (2012). *Haplotype-based variant detection from short-read sequencing* (arXiv:1207.3907). arXiv. http://arxiv.org/abs/1207.3907

Grant, W. S. (2015). Problems and Cautions With Sequence Mismatch Analysis and Bayesian Skyline Plots to Infer Historical Demography. *Journal of Heredity*, *106*(4), 333–346. https://doi.org/10.1093/jhered/esv020

Haberl, H., Erb, K. H., Krausmann, F., Gaube, V., Bondeau, A., Plutzar, C., Gingrich, S., Lucht, W., & Fischer-Kowalski, M. (2007). Quantifying and mapping the human appropriation of net primary production in earth's terrestrial ecosystems. *Proceedings of the National Academy of Sciences*, *104*(31), 12942–12947. https://doi.org/10.1073/pnas.0704243104

Hadany, L., & Beker, T. (2003). On the Evolutionary Advantage of Fitness-Associated Recombination. *Genetics*, *165*(4), 2167–2179. https://doi.org/10.1093/genetics/165.4.2167

Hadfield, J., Croucher, N. J., Goater, R. J., Abudahab, K., Aanensen, D. M., & Harris, S. R. (2018). Phandango: An interactive viewer for bacterial population genomics. *Bioinformatics*, *34*(2), 292–293. https://doi.org/10.1093/bioinformatics/btx610

Harmon, L., Weir, J., Brock, C., Glor, R., Challenger, W., Hunt, G., FitzJohn, R., Pennell, M., Slater, G., Brown, J., & Uyeda, J. (2015). Package 'geiger'. Retrieved from: https:// CRAN.R-project.org/package=geiger

Hašler, P., Dvořák, P., Johansen, J. R., Kitner, M., Ondřej, V., & Poulíčková, A. (2012). Morphological and molecular study of epipelic filamentous genera Phormidium, Microcoleus and Geitlerinema (Oscillatoriales, Cyanophyta/Cyanobacteria). *Fottea*, *12*(2), 341–356. https://doi.org/10.5507/fot.2012.024

Hays, J. D., Imbrie, J., & Shackleton, N. J. (1976). Variations in the Earth's Orbit: Pacemaker of the Ice Ages: For 500,000 years, major climatic changes have followed variations in obliquity and precession. *science*, *194*(4270), 1121-1132.

Held, N. L., Herrera, A., Cadillo-Quiroz, H., & Whitaker, R. J. (2010). CRISPR Associated Diversity within a Population of Sulfolobus islandicus. *PLoS ONE*, *5*(9), e12988. https://doi.org/10.1371/journal.pone.0012988

Helfrich, P., Rieb, E., Abrami, G., Lücking, A., & Mehler, A. (2018). TreeAnnotator: Versatile Visual Annotation of Hierarchical Text Relations. *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018).*

Herbert, T. D., Lawrence, K. T., Tzanova, A., Peterson, L. C., Caballero-Gill, R., & Kelly, C. S. (2016). Late Miocene global cooling and the rise of modern ecosystems. *Nature Geoscience*, *9*(11), 843–847. https://doi.org/10.1038/ngeo2813

Hijmans, R. J., Williams, E., Vennes, C., & Hijmans, M. R. J. (2017). Package 'geosphere'. *Spherical trigonometry*, *1*(7), 1-45.

Huang, J., Yu, H., Guan, X., Wang, G., & Guo, R. (2016). Accelerated dryland expansion under climate change. *Nature Climate Change*, *6*(2), 166–171. https://doi.org/10.1038/nclimate2837

Hugenholtz, P., Skarshewski, A., & Parks, D. H. (2016). Genome-Based Microbial Taxonomy Coming of Age. *Cold Spring Harbor Perspectives in Biology*, *8*(6), a018085. https://doi.org/10.1101/cshperspect.a018085

Jarmer, H., Berka, R., Knudsen, S., & Saxild, H. H. (2002). Transcriptome analysis documents induced competence of *Bacillus subtilis* during nitrogen limiting conditions. *FEMS Microbiology Letters*, *206*(2), 197–200. https://doi.org/10.1111/j.1574-6968.2002.tb11009.x

Jombart, T. (2008). *adegenet*: A R package for the multivariate analysis of genetic markers. *Bioinformatics*, *24*(11), 1403–1405. https://doi.org/10.1093/bioinformatics/btn129

Jombart, T., Kamvar, Z. N., Collins, C., Lustrik, R., Beugin, M. P., Knaus, B. J., & Jombart, M. T. (2018). Package 'adegenet'. Retrieved from: https:// CRAN.R-project.org/package=adegenet

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods*, *14*(6), 587–589. https://doi.org/10.1038/nmeth.4285

Kashtan, N., Roggensack, S. E., Rodrigue, S., Thompson, J. W., Biller, S. J., Coe, A., Ding, H., Marttinen, P., Malmstrom, R. R., Stocker, R., Follows, M. J., Stepanauskas, R., & Chisholm, S. W. (2014). Single-Cell Genomics Reveals Hundreds of Coexisting Subpopulations in Wild *Prochlorococcus*. *Science*, *344*(6182), 416–420. https://doi.org/10.1126/science.1248575

Katoh, K. (2002). MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066. https://doi.org/10.1093/nar/gkf436

Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kembel, S. W., Cowan, P. D., Helmus, M. R., Cornwell, W. K., Morlon, H., Ackerly, D. D., Blomberg, S. P., & Webb, C. O. (2010). Picante: R tools for integrating phylogenies and ecology. *Bioinformatics*, *26*(11), 1463–1464. https://doi.org/10.1093/bioinformatics/btq166

Kistler, L. (2012). Ancient DNA Extraction from Plants. In: B, Shapiro, & M. Hofreiter (Eds.) *Ancient DNA: Methods in Molecular Biology* (71-79). *Humana Press*. https://doi.org/10.1007/978-1-61779-516-9_10

Kollár, J., Poulíčková, A., & Dvořák, P. (2022). On the relativity of species, or the probabilistic solution to the species problem. *Molecular Ecology*, *31*(2), 411–418. https://doi.org/10.1111/mec.16218

Konstantinidis, K. T., & Tiedje, J. M. (2005). Genomic insights that advance the species definition for prokaryotes. *Proceedings of the National Academy of Sciences*, *102*(7), 2567–2572. https://doi.org/10.1073/pnas.0409727102

Koonin, E. V., Makarova, K. S., & Wolf, Y. I. (2021). Evolution of Microbial Genomics: Conceptual Shifts over a Quarter Century. *Trends in Microbiology*, *29*(7), 582–592. https://doi.org/10.1016/j.tim.2021.01.005

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, *47*(260), 583-621.

