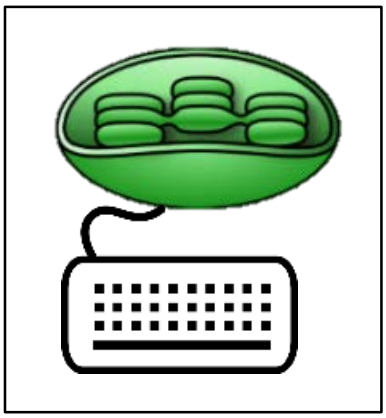
DOTTORATO DI RICERCA IN ETOLOGIA, ECOLOGIA, ANTROPOLOGIA E BIOSISTEMATICA

CICLO XXVIII

Insights into the Genome of Cyanobacteria and Plastids: tRNA Gene Content and Codon Usage

Tesi di

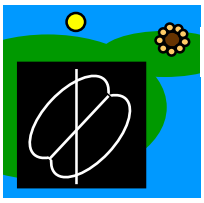
Mattia Belli



Coordinatore Prof. Alberto Ugolini

Tutor Prof. Alessio Papini

2015



UNIVERSITÀ
DEGLI STUDI
FIRENZE

DOTTORATO DI RICERCA IN ETOLOGIA, ECOLOGIA, ANTROPOLOGIA E BIOSISTEMATICA

CICLO XXVIII

COORDINATORE Prof. Ugolini Alberto

Insights into the Genome of Cyanobacteria and Plastids: tRNA Gene Content and Codon Usage

Settore Scientifico Disciplinare BIO/01

Dottorando
Dott. Belli Mattia

_____
(firma)

Tutore
Prof. Papini Alessio

_____
(firma)

Coordinatore
Prof. Ugolini Alberto

_____
(firma)

Anni 2012/2015

Acknowledgements

First of all, I would like to express my gratitude to Professor Alessio Papini for his friendly and professional supervision. I also thank the Department of Biology and all members of the Laboratory of General Botany who contributed greatly to my education and provided a friendly environment for my PhD research. In particular, I would like to thank Prof. Laura Maleci, Prof. Nicola Longo, Dr. Claudia Giuliani, Corrado Tani, Gabriele Tani, and Pietro Di Falco. Besides, I would also like to thank Prof. Giovanni Delfino for the many inspiring conversations we had together.

A special acknowledgment to my parents Ughetta and Franco, who raised me with love and gave me always everything I needed to follow my interests.

Lastly, I would like to express my sincere gratitude to my beloved Francesca for having always taken care of me with patience and love during these years

Contents

Summary

As described by the endosymbiotic theory, plastids originated from a process involving the engulfment of a cyanobacteria-like cell by a eukaryotic ancestor. Over time, plastid genomes or “plastomes” underwent a drastic reduction, which led to the conservation of a minimal expression apparatus and a set of specific coding sequences. It has been estimated that more than 95% of the plastid proteome would be expressed by the nuclear genome. On contrary, no evidence of import of rRNAs and tRNAs molecules into plastids have been reported so far.

Since a number of plastid genes encode for highly expressed proteins, their sequences could have been evolved to be efficiently translated in the plastids. Assuming that translationally optimal codons are related to the most abundant tRNA isoacceptors and that the abundance of each tRNA can be inferred by its gene copy number, a relationship should exists between codon usage and tRNA gene content.

In this PhD thesis, the tRNA gene content and its correlation with the codon usage in protein-coding genes were studied in about 600 plastid genomes and 80 cyanobacteria genomes. In order to analyse the several thousands of GenBank annotations, a set of suitable computer programs called “tRNA tools” was developed and freely distributed.

Firstly, despite the reduced number of tRNA genes, most plastid genomes can effectively translate all the codons corresponding to the 20 standard amino acids by using both the wobble and the extended wobble rules (superwobble and “two out of three”). However, a few plastome sequences do not have enough tRNAs to decode all the standard amino acids. A “defective” set of tRNA genes may be ascribed to a loss of plastid functionality or, alternatively, to the existence of tRNA import into plastids.

Data on the tRNA gene content were used to evaluate distinctive traits among different groups of eukaryotic organisms, not only among the main kingdoms of life as shown in previous studies. Noteworthy, a set of 23 tRNA genes was recognised as universally shared by most of the plastid genomes, thus extending the results obtained from *in vivo* experiments made on *Nicotiana tabacum* to a wide range of organisms.

In addition, tRNA(ACG)-Arg was found to be the only tRNA available to decode Arginine codons in Alveolata, Excavata, and Streptophyta, thus confirming the essentiality of the “two out of three” rule. Besides, a number of plastome sequences lack the genes coding for tRNA-Glu which is involved in the pyrrole biosynthesis. Although annotation errors cannot be excluded, the essentiality of tRNA-Glu in plastids could again suggest the occurrence of tRNA import mechanisms.

Finally, the hypothesis that synonymous codons related to the most abundant tRNA isoacceptors could have been preferentially conserved during evolution was tested by estimating the correlation between codon usage and tRNA gene copy number.

At a global genome-level, the results obtained in this thesis showed that the tRNA gene copy number and the codon usage generally do not correlate when only the standard pairings are taken into account. The results drastically change when the wobbling and superwobbling mechanisms are considered, showing from weak to moderate correlation among all groups. Conversely, plastomes with a reduced tRNA gene set such as *Rhizanthella gardneri* and *Selaginella moellendorfii* do not show significant correlation even when considering the superwobble rules.

The correlation was further analysed also at singe-gene level, in order to have a deeper insight into the genomes of each group of plastids. As occurred at global genome level, none of the genes analysed in the plastomes of Cyanobacteria, Rhizaria, Glaucophyta, and Streptophyta, showed significant correlation between codon usage and tRNA gene content considering only the standard pairings. In the other groups, the genes that obtained a significant correlation encode subunits of photosystem I, photosystem II, RuBisCO enzyme, ATPase, or Ribosome.

Among all the groups of plastids, the *psbA* gene resulted to have the most optimized codon composition in term of correlation with the tRNA gene content by considering only the standard pairing rules, which are supposed to be the most efficient for translating the codons. Conversely, genes with a lower rate of expression seem to be not optimized in this sense and rely more on the wobble and superwobble rules to be effectively translated.

Abbreviations

Abbreviation	Meaning
95% CI	95% confidence interval
A	Adenine
C	Cytosine
G	Guanine
T (or U)	Thymine (or Uracil)
N	Any nucleotide base
AA	Amino Acid
tRNA isoacceptors	tRNAs decoding all the codons that specify for the same amino acid

1 INTRODUCTION

1.1 Oxygenic photosynthetic prokaryotes

1.1.1 *Origin of photosynthetic autotrophy.*

The appearance of autotrophic cells is a landmark in the evolution of life on Earth. The organisms that populated our world more than 3.5 billion years ago (earliest fossils found so far) were monocellular and heterotrophic, that is their energy metabolism relied on the organic compounds already present in the environment. The competition for food pushed the heterotrophic cells towards the developing of strategies allowing them to use more efficiently the limited organic compounds available.

Autotrophic organisms broke the dependence on finding energy-rich compounds by self-producing organic molecules from inorganic molecules. Approximately 3.2 to 3.5 billion years ago, the photosynthetic organisms evolved a complex system able to capture and convert the light energy from the sun into chemical energy (De Marais 2000). The spread of photosynthetic organisms had a dramatic impact also on the heterotrophic organisms due to the production of new organic molecules. Probably the earliest forms of photosynthesis were anoxygenic and used molecular hydrogen, sulfides, or organic molecules as sources of electrons (Blankenship 2010). Later, oxygenic photosynthesis evolved in order to use water as electron donor and releasing free oxygen. As a consequence, the photosynthetic activities caused a drastic change in the atmosphere composition. This phenomenon was termed the Great Oxidation Event (GOE), which occurred 2.4-2.3 billion years ago and was fundamental for the evolution of aerobic organisms (Holland 2006).

According to the fossil record, 2 billion years ago a new step in the evolution occurred when the eukaryotic cells appeared. Eukaryotes were different from prokaryotes in many aspects, among which the compartmentation of metabolic activities in organelles such as mitochondria and chloroplasts (in plants and algae). The “endosymbiotic theory” describes the origin of mitochondria and plastids (chloroplasts and related organelles) as the result of a symbiotic relationship in which one cell, the symbiont, lives within another cell, the host. Nowadays, it is widely accepted that mitochondria originated from alpha-proteobacteria whereas chloroplasts and other plastids from cyanobacteria (Hedges et al. 2001; Yoon et al. 2004).

1.1.2 *Cyanobacteria*

Among the most abundant organisms on Earth are the group of photosynthetic prokaryotes called cyanobacteria, also known as cyanophyta, cyanophyceae, blue-green algae, and several other names. The earliest fossils showing traces of cyanobacteria are represented by stromatolites, layered deposit of calcium produced by colonies of cyanobacteria. Studies made on stromatolites suggest that cyanobacteria were already present more than 3 billion years ago (Schopf 2002). However, organisms morphologically and chemically identified as cyanobacteria with a certain degree of confidence were observed in more recent fossils dating back 2-2.5 billion years ago (Knoll 2003). Despite the ongoing discussion on their real age, it is known that since those ancient times cyanobacteria have had a great influence on the ecology and the evolution of life on our planet.

The estimated number of known cyanobacteria ranges from 2000 to 8000 species (Nabout et al. 2013), many of them living in symbiosis with other organisms. Cyanobacteria can live in many different aquatic and terrestrial environments, also under extreme conditions as the sea ice of Antarctica (Adams et al. 2012). All cyanobacteria synthetize and use chlorophyll *a* as main pigment to capture the light energy, together with accessory pigments such as carotenoids and phycobilins. The phycocyanin, which is blue, and the phycoerythrin, which is red, are the phycobilin pigments used by cyanobacteria. Cyanobacteria have been called blue-green algae due to the colours resulting from the combination of these pigments, but actually they may have different colours and, moreover cyanobacteria are no longer classified as algae. Photosynthesis takes place in the numerous layers of membranes within the cells of cyanobacteria, resembling what happens in the thylakoids of plastids (Mullineaux 2005).