Lawniczak, M. K. N., Emrich, S. J., Holloway, A. K., Regier, A. P., Olson, M., White, B., Redmond, S., Fulton, L., Appelbaum, E., Godfrey, J., Farmer, C., Chinwalla, A., Yang, S.-P., Minx, P., Nelson, J., Kyung, K., Walenz, B. P., Garcia-Hernandez, E., Aguiar, M., … Besansky, N. J. (2010). Widespread Divergence Between Incipient *Anopheles gambiae* Species Revealed by Whole Genome Sequences. *Science*, *330*(6003), 512–514. https://doi.org/10.1126/science.1195755

Lee-Yaw, J. A., Grassa, C. J., Joly, S., Andrew, R. L., & Rieseberg, L. H. (2019). An evaluation of alternative explanations for widespread cytonuclear discordance in annual sunflowers (Helianthus). *New Phytologist*, *221*(1), 515-526. https://doi.org/10.1111/nph.15386

Levsen, N. D., Tiffin, P., & Olson, M. S. (2012). Pleistocene Speciation in the Genus Populus (Salicaceae). *Systematic Biology*, *61*(3), 401. https://doi.org/10.1093/sysbio/syr120

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer research*, *27*(2_Part_1), 209-220.

Martin, S. H., Dasmahapatra, K. K., Nadeau, N. J., Salazar, C., Walters, J. R., Simpson, F., Blaxter, M., Manica, A., Mallet, J., & Jiggins, C. D. (2013). Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Research*, *23*(11), 1817–1828. https://doi.org/10.1101/gr.159426.113

Mayr, E. (1942). Systematics and the Origin of Species. New York, NY: Columbia University Press.

Meyer, M., & Kircher, M. (2010). Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. *Cold Spring Harbor Protocols*, *2010*(6), pdb.prot5448. https://doi.org/10.1101/pdb.prot5448

Michel, A. P., Sim, S., Powell, T. H. Q., Taylor, M. S., Nosil, P., & Feder, J. L. (2010). Widespread genomic divergence during sympatric speciation. *Proceedings of the National Academy of Sciences*, *107*(21), 9724–9729. https://doi.org/10.1073/pnas.1000939107

Mirabello, L., & Conn, J. E. (2008). Population analysis using the nuclear white gene detects Pliocene/Pleistocene lineage divergence within Anopheles nuneztovari in South America. *Medical and Veterinary Entomology*, *22*(2), 109–119. https://doi.org/10.1111/j.1365-2915.2008.00731.x

Mirarab, S., & Warnow, T. (2015). ASTRAL-II: Coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics*, *31*(12), i44–i52. https://doi.org/10.1093/bioinformatics/btv234

Murik, O., Oren, N., Shotland, Y., Raanan, H., Treves, H., Kedem, I., Keren, N., Hagemann, M., Pade, N., & Kaplan, A. (2017). What distinguishes cyanobacteria able to revive after desiccation from those that cannot: The genome aspect: Desiccation Resistance Genes in Cyanobacteria. *Environmental Microbiology*, *19*(2), 535–550. https://doi.org/10.1111/1462-2920.13486

Nelson, C., Giraldo-Silva, A., & Garcia-Pichel, F. (2021). A symbiotic nutrient exchange within the cyanosphere microbiome of the biocrust cyanobacterium, Microcoleus vaginatus. *The ISME Journal*, *15*(1), 282–292. https://doi.org/10.1038/s41396-020-00781-1

Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274. https://doi.org/10.1093/molbev/msu300

Nies, F., Mielke, M., Pochert, J., & Lamparter, T. (2020). Natural transformation of the filamentous cyanobacterium Phormidium lacuna. *PLOS ONE*, *15*(6), e0234440. https://doi.org/10.1371/journal.pone.0234440

Nosil, P. (2012). *Ecological speciation*. Oxford University Press.

Nosil, P., & Feder, J. L. (2012). Genomic divergence during speciation: Causes and consequences. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1587), 332–342. https://doi.org/10.1098/rstb.2011.0263

Nosil, P., Feder, J. L., Flaxman, S. M., & Gompert, Z. (2017). Tipping points in the dynamics of speciation. *Nature Ecology & Evolution*, *1*(2), 0001. https://doi.org/10.1038/s41559-016-0001

Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., Mcglinn, D. & Wagner, H. (2016). Vegan: Community ecology package. Retrieved from https:// CRAN.R-project.org/package=vegan

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., Fookes, M., Falush, D., Keane, J. A., & Parkhill, J. (2015). Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, *31*(22), 3691–3693. https://doi.org/10.1093/bioinformatics/btv421

Pagel, M. (1999). Inferring the historical patterns of biological evolution. *Nature*, *401*(6756), 877–884. https://doi.org/10.1038/44766

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, *25*(7), 1043–1055. https://doi.org/10.1101/gr.186072.114

Pennisi, E. (2014). Disputed islands. *Science*, *345*(6197), 611–613. https://doi.org/10.1126/science.345.6197.611

Pérez-Carrascal, O. M., Terrat, Y., Giani, A., Fortin, N., Greer, C. W., Tromas, N., & Shapiro, B. J. (2019). Coherence of Microcystis species revealed through population genomics. *The ISME Journal*, *13*(12), 2887–2900. https://doi.org/10.1038/s41396-019-0481-1