A Gram-negative type of cell wall surrounds cyanobacteria but the peptidoglycan layer is considerably thicker than that of most gram-negative bacteria (Hoiczyk and Hansel 2000). Many cyanobacteria produce a mucilaginous sheath composed mainly by polysaccharides but it may also contain polypeptides. The sheath can bind groups of cells together, so cyanobacteria can be found as single cells, associated in filaments or colonies, where each cell still lives independently from the others. Moreover, intercellular connections or microplasmodesmata (name derived from the more advanced intercellular communication channels of green plants) can also be present between cells, thus resembling simple forms of multicellular organisms (Nürnberg et al. 2014; Schirrmeister et al. 2011). Cyanobacteria show a great morphological variability, ranging from unicellular to filamentous forms (Figure 1).

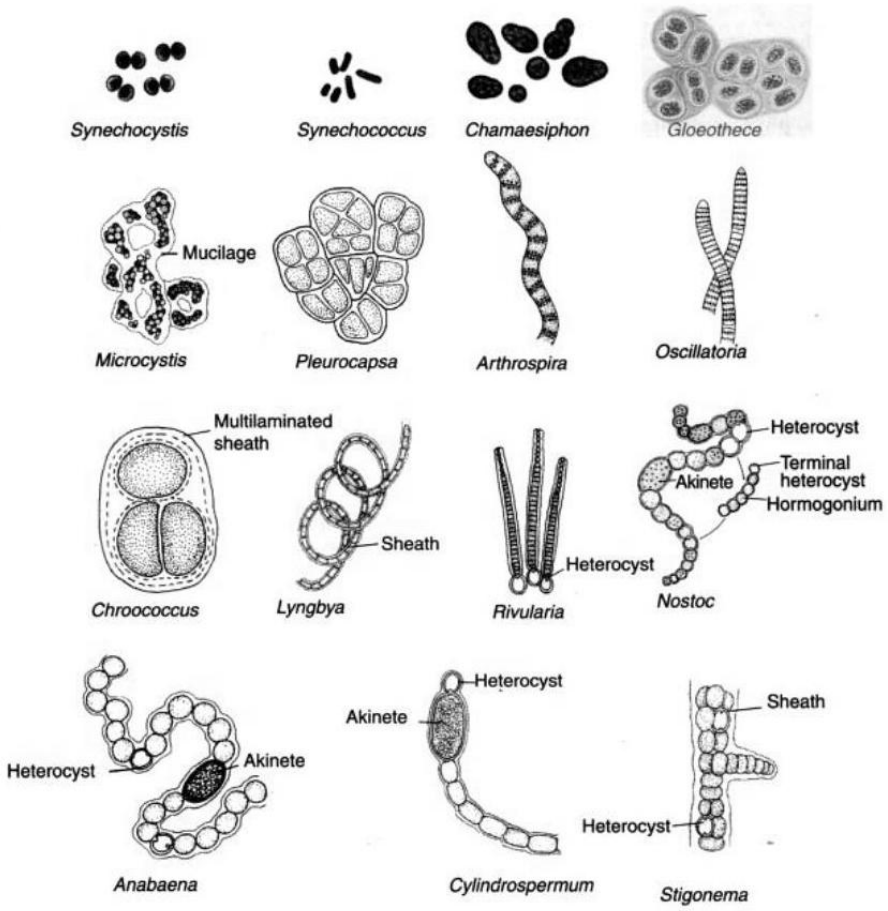


Figure 1 - Morphology of different cyanobacteria. Source: Sumbali G. 2009 Principles of Microbiology.

The reproductive strategies adopted by cyanobacteria vary from binary fission to fragmentation (filaments), budding, and multiple fission. A number of cyanobacteria can also form specialized spore-like cells called akinetes, which can live in a resting state and germinate under favourable conditions (Kaplan-Levy et al. 2010).

Some filamentous cyanobacteria are able to move by gliding and rotating. Many cyanobacteria have gas vacuoles composed by gas vesicles within their cells, which allow them to float in the water. The “blooms” occur when a large number of cyanobacteria floats to the water surface under certain

conditions, such as fluctuations of temperatures, nutrients, or oxygen (Paerl and Huisman 2009).

Many cyanobacteria can fix atmospheric nitrogen by converting it into ammonium, a source of molecular nitrogen (Whitton and Potts 2000). Since the enzyme nitrogenase is strongly inhibited by oxygen, in filamentous cyanobacteria, the nitrogen-fixing reactions occur in specialized cells called heterocysts. In heterocysts, the lack of the oxygenic photosystem II and the presence of mechanisms that keep low the concentration of oxygen create an anaerobic environment ideal for fixing nitrogen. The nitrogen fixation activity carried out by cyanobacteria is fundamental in many environments and in agriculture. For example, the association of the cyanobacterium *Anabaena azollae* with *Azolla*, a fern that grows in rice fields, allows the cultivation of rice with a limited use of fertilizers (Bocchi and Malgioglio 2010).

The ability of cyanobacteria to fix nitrogen and carbon makes them the ideal partner for symbiotic associations, including the endosymbiosis. Cyanobacteria are often found as symbionts of a wide range of eukaryotic organisms such as fungi (lichens), animals (corals, sponges), plants, and protists (Adams et al. 2012).

According to studies that examined cellular and molecular aspects of pure cultures, cyanobacteria have been historically divided into five subsections corresponding to five orders (Rippka et al. 1979): Chroococcales, Pleurocapsales, Oscillatoriales, Nostocales, and Stigonematales (Table 1).

Table 1- Classification of cyanobacteria according to Rippka et al. 1979 [a]

Subsection	Order	Features	Examples of genera
I	Chroococcales	Unicellular or aggregates, reproduce by binary fission or budding	*Prochlorococcus, Synechococcus, Synechocystis Gloeobacter*
II	Pleurocapsales	Unicellular, form aggregates, reproduce by multiple fission	*Pleurocapsa*
III	Oscillatoriales	Filamentous, undifferentiated, reproduce by binary division	*Trichodesmium, Lyngbya*
IV	Nostocales	Filamentous, cell differentiation, reproduce by binary fission	*Anabaena, Nostoc*
V	Stigonematales	Filamentous, branching, cell differentiation	*Chlorogloeopsis, Fischerella*

[a] Adapted from Sarma 2012 and Slepecky and Hemphill 2006.

Subsequent studies based on phylogenetic analyses showed that this classification may be unreliable and proposed new grouping systems (Hoffmann et al. 2005; Komárek J., Kaštovský J. 2014; Dvořák et al. 2015).

Phylogenetic analyses suggest that all the existing cyanobacteria may descend from a common ancestor and that *Gloeobacter violaceus* may be its closest living descendant (Schirrmeister et al. 2015). Supporting this hypothesis, *G. violaceus* is the only cyanobacterium that lacks thylakoid membranes and additionally, molecular studies showed that it split off early from the other cyanobacteria (Kauff and Büdel 2011).

Prochlorophytes are a group of small photosynthetic bacteria, among them *Prochlorococcus* is probably the most abundant photosynthetic organisms on Earth (Flombaum et al. 2013). The light-harvesting system of prochlorophytes differs from that of “normal” cyanobacteria as it contains chlorophyll *b* in addition to chlorophyll *a* and lacks phycobilins. Despite their distinctive features, phylogenetic studies assigned prochlorophytes to the cyanobacterial clade (Lewin 2002). Another notable feature of Prochlorococcus

is its minimal genome size, much smaller than that of other cyanobacteria and very adapted to their specific niches (Biller et al. 2014). Since they have chlorophyll *b*, prochlorophytes have supposed to be the precursor of chloroplasts for many years but, phylogenetic analyses have not been in agreement with this hypothesis (Tomitani et al. 1999). The current idea is that probably cyanobacteria (including prochlorophytes) and eukaryotic chloroplasts evolved from a common ancestor containing phycobilins and various types of chlorophyll.

1.2 Plastids

1.2.1 *Brief description*

Plastids are a family of organelles found in the cells of plants and algae, several marine molluscs and also in some parasitic protists. Among the several functions that are accomplished by plastids, the most popular is photosynthesis, despite not all plastids carry out the photosynthesis. Other functions of plastids are: amino acid and lipid synthesis, carbohydrates (starch) and fat storage, fruit and flower coloration, iron storage, gravity sensing, stomatal functioning, and environmental perception (Jarvis and López-Juez 2013).

In particular, some activities are specific to different classes of plastids, and these organelles are dynamic so that as the cell metabolism changes, plastids may also change by altering their membrane and protein configuration. Different types of plastids are characterized also by the absence or the presence of specific pigments.

All plastids derive from an endosymbiotic process occurred about 1.6 billion years ago, which gave rise to three main lineages of plastids: glaucophytes, red algae, and green algae.

1.2.2 *Symbiogenesis of plastids*

Symbiosis represents an intimate association between different organisms (Schwab 2012) and it is a common phenomenon that occurs in nature. For example, since the ability of using sunlight energy represents a remarkable advantage to survive in several environments, many non-photosynthetic organisms have built transient or even permanent symbiotic relationships with photosynthetic partners. For example, lichens result from a symbiotic relationship between non-photosynthetic organisms (the fungi) and photosynthetic partners such as cyanobacteria, algae, or even both.