Pfeifer, B., Wittelsbürger, U., Ramos-Onsins, S. E., & Lercher, M. J. (2014). PopGenome: An Efficient Swiss Army Knife for Population Genomic Analyses in R. *Molecular Biology and Evolution*, *31*(7), 1929–1936. https://doi.org/10.1093/molbev/msu136

Polz, M. F., Alm, E. J., & Hanage, W. P. (2013). Horizontal gene transfer and the evolution of bacterial and archaeal population structure. *Trends in Genetics*, *29*(3), 170–175. https://doi.org/10.1016/j.tig.2012.12.006

Posada, D., & Crandall, K. A. (1998). MODELTEST: Testing the model of DNA substitution. *Bioinformatics*, *14*(9), 817–818. https://doi.org/10.1093/bioinformatics/14.9.817

Prăvălie, R. (2016). Drylands extent and environmental issues. A global approach. *Earth-Science Reviews*, *161*, 259–278. https://doi.org/10.1016/j.earscirev.2016.08.003

QGIS.org, %Y. QGIS Geographic Information System. QGIS Association. http://www.qgis.org

Rajeev, L., Da Rocha, U. N., Klitgord, N., Luning, E. G., Fortney, J., Axen, S. D., Shih, P. M., Bouskill, N. J., Bowen, B. P., Kerfeld, C. A., Garcia-Pichel, F., Brodie, E. L., Northen, T. R., & Mukhopadhyay, A. (2013). Dynamic cyanobacterial response to hydration and dehydration in a desert biological soil crust. *The ISME Journal*, *7*(11), 2178–2191. https://doi.org/10.1038/ismej.2013.83

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., & Suchard, M. A. (2018). Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Systematic Biology*, *67*(5), 901–904. https://doi.org/10.1093/sysbio/syy032

Redfield, R. J. (1993). Genes for Breakfast: The Have-Your-Cake and-Eat-lt-Too of Bacterial Transformation. *Journal of Heredity*, *84*(5), 400–404. https://doi.org/10.1093/oxfordjournals.jhered.a111361

Renaut, S., Owens, G. L., & Rieseberg, L. H. (2014). Shared selective pressure and local genomic landscape lead to repeatable patterns of genomic divergence in sunflowers. *Molecular Ecology*, *23*(2), 311–324. https://doi.org/10.1111/mec.12600

Reno, M. L., Held, N. L., Fields, C. J., Burke, P. V., & Whitaker, R. J. (2009). Biogeography of the *Sulfolobus islandicus* pan-genome. *Proceedings of the National Academy of Sciences*, *106*(21), 8605–8610. https://doi.org/10.1073/pnas.0808945106

Ribeiro, K. F., Ferrero, A. P., Duarte, L., Turchetto-Zolet, A. C., & Crossetti, L. O. (2020). Comparative phylogeography of two free-living cosmopolitan cyanobacteria: Insights on biogeographic and latitudinal distribution. *Journal of Biogeography*, *47*(5), 1106–1118. https://doi.org/10.1111/jbi.13785

Richter, M., & Rosselló-Móra, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. *Proceedings of the National Academy of Sciences*, *106*(45), 19126–19131. https://doi.org/10.1073/pnas.0906412106

Rosselló-Mora, R., & Amann, R. (2001). The species concept for prokaryotes. *FEMS Microbiology Reviews*, *25*(1), 39–67. https://doi.org/10.1016/S0168-6445(00)00040-1

Roux, C., Fraïsse, C., Romiguier, J., Anciaux, Y., Galtier, N., & Bierne, N. (2016). Shedding Light on the Grey Zone of Speciation along a Continuum of Genomic Divergence. *PLOS Biology*, *14*(12), e2000234. https://doi.org/10.1371/journal.pbio.2000234

Schoustra, S., Rundle, H. D., Dali, R., & Kassen, R. (2010). Fitness-Associated Sexual Reproduction in a Filamentous Fungus. *Current Biology*, *20*(15), 1350–1355. https://doi.org/10.1016/j.cub.2010.05.060

Seeman, T. Snippy. Available online: https://github.com/tseeman/snippy

Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068–2069. https://doi.org/10.1093/bioinformatics/btu153

Shapiro, B. J. (2018). What Microbial Population Genomics Has Taught Us About Speciation. In: M. Polz, & O. Rajora (Eds.), *Population Genomics: Microorganisms. Population Genomics.* Springer, Cham. https://doi.org/10.1007/13836_2018_10

Shapiro, B. J., & Polz, M. F. (2015). Microbial Speciation. *Cold Spring Harbor Perspectives in Biology*, *7*(10), a018143. https://doi.org/10.1101/cshperspect.a018143

Shapiro, B. J., Friedman, J., Cordero, O. X., Preheim, S. P., Timberlake, S. C., Szabó, G., Polz, M. F., & Alm, E. J. (2012). Population Genomics of Early Events in the Ecological Differentiation of Bacteria. *Science*, *336*(6077), 48–51. https://doi.org/10.1126/science.1218198

Škaloud, P., Škaloudová, M., Doskočilová, P., Kim, J. I., Shin, W., & Dvořák, P. (2019). Speciation in protists: Spatial and ecological divergence processes cause rapid species diversification in a freshwater chrysophyte. *Molecular Ecology*, *28*(5), 1084–1095. https://doi.org/10.1111/mec.15011

Stamatakis, A. (2006). RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, *22*(21), 2688–2690. https://doi.org/10.1093/bioinformatics/btl446

Stankowski, S., & Ravinet, M. (2021). Defining the speciation continuum. *Evolution*, *75*(6), 1256–1273. https://doi.org/10.1111/evo.14215

Stanojković, A., Skoupý, S., Hašler, P., Poulíčková, A., & Dvořák, P. (2022a). Geography and climate drive the distribution and diversification of the cosmopolitan cyanobacterium Microcoleus (Oscillatoriales, Cyanobacteria). *European Journal of Phycology*, *57*(4), 396-405. https://doi.org/10.1080/09670262.2021.2007420

Stanojković, A., Skoupý, S., Škaloud, P., & Dvořák, P. (2022b). High genomic differentiation and limited gene flow indicate recent cryptic speciation within the genus Laspinema (cyanobacteria). *Frontiers in Microbiology*, *13*, 977454. https://doi.org/10.3389/fmicb.2022.977454

Sun, J., & Windley, B. F. (2015). Onset of aridification by 34 Ma across the Eocene-Oligocene transition in Central Asia. *Geology*, *43*(11), 1015–1018. https://doi.org/10.1130/G37165.1

Takeda, S., Fujisawa, Y., Matsubara, M., Aiba, H., & Mizuno, T. (2001). A novel feature of the multistep phosphorelay in Escherichia coli: A revised model of the RcsC→YojN→RcsB signalling pathway implicated in capsular synthesis and swarming behaviour. *Molecular Microbiology*, *40*(2), 440–450. https://doi.org/10.1046/j.1365-2958.2001.02393.x

Turner, T. L., & Hahn, M. W. (2010). Genomic islands *of* speciation or genomic islands *and* speciation? *Molecular Ecology*, *19*(5), 848–850. https://doi.org/10.1111/j.1365-294X.2010.04532.x

Vos, M., & Didelot, X. (2009). A comparison of homologous recombination rates in bacteria and archaea. *The ISME Journal*, *3*(2), 199–208. https://doi.org/10.1038/ismej.2008.93

Vos, M., Wolf, A. B., Jennings, S. J., & Kowalchuk, G. A. (2013). Micro-scale determinants of bacterial diversity in soil. *FEMS Microbiology Reviews*, *37*(6), 936–954. https://doi.org/10.1111/1574-6976.12023

Ward, D. M., Cohan, F. M., Bhaya, D., Heidelberg, J. F., Kühl, M., & Grossman, A. (2008). Genomics, environmental genomics and the issue of microbial species. *Heredity*, *100*(2), 207–219. https://doi.org/10.1038/sj.hdy.6801011

Weir, J. T., & Schluter, D. (2004). Ice sheets promote speciation in boreal birds. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *271*(1551), 1881–1887. https://doi.org/10.1098/rspb.2004.2803

Whelan, F. J., Rusilowicz, M., & McInerney, J. O. (2020). Coinfinder: Detecting significant associations and dissociations in pangenomes. *Microbial Genomics*, *6*(3). https://doi.org/10.1099/mgen.0.000338

Whitaker, R. J., & Banfield, J. F. (2006). Population genomics in natural microbial communities. *Trends in Ecology & Evolution*, *21*(9), 508–516. https://doi.org/10.1016/j.tree.2006.07.001

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag. Retrieved from: https://CRAN.R-project.org/package=ggplot2

Wickham, H. (2017). reshape2: flexibly reshape data: a reboot of the reshape package. Retrieved from: https:// CRAN.R-project.org/package=reshape2

Wielgoss, S., Didelot, X., Chaudhuri, R. R., Liu, X., Weedall, G. D., Velicer, G. J., & Vos, M. (2016). A barrier to homologous recombination between sympatric strains of the cooperative soil bacterium Myxococcus xanthus. *The ISME Journal*, *10*(10), 2468–2477. https://doi.org/10.1038/ismej.2016.34

Wu, C.-I. (2001). The genic view of the process of speciation: Genic view of the process of speciation. *Journal of Evolutionary Biology*, *14*(6), 851–865. https://doi.org/10.1046/j.1420-9101.2001.00335.x

Wu, Y.-W., Simmons, B. A., & Singer, S. W. (2016). MaxBin 2.0: An automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, *32*(4), 605–607. https://doi.org/10.1093/bioinformatics/btv638

Yu, Y., Harris, A. J., Blair, C., & He, X. (2015). RASP (Reconstruct Ancestral State in Phylogenies): A tool for historical biogeography. *Molecular Phylogenetics and Evolution*, *87*, 46–49. https://doi.org/10.1016/j.ympev.2015.03.008

Zachos, J., Pagani, M., Sloan, L., Thomas, E., & Billups, K. (2001). Trends, rhythms, and aberrations in global climate 65 Ma to present. *Science*, *292*(5517), 686-693.

Zhang, C., Rabiee, M., Sayyari, E., & Mirarab, S. (2018). ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, *19*(S6), 153. https://doi.org/10.1186/s12859-018-2129-y

Palacký University Olomouc

Faculty of Science

Department of Botany

Genome variability in the evolution of microorganisms

Cyanobacteria: Painting the speciation continuum

Summary of the Doctoral Thesis

By

Mgr. Aleksandar Stanojković

Supervisor: doc. Mgr. Petr Dvořák, Ph.D.

Olomouc 2023

This doctoral thesis was carried out within the framework of internal doctoral studies at the Department of Botany, Faculty of Science, Palacký University Olomouc, from 2018-2023.

Aspirant: **Mgr. Aleksandar Stanojković**

Department of Botany, Faculty of Science, Palacký University Olomouc

Supervisor: **doc. Mgr. Petr Dvořák, Ph.D.**

Department of Botany, Faculty of Science, Palacký University Olomouc

Opponents: **prof. Jeffrey R. Johansen, Ph.D.**

Department of Biology, John Carroll University, USA

doc. RNDr. Linda Nedbalová, Ph.D.

Department of Ecology, Faculty of Science, Charles University

prof. Ing. Petr Smýkal, Ph.D.

Department of Botany, Faculty of Science, Palacký University Olomouc

The summary of the doctoral thesis has been written by the Department of Botany, Faculty of Science, Palacký University Olomouc.

The summary of the doctoral thesis was sent out on ………………………………………

The oral defense of the thesis takes place in front of the Commission for doctoral thesis for the Study Subject Botany in the conference room at the Department of Botany, Faculty of Science, Palacký University Olomouc, Olomouc – Holice, Šlechtitelů 27, on ……………………………

The doctoral thesis is available in the Library of Biological Department of the Faculty of Science at Palacký University Olomouc, Olomouc – Holice, Šlechtitelů 27.

prof. Ing. Aleš Lebeda, DrSc.

Chairman of the Commission for the Ph.D.