During the evolution, some symbiotic associations between organisms became so strong that they eventually merged together to form a new organism. Subsequently, many unnecessary genes were simply lost by the symbiont whereas others were transferred to the host nucleus. This extremely rare phenomenon is called symbiogenesis and, according to the current knowledge it explains the existence of all the current eukaryotes (Margulis 1970).

The endosymbiotic theory was first suggested in the beginning of the 20th century (Mereschkowsky 1905; Mereschowsky 1910) but an accepted theory was formulated only after the advent of the molecular biology (Margulis 1970). Endosymbiosis is defined as the permanent symbiotic relationship between two organisms implying the incorporation of one into another.

According to the endosymbiotic theory, mitochondria originated about 2.2 billion years ago from the engulfment of an alpha-proteobacterium ancestor by the ancient eukaryotic progenitor (Bullerwell 2011; Hedges et al. 2001). Plastids originated later, and in fact all eukaryotes have mitochondria while only some eukaryotes have plastids. Plastids could have arisen either from a single endosymbiotic event (McFadden and van Dooren 2004) or from different and independent endosymbiotic events (Larkum et al. 2007). Several studies support the hypothesis of a single endosymbiotic event, and suggest that 1.6 billion years ago a cyanobacterium-like organisms was incorporated into a eukaryotic organism (Yoon et al. 2004).

The primary endosymbiosis gave rise to the main groups of plastids, namely Glaucophyta, Rhodophyta, and Chlorophyta (McFadden and van Dooren 2004). Chloroplasts of plants evolved from the plastids of Chlorophyta at a later stage, whilst subsequent events of secondary endosymbiosis led to the formation of the plastids found in euglenids, chloroarachniophytes, dinoflagellates, cryptophytes and other heterokont algae (McFadden 2001). Figure 2 shows schematically the evolution of plastids.

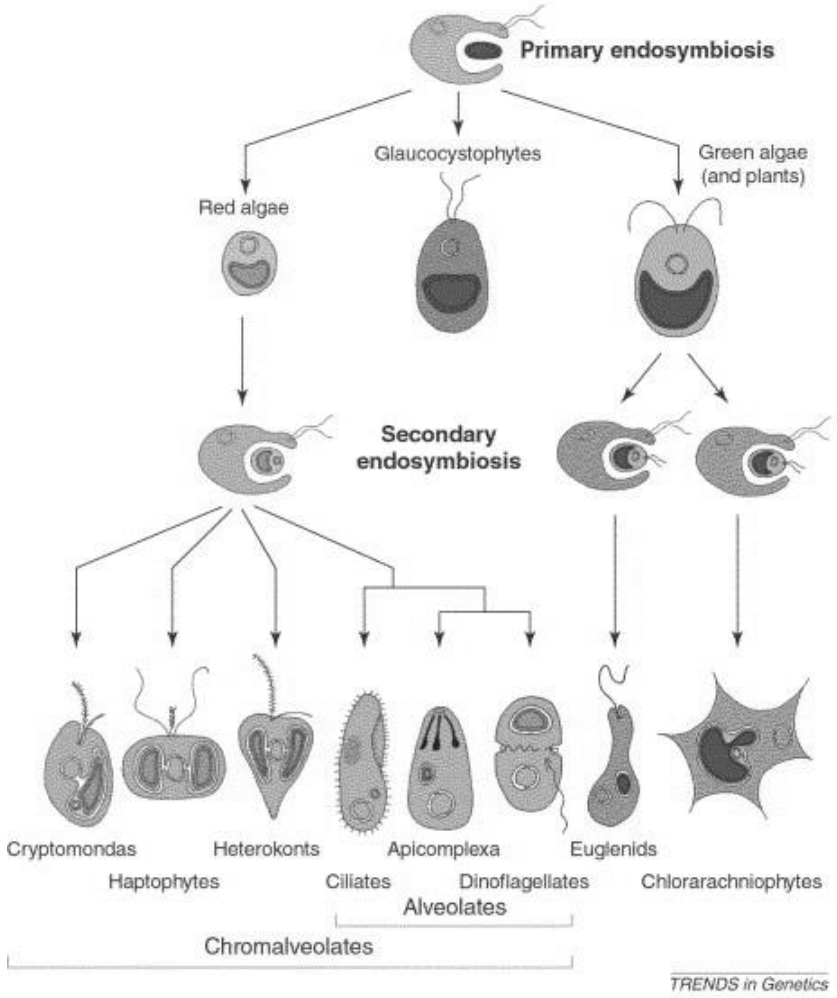


Figure 2 - Origin and evolution of plastids. From Fig 3 of Archibald and Keeling 2002.

Phylogenetic analysis suggested that the Glaucophyta was the first lineage to diverge, thus its plastid-like photosynthetic organelles (sometimes referred to as cyanelles) should resemble the cyanobacterial ancestor more than those of green and red algae lineages. Indeed, plastids of glaucophytes retain ancestral cyanobacterial features such as the presence of a peptidoglycan wall between their two membranes (Pfanzagl et al. 1996). Interestingly, no examples of secondary endosymbiosis involving glaucophytes have been reported so far.