Theses for Study Subject Botany

Faculty of Science, Palacký University Olomouc

Contents

1. Introduction

Cyanobacteria are one of the oldest photosynthetic organisms. Although the time of their origin is still being questioned, it is thought that they are almost 3.5 billion years old (Schopf & Packer, 1987). During the billions of years of dominance on the Earth, they spread and inhabited myriad environments such as marine, freshwater, and terrestrial, but also extreme ones (like hot and cold deserts) (Büdel, 2011). The secret of cyanobacterial endurance is woven into their genetic material, which underwent many changes over time and space. Understanding the changes in genetic variability is paramount to discovering the mechanisms governing species emergence and, in that way, life diversity.

Genetic variation is fundamental for the onset of speciation, and processes governing it are continuous rather than sudden (Shapiro, 2018; Koonin & Wolf, 2012). Mutations and gene flow mediated by homology-(in)dependent mechanisms are motors introducing genetic changes to microbial populations, and the interplay of selection, genetic drift, and recombination dictate the magnitude of divergence between the populations (Reno et al., 2009; Shapiro et al., 2012).

The effect of geographic isolation was thought to be negligible in microbes, which is explained by the notion of microbial ubiquity coined by Baas-Becking (1934). Although some appear to be cosmopolitan (e.g., *Microcystis*; van Gremberghe et al., 2011), restricted distribution has been reported in others (e.g., *Cylindrospermopsis raciborskii*; Ribeiro et al., 2020; *M. vaginatus*; Dvořák et al., 2012; Papers I and III). The significant congruence between geographical distance and phylogenetic relationships provided evidence of possible allopatric speciation in microbes. Further, many have highlighted the importance of environmental factors, such as light (e.g., *Prochlorococcus*, Kashtan et al., 2014; *Laspinema*, Paper II), temperature (e.g., *Microcoleus*, Paper I), or nutrient uptake (e.g., *Vibrio*, Shapiro et al., 2012), as drivers of microbial diversification. While at the early stages of speciation, ecological heterogeneity drives the emergence of niche-specific genes (i.e., adaptive genes), in the later stages, it ensures the establishment of barriers to gene flow. Thus, both isolation by distance and isolation by the environment are significant contributors to the species' emergence in cyanobacteria.

A recent rise of the idea that speciation does not have a distinct starting- or endpoint but is instead a dynamic and continuous process signifies a profound shift in how we conceptualize and observe

species (Nosil & Feder, 2012; Shapiro & Polz, 2014; Kollár et al., 2022). Hence, we can regard speciation as a continuum, which according to Stankowski & Ravinet (2021), is "a continuum of reproductive isolation"; or in microbes, it may be "a continuum of a barrier to gene flow". Despite the mainly asexual nature of prokaryotes, gene flow still frequently occurs among them, with the potential to act well beyond species boundaries (Bobay & Ochman, 2017). Continual gene flow ensures that populations merge and divide over time, resulting in incipient species at intermediate stages of speciation ("grey speciation zone"; Roux et al., 2016; Kollár et al., 2022). Expressly, incipient species exchange genes at various rates and may take millions of years to diverge (Coyne & Orr, 2004). Nevertheless, restricted gene flow between populations keeps them separate and promotes permanent genetic, ecological, and, ultimately, morphological divergence, resulting in what species are (Shapiro et al., 2012; Kollár et al., 2022).

Perhaps, the most promising way to study speciation is through the prism of population genomics, phylogenomics, and gene flow, as outlined by several authors (Shapiro & Polz, 2015; Bobay & Ochman, 2017; Arevalo et al., 2018). The main idea behind such a multifaceted approach is to study microbial species as close as possible to reality, where they can be anywhere on the spectrum of an early to a full divergence. Considering the evolutionary forces mentioned previously with the estimates of the extent of gene flow between and within populations allows us to paint the speciation continuum and position incipient species along it. This is important because it gives us insight into the progress of speciation as well as into the barrier to gene flow which maintains them genetically and ecologically distinct (Stankowski & Ravinet, 2021). Detailed speciation stages are in the thesis.

This thesis focused on filamentous cyanobacteria thriving in soil crusts – *Microcoleus* and *Laspinema*. *Microcoleus* is a well-known genus first described in the 19th century by Desmazières, which was later redefined by Gomont (1892). It is a terrestrial, bundle-forming cyanobacterium occurring in biocrusts (Garcia-Pichel & Wojciechowski, 2009). *Microcoleus* can serve as a perfect model system for studying distribution and divergence patterns in cyanobacteria. The first reason is its representation in all parts of the world and includes many recently diverged groups. Then, it is easily cultivated and distinguished from other cyanobacteria by morphological features. *Laspinema* is a newly erected genus within the recently established family of *Laspinemaceae* (Heidari et al., 2018). The genus occurs in diverse terrestrial habitats, from ones with extremely

high temperatures and radioactive contamination to microbial mats of puddles, e.g., hyposaline puddles (Dadheech et al., 2013), saline ponds (Casamatta et al., 2005), thermal springs and muds (Heidari et al., 2018; Duval et al., 2020). This genus is still undersampled and consists of recently diversified species, making it an ideal organism for delimitation and evolutionary studies.

2. Aims

The underlying goal of my thesis has been to characterize evolutionary forces governing genetic variation that might lead to the diversification of microbial species from a global to a local scale. Specifically, I have studied speciation patterns by focusing on genome-wide signatures of diversity and divergence. My focus was on cryptic filamentous cyanobacteria - *Microcoleus* and *Laspinema*.

For **Paper I**, we aimed to establish a large collection culture of globally distributed strains on a population level, a rarely used approach for cyanobacteria. I sequenced 16S rRNA and 16S-23S ITS to investigate drivers of diversification and biogeographic distribution of *Microcoleus*. Additionally, the inferred global population structure of *Microcoleus* served as a priori information for subsequent genome analyses in Paper III.

In **Paper II**, I shifted focus to explore patterns of genetic variability and evolutionary mechanisms generating it on a local scale. Specifically, I scanned the genome for signatures of selection and recombination to seek specific adaptive genes responsible for the differentiation of diverging species of *Laspinema*.

Using the population genomics approach developed in Paper II and the dataset from Paper I, I sought to investigate local and global patterns of genomic diversity and the emergence of the continuum of *Microcoleus* species. I examined genome-wide inter- and intraspecific signatures of local selection and homologous recombination (HR) to provide a better understanding of the potential cause of species' divergence. Ultimately, I estimated whether *Microcoleus* lineages are on separate evolutionary trajectories, i.e., placed them on the speciation continuum (**Paper III**).

3. Results

PAPER I. Geography and climate drive the evolutionary and biogeographic patterns in *Microcoleus*

Despite the extensive diversity of bacteria and their importance to the fundamental functioning of terrestrial ecosystems, their distribution and diversification patterns are still not fully known. This study is focused on investigating the phylogeographic structure and the underlying drivers of diversification among populations of the cyanobacterium *Microcoleus vaginatus*. Additionally, the inferred global population structure of *Microcoleus* served as a priori information for subsequent genome analyses in Paper III.

- A large, global collection culture of 495 closely related *Microcoleus* strains originating from all continents besides South America has been established
- Characterization of unialgal cultures based on morphology and sequencing of the 16S rRNA gene and 16S-23S ITS
- The phylogenetic reconstruction revealed 13 genetically distinct lineages of *Microcoleus* that could represent novel cyanobacterial species
- Up to four distinct lineages coexisted at the same sampling site, indicating fine differences in microniche exploitation between the lineages
- *Microcoleus* exhibited a distinct phylogeographic structure within the respective lineages
- Geographic distance and contemporary climatic conditions (temperature and precipitation) play significant roles in shaping the distribution and diversification of *Microcoleus*

PAPER II. Indications of recent, cryptic speciation within cyanobacterium *Laspinema*

The coexistence of multiple closely related but distinct lineages in sympatry appears to be a characteristic feature of free-living (cyano)bacteria (e.g., Kashtan et al., 2014). Molecular markers traditionally used in the study of cyanobacteria (16S rRNA and 16S-23S ITS) are not sensitive enough to capture fine genetic diversity between cryptic, closely related species and infer drivers of variability. In **Paper II**, we utilized the population genomics approach to explore the extent of local genomic diversity in a centimeter-scaled habitat (puddle) and the mechanisms responsible for the sympatric speciation of *Laspinema thermale*.

- Two cryptic lineages of *Laspinema* were found, and they likely represent novel cyanobacterial species
- The two species were highly genetically differentiated over the whole genome and separated by the barrier to gene flow
- In genomic regions of high divergence (elevated F_{ST}) and suppressed HR, we found 26 annotated potentially adaptive genes. They were associated with physiological processes associated with nutrient uptake (nitrogen, phosphorous, iron) as well as response to various environmental stresses and high/low light conditions
- Tentative adaptive genes exhibited a weak diversifying selection
- HR occurred more frequently within than between the two species, making the biological species concept applicable
- Overall, the diversification of recently diversified *Laspinema* species might have been affected by both genomic and ecological processes

PAPER III. Exploring the emergence of the global continuum of *Microcoleus* species

Microbial species emerge through a dynamic and continuous process called speciation, which involves the evolution of genetic and ecological differentiation across the genome (Shapiro & Polz, 2014). The often gradual acquisition of genetic variants within a microbial population is mediated by the complex coaction of evolutionary forces such as selection, gene flow, mutations, and genetic drift (Reno et al., 2009). However, understanding the relative importance of mechanisms by which genetic differences accumulate over time and subsequently lead to the emergence of new species remains unclear. **Paper III** illustrates the largest speciation continuum observed and the most comprehensive examination of genomic differentiation on a local and global scale in free-living terrestrial prokaryotes, especially cyanobacteria.

- I recreated the population structure inferred in Paper I by a careful selection of 202 closely related *Microcoleus* strains and their subsequent characterization by whole-genome sequencing

- *Microcoleus* represents a continuum of at least 13 species with varying levels of genetic and ecological differentiation

- The diversification of *Microcoleus* commenced before 29.6 million years ago in the Eocene/Oligocene, which coincided with the global expansion of arid regions and continued throughout the Miocene/Pliocene (13.7-4.7 million years ago)

- The demographic changes within *Microcoleus* were likely triggered by the global aridity and climate shifts in the Pleistocene (less than 500 thousand years ago)

- Geography and environmental conditions (precipitation, soil properties, human activity) played a significant role in *Microcoleus* diversification

- Genomic regions of elevated genetic differentiation (high F_{ST} and D_{XY}) contained 28 genes associated with stress response and biosynthesis. The genes were under the strong influence of positive and negative selection, confirming ecological differentiation

- *Microcoleus* species underwent preferential gene flow within species rather than between, underscoring the barriers to gene flow

- The speciation in microbes is driven by the coaction of periodic allopatric speciation, ecological speciation with a selection acting on loci in the accessory genome, and extensive gene flow

4. Conclusions

The broad view of this dissertation was to open the doors into the wondrous world of cyanobacteria – what factors governed the species' emergence and the global competitive success in highly multiplexed terrestrial ecosystems? Overall, the results of this thesis provide the following:

- The chapters of this thesis can be seen as a contribution toward guiding cyanobacteriologists and microbiologists to apply various approaches and concepts to delineate the microbial species, taking into consideration the relative importance of evolutionary forces to their divergence.

- Population-level sequencing revealed a vast genetic diversity of filamentous cyanobacteria, which is reflected in the discovery of at least 12 incipient species of *Microcoleus* on a global scale and at least two *Laspinema*.

- Following the patterns of genome-wide diversity and the interplay of evolutionary forces allowed us to capture the speciation continuum and order *Microcoleus* and *Laspinema* species along it. This is important for understanding what forces might have promoted or inhibited gene flow between species, which can eventually lead to the divergence of species over time. Additionally, the speciation continuum reveals the commonality of intermediate speciation stages in natural populations of cyanobacteria.

- The diversification of soil cyanobacteria *Microcoleus* and *Laspinema* is governed by evolutionary forces operating in tandem: abiotic factors (precipitation, soil properties, UV radiation, human activities), geography, gene flow, neutral processes, and selection.