Cyanophora paradoxa is considered the model organism of glaucophytes, indeed both the plastid and nuclear genomes have been sequenced (Stirewalt et al. 1995; Price et al. 2012).

Secondary endosymbiosis occurs when an organism derived by a primary endosymbiosis event is incorporated into another organism. While primary endosymbiosis led to plastids surrounded by two membranes, the plastids originated by secondary endosymbiosis can be surrounded by more than two membranes as occurs, for example, in euglenids and chloroarachniophytes (Table 2).

Together with cryptophytes, chloroarachniophytes are characterised by the presence of a vestigial nucleus called nucleomorph between plastidial membranes, supporting the hypothesis of a secondary endosymbiosis (Gilson et al. 2006). Furthermore, the plastids of dinoflagellates and stramenopiles might be originated by tertiary or multiple endosymbiotic events (Morden and Sherwood 2002).

Table 2 Characteristics of primary and secondary plastids. From table 1.1 of Sharkey et al. 2012.

		Chl			Pb	Mb	Nm
Primary	Secondary	*a*	*b*	*c*	PBP[a]	Mem[b]	Nm[c]
Chlorophyta		X	X			2	
	Chloroarachniophyta	X	X			4	X
	Euglenophytes	X	X			3	
Rhodophyta		X			X	2	
	Cryptophytes	X		X	X	4	X
	Stramenophytes	X		X		4	
	Haptophytes	X		X		4	
	Dinoflagellates	X		X		3	
	Apicomplexan					4	
Glaucophyta		X			X	2	

[a]PBP: phycobiliproteins
[b]Mem: number of membranes surrounding the chloroplast
[c]Nm: presence of nucleomorph

Paulinella chromatophora and *Rhopalodia gibba* are interesting cases of recent and ongoing endosymbiosis. The freshwater amoeba *Paulinella chromatophora* harbours a cyanobacterium-like symbiont, called "chromatophore", which is related to the *Prochlorococcus* and *Synechococcus* genera (Nowack et al. 2008). The number of endosymbionts that reside in the cytoplasm of *P. chromatophora* is strictly regulated, suggesting a higher level

of interaction with the host, even though neither gene loss nor gene transfer have been revealed (Yoon et al. 2006).

R. gibba is a diatom that, like *P. chromatophora*, contains a cyanobacterial symbiont in addition to a plastid derived from the lineage of red algae. In this case, the endosymbiont has apparently lost its photosynthetic capability but, on the other hand, it is still able to fix nitrogen (Prechtl et al. 2004).

Noteworthy are the apicomplexans *Plasmodium falciparum* (the causal agent of malaria) and *Toxoplasma gondii* (the causal agent of toxoplasmosis), which contain plastids known as "apicoplasts" derived from the red lineage (*Rhodophyt*) but that do not have any photosynthetic function (Lim and McFadden 2010). Similarly, the plastids of some parasite plants of the genus *Cuscuta* lost functionality, so that they partially or completely rely on their photosynthetic host.

1.2.3 *Plastids of the Chlorophyta lineage*

Plastids of green algae and land plants are surrounded by an envelope of two membranes, the outer membrane and the inner membrane separated by an intermembrane space (Figure 3). The space delimited by the inner membrane contains a fluid matrix called "stroma" in which is present a third membrane system differentially developed according to the plastid type.

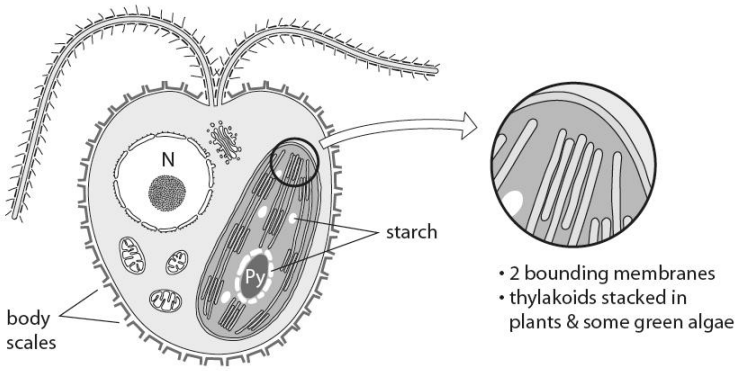


Figure 3 - Plastids of plants. From Kim and Archibald 2009. N = nucleus, Py = pyrenoid.

In multicellular plants, plastids can differentiate in different types, among which the most common are: proplastids, etioplasts, chloroplasts,

amyloplasts, leucoplasts, and chromoplasts (Figure 4) (Pyke 2007; Jarvis and López-Juez 2013).