- Estimating the extent and frequency of gene flow within and between incipient species helped us designate two *Laspinema* species as having a very high probability of being on separate evolutionary trajectories. The global collection of *Microcoleus* species encompassed species at various stages of separation, from early to later stages.

- Future physiological, transcriptomic, and epigenomic explorations will deepen our understanding of the adaptive genes functions and mechanisms that cyanobacteria utilize to ensure the present-day dominance in terrestrial ecosystems

In the current era of genomics dominance, we have already made critical steps forward in the research of prokaryotic evolution. Now it is time for cyanobacteria to be in the spotlight. With this PhD thesis finished, it can serve as a starting point for an even more thorough exploration of the evolutionary secrets embedded deep into the cyanobacterial genome. After all, we have much work ahead to find the answer to the "Great Question" of microbial "Life, the Universe, and Everything".

5. References

Arevalo, P., VanInsberghe, D., & Polz, M. F. 2018. A reverse ecology framework for bacteria and archaea. In: Polz, M. F. & Rajora, O. P. (Eds.) *Population Genomics: Microorganisms*. Springer, Cham, 77-96. doi: 10.1007/978-3-030-04756-6

Baas-Becking, L. G. M. 1934. Geobiologie of Inleiding tot de Milieukunde. W. P. Van Stockum & Zoon, The Hague.

Bobay, L. M. & Ochman, H. 2017. The evolution of bacterial genome architecture. *Frontiers in genetics*, 8, 72. doi: 10.3389/fgene.2017.00072

Büdel, B., 2011. Cyanobacteria: habitats and species. In: Caldwell, M. M., Heldmaier, G., Jackson, R. B., Lange, O. L., Mooney, H. A., Schulze, E. D., & Sommer, U. (Eds.) *Plant Desiccation Tolerance*, Springer, Berlin, Heidelberg, 11-21. doi: 10.1007/978-3-642-19106-0

Casamatta, D. A., Johansen, J. R., Vis, M. L., & Broadwater, S. T. 2005. Molecular and morphological characterization of ten polar and near-polar strains within the Oscillatoriales (Cyanobacteria) 1. *Journal of Phycology*, 41, 421-438. doi: 10.1111/j.1529-8817.2005.04062.x

Coyne, J. A. & Orr, H. A. 2004. *Speciation*. Sunderland, MA.

Dadheech, P. K., Glöckner, G., Casper, P., Kotut, K., Mazzoni, C. J., Mbedi, S., & Krienitz, L. 2013. Cyanobacterial diversity in the hot spring, pelagic and benthic habitats of a tropical soda lake. *FEMS Microbiology Ecology*, 85, 389-401. doi: 10.1111/1574-6941.12128

Duval, C., Hamlaoui, S., Piquet, B., Toutirais, G., Yepremian, C., Reinhart, A., Duperron, S., Marie, B., Demay, J., & Bernard, C. 2020. Characterization of cyanobacteria isolated from thermal muds of Balarucles-Bains (France) and description of a new genus and species *Pseudochroococcus couteii*. bioRxiv. doi: 10.1101/2020.12.12.422513

Dvořák, P., Hašler, P., & Poulíčková, A. 2012. Phylogeography of the *Microcoleus vaginatus* (cyanobacteria) from three continents–a spatial and temporal characterization. *PLoS One*, 7, e40153. doi: 10.1371/journal.pone.0040153

Garcia-Pichel, F. & Wojciechowski, M. F. (2009). The evolution of a capacity to build supra-cellular ropes enabled filamentous cyanobacteria to colonize highly erodible substrates. *PLoS ONE*, 4, e7801. doi: 10.1371/journal.pone.0007801

Gomont, M. 1892. Monographie des Oscillariées (Nostocacées homocystées). Annales des Sciences Naturelles, Botanique, 7, 263–368.

Heidari, F., Hauer, T., Zima, J. R. H., & Riahi, H. 2018. New simple trichal cyanobacterial taxa isolated from radioactive thermal springs. *Fottea*, 18, 137–149. doi: 10.5507/fot.2017.024

Kashtan, N., Roggensack, S. E., Rodrigue, S., Thompson, J. W., Biller, S. J., Coe, A., Ding, H., Marttinen, P., Malmstrom, R. R., Stocker, R., Follows, M. J., Stepanauskas, R., & Chisholm, S. W.2014. Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus. Science*, 344, 416-420. doi: 10.1126/science.1248575

Kollár, J., Poulíčková, A., & Dvořák, P. 2022. On the relativity of species, or the probabilistic solution to the species problem. *Molecular Ecology*, 31, 411–418. doi: 10.1111/mec.16218

Koonin, E.V. & Wolf, Y.I. 2012. Evolution of microbes and viruses: a paradigm shift in evolutionary biology?. *Frontiers in Cellular and Infection Microbiology*, 2, 119. doi: doi.org/10.3389/fcimb.2012.00119

Nosil, P. & Feder, J.L. 2012. Genomic divergence during speciation: causes and consequences. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367, 332-342. doi: 10.1098/rstb.2011.0263

Reno, M. L., Held, N. L., Fields, C. J., Burke, P. V., & Whitaker, R. J. 2009. Biogeography of the *Sulfolobus islandicus* pan-genome. *Proceedings of the National Academy of Sciences*, 106, 8605–8610. doi: 10.1073/pnas.0808945106

Ribeiro, K. F., Ferrero, A. P., Duarte, L., Turchetto-Zolet, A. C., & Crossetti, L. O. 2020. Comparative phylogeography of two free-living cosmopolitan cyanobacteria: Insights on biogeographic and latitudinal distribution. *Journal of Biogeography*, 47, 1106-1118. doi: 10.1111/jbi.13785

Roux, C., Fraisse, C., Romiguier, J., Anciaux, Y., Galtier, N., & Bierne, N. 2016. Shedding light on the grey zone of speciation along a continuum of genomic divergence. *PLoS Biology*, 14, e2000234. doi: 10.1371/journal.pbio.2000234