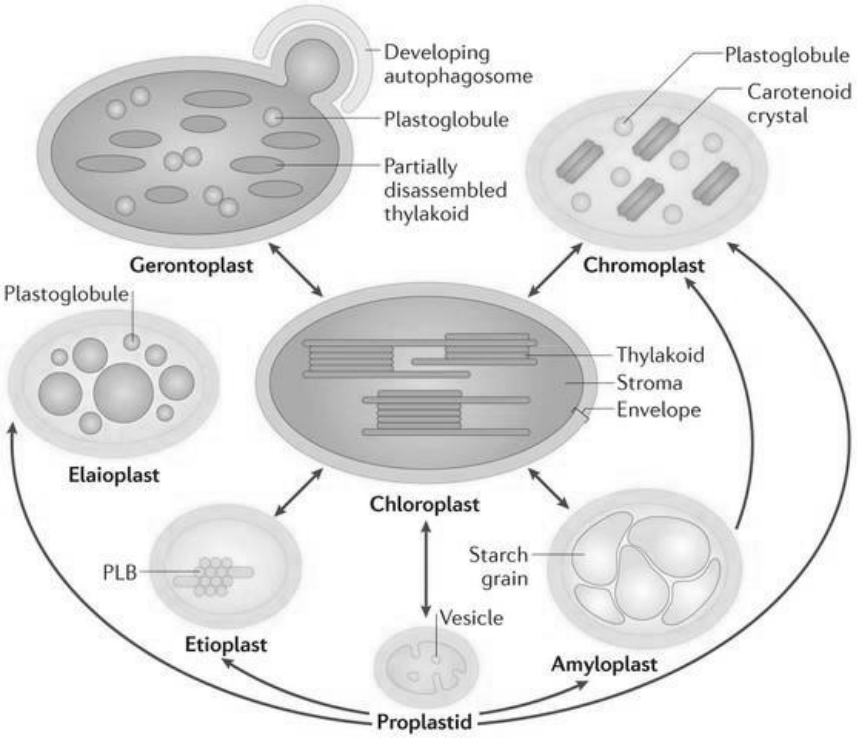


Figure 4 - Plastids differentiation. From Jarvis and López-Juez 2013. PLB = prolamellar bodies.

Proplastids are undifferentiated precursors of the other plastids (Marinos 1967). They are small and nearly colourless organelles which are found in young and rapidly dividing cells (meristematic cells) of roots and shots. In absence of light, proplastids develop into etioplasts, the precursors of chloroplasts that contain prolamellar bodies (Domanskii et al. 2003). When exposed to light, etioplasts differentiate into chloroplasts and the prolamellar bodies develop into thylakoids.

Chloroplasts are the site of photosynthesis, so they are present where the photosynthetic reactions occur. A typical mesophyll cell in a leaf may have up to 30-40 chloroplasts. Commonly, chloroplasts are described as disk-shaped organelles with a diameter of 3-6 µm, but their shape can change. They are called “green plastids” due to the green colour caused by the presence of chlorophyll in their thylakoid membranes, which are densely packed and form

column-like structures called “grana”. Chloroplasts contain also carotenoids, accessory pigments which, in association with chlorophylls, are involved in the capture of light. During the active phases of photosynthesis, sugars may be accumulated in the stroma of chloroplasts as grains of starch.

Chromoplasts lack chlorophyll but, like chloroplasts, synthetize and contain carotenoids which are responsible for the characteristic colours of flowers, fruits and other parts of the plants. Colours have a remarkable function in plant reproduction as they attract insects and other animals essential for the cross-pollination or the dispersal of seeds.

In contrast to chloroplasts, leucoplasts lack pigments and a complex membrane system made up of thylakoids, indeed their primary functions are synthesis and storage. A variety of lipids is synthetized in leucoplasts, and then stored locally or transferred to other cell compartments. When storage is the main function, leucoplasts specialise into different forms, such as amyloplasts (starch storage), elaioplasts (fat storage) and proteinoplasts (protein storage).

1.2.4 *Primary plastids of the Glaucophyta lineage*

The photosynthetic plastids of glaucocystophitic algae are called “muroplasts” or “cyanelles” or “cyanoplasts”. The name muroplasts highlights the existence of a peptidoglycan cell wall similar to that of prokaryotes. Muroplasts are also frequently called “cyanelles”, although some authors argue that this term should be used only to identify endosymbiotic cyanobacteria and not plastids (Wise and Hoober 2007).

Cyanelles are surrounded by two membranes separated by a layer of peptidoglycans, likely the remnant of an ancient prokaryotic cell wall (Kugrens et al. 1999) (Figure 5). Like red algae and cyanobacteria, the thylakoids inside cyanelles are unstacked and covered with the light-harvesting pigments chlorophyll *a* and phycobilins (Keeling 2004). Another cyanobacterial character of cyanelles is the presence of a “carboxysome”, an inclusion rich of the enzyme Rubisco (Kugrens et al. 1999).