Schopf, J. W. & Packer, B. M. 1987. Early Archean (3.3-billion to 3.5-billion-year-old) microfossils from Warrawoona Group, Australia. *Science*, 237, 70-73. doi: 10.1126/science.11539686

Shapiro, B. J. & Polz, M. F. 2014. Ordering microbial diversity into ecologically and genetically cohesive units. *Trends in Microbiology*, 22, 235-247. doi: 10.1016/j.tim.2014.02.006

Shapiro, B. J. & Polz, M. F. 2015. Microbial speciation. *Cold Spring Harbor Perspectives in Biology*, 7, a018143. doi: 10.1101/cshperspect.a018143

Shapiro, B. J. 2018. What microbial population genomics has taught us about speciation. In: Polz, M. F. & Rajora, O. P. (Eds.) *Population Genomics: Microorganisms*. Springer, Cham, 31-47. doi: 10.1007/978-3-030-04756-6

Shapiro, B. J., Friedman, J., Cordero, O. X., Preheim, S. P., Timberlake, S. C., Szabó, G., Polz, M. F., & Alm, E. J. 2012. Population genomics of early events in the ecological differentiation of bacteria. *Science*, 336, 48–51. doi: 10.1126/science.1218198

Stankowski, S. & Ravinet, M. 2021. Defining the speciation continuum. *Evolution*, 75, 1256-1273. doi: 10.1111/evo.14215

Van Gremberghe, I., Leliaert, F., Mergeay, J., Vanormelingen, P., Van der Gucht, K., Debeer, A. E., Lacerot, G., Meester, L. D., & Vyverman, W. 2011. Lack of phylogeographic structure in the freshwater cyanobacterium *Microcystis aeruginosa* suggests global dispersal. *PLoS ONE*, 6, e19561. doi: 10.1371/journal.pone.0019561

6. Abstract

Genetic variation determines the potential for species to evolve. The complex interplay of evolutionary forces, such as gene flow, mutations, recombination, and selection, shape the extent of variability within a species. The concerted action of these forces dictates the degree of genetic differentiation among different microbial taxa as speciation unfolds, ultimately resulting in their divergence. However, the lack of adequate species concept, promiscuous gene exchange, and high cryptic diversity of microbes hampered the speciation study in cyanobacteria. In this thesis, I studied patterns of speciation in non-model free-living soil cyanobacteria *Microcoleus* and *Laspinema* on a local and global scale by searching for genome-wide hallmarks of ongoing differentiation and divergence. Employing population-level sampling from all continents besides South America, 500 closely related cyanobacterial strains were obtained. They were characterized based on morphology, sequenced markers 16S rRNA and 16S-23S ITS, and genomes (210 strains), which were subsequently used for the population genomics analyses. We found at least 12 distinct species at different points along a continuum of divergence in the global collection of *Microcoleus*, with up to four coexisting in sympatry. A significant influence of abiotic environmental factors (e.g., soil, climate, UV light), homologous recombination, and geography contributed to the diversification of *Microcoleus* in terrestrial soil systems. Furthermore, an ongoing divergence was captured between *Microcoleus* and *Laspinema* species in a sympatric setting. The speciation of these cyanobacteria was likely governed by adaptation to novel yet unexplored microniches in soil systems, particularly adaptation to varied light conditions and stress stimuli. In aggregate, genome-wide signatures of genetic differentiation, homologous recombination, and selection allowed us to place the diverging species on the speciation continuum, although the boundary remains blurry. Further studies are needed to unravel the functions of the genes under selection pressures as well as the nature of emerging barriers to gene flow between the species. Overall, the results of this thesis provide a deeper understanding of the genetic diversity that underlies ongoing speciation in terrestrial cyanobacteria and elucidates mechanisms contributing to the rise of new cyanobacterial species.

7. Souhrn (Summary in Czech)

Míra genetické variability určuje potenciál pro vývoj druhů. Složitá souhra evolučních sil, jako je tok genů, mutace, rekombinace a selekce, utváří rozsah variability v rámci druhu. Společné působení těchto sil určuje míru genetické diferenciace mezi různými mikrobiálními taxony v průběhu speciace, což nakonec vede k jejich divergenci. Nicméně neadekvátní druhové koncepty, častá výměna genů a vysoká kryptická diverzita mikrobů brzdily studium speciace u sinic. V této práci jsem studoval speciaci u nemodelových volně žijících půdních sinic *Microcoleus* a *Laspinema* v lokálním i globálním měřítku charakterizací probíhající diferenciace a divergence na urovní celého genomu. Na základě vzorkování na úrovni populací ze všech kontinentů, kromě Jižní Ameriky, bylo získáno 500 blízce příbuzných kmenů sinic. Byly charakterizovány na základě morfologie, sekvenovaných markerů 16S rRNA, 16S-23S ITS a genomů (210 kmenů), které byly následně použity pro analýzy populační genomiky. V globální sbírce *Microcoleus* jsme nalezli nejméně 12 odlišných druhů v různých fázích speciačního kontinua, přičemž až čtyři druhy koexistovaly v sympatrii. K diverzifikaci *Microcoleus* v terestrických půdních systémech přispěl významný vliv abiotických faktorů prostředí (např. půda, klima, UV záření), homologní rekombinace a geografická diferenciace. Dále byla zachycena probíhající divergence mezi druhy *Microcoleus* a *Laspinema* v sympatrickém prostředí. Speciace těchto sinic byla pravděpodobně poháněna adaptací na nové, dosud neprozkoumané mikroniky v půdních systémech, zejména adaptací na rozmanité světelné podmínky a stresové podněty. Genetická diferenciace, homologní rekombinace a selekce nám umožnily zařadit divergující druhy do speciačního kontinua, i když hranice zůstává nejasná. K odhalení funkcí genů pod selekčním tlakem a také povahy vznikajících bariér toku genů mezi druhy jsou zapotřebí další studie. Celkově výsledky této práce poskytují hlubší pochopení genetické rozmanitosti, která je základem probíhající speciace u suchozemských sinic, a objasňují mechanismy přispívající ke vzniku nových druhů sinic.