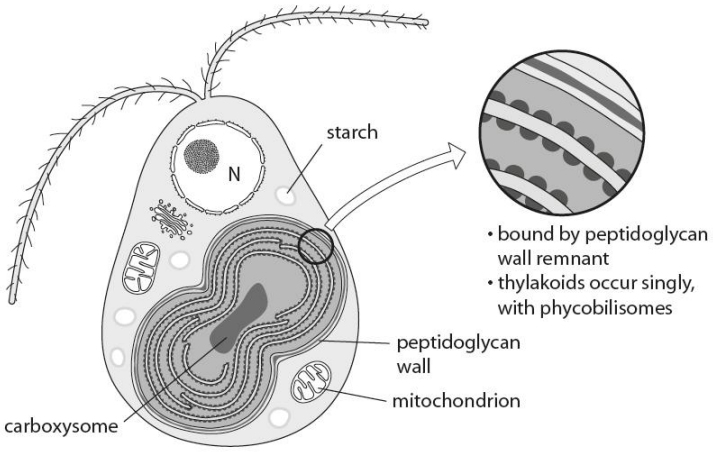


Figure 5 - Plastids of glaucophytes. From Kim and Archibald 2009. *N* is the nucleus.

1.2.5 *Plastids of the Rhodophyta lineage*

The plastids found in red algae are called “rhodoplasts”. Like cyanobacteria and cyanelles, they contain chlorophyll *a* and phycobilins such as phycocyanin and phycoerythrin (Tomitani et al. 1999).

Rhodoplasts are bound by a double membrane and show different morphology from stellate to ovoid (Figure 6). They contain unstacked thylakoids and may contain a “pyrenoid”, which has a function analogous to that of carboxysomes (Giordano et al. 2005).

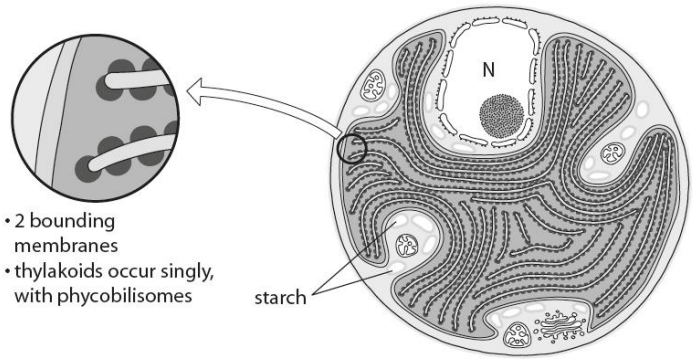


Figure 6 - Plastids of rhodophytes. From Kim and Archibald 2009. *N* is the nucleus.

1.3 Plastid Genomes

1.3.1 *Description*

Plastids harbour their own genome, or "plastome", which generally consists of a circular molecule of double-stranded DNA. Each plastid contains several copies of the plastome, so that photosynthetically active cells can have up to 10,000 plastome copies per cell (Boffey and Leech 1982). Plastid DNA or ptDNA was first isolated in the 1960s from *Chlamydomonas* (Sager and Ishida 1963). Subsequently, with the advent of DNA analysis, the plastomes of several species were mapped (Palmer 1985). The first completely sequenced plastomes have been those of *Merchantia polymorpha* (Ohyama et al. 1986) and *Nicotiana tabacum* (Shinozaki et al. 1986). To date, the NCBI Genome database lists nearly 1,000 complete plastome sequences (Table 3), mostly belonging to green plants (Viridiplantae), in particular to Streptophyta (Embryophyta and Charophyta) → land plants (Embryophyta) → vascular plants (Tracheophyta) → seed plants (Gymnospermae) → flowering plants (Angiospermae).

Table 3 – Number of the annotated plastome sequences available on NCBI Genome.

GROUP	N. SEQ
Viridiplantae	864
Stramenopiles	35
Rhodophyta	14
Alveolata	13
Euglenozoa	10
Cryptophyta	4
Haptophyceae	4
Rhizaria	2
Glaucocystophyceae	1
TOTAL	947

Generally, plastid genomes associate with proteins and RNA to form structures called “nucleoids” (Krupinska et al. 2013). Plastomes are commonly represented as circular DNA molecules (Figure 7), but actually also different linear forms exist, as concatenated or brunched multimers (Bock 2007b). In dinoflagellate algae, the plastid genome is fragmented in mini-circles, each containing a single gene (Zhang et al. 1999). Along with the nucleoid structure, the conformation of ptDNA is likely to play an important role in plastome replication, although an exhaustive replication model is still under study (Bock 2007b).

As shown in Figure 7, plastid genomes are characterised by the presence of two inverted repeats (IR) which define a large single copy region (LSC) and a small single copy region (SSC). Except for their opposite orientation, the DNA sequences of the IRs are identical, so that the genes contained within them are present in two copies for each plastome. The genes located in the IRs are mainly highly expressed genes which code for ribosomal RNAs, ribosomal proteins, and transfer RNAs. The two inverted repeats could be important either to increase the gene dosage or to stabilise the genome, but other explanations are still debated (Palmer and Thompson 1982). Differently from land plants, some green and red algae have only one IR whereas others have direct repeats instead of inverted repeats (Bock 2007a).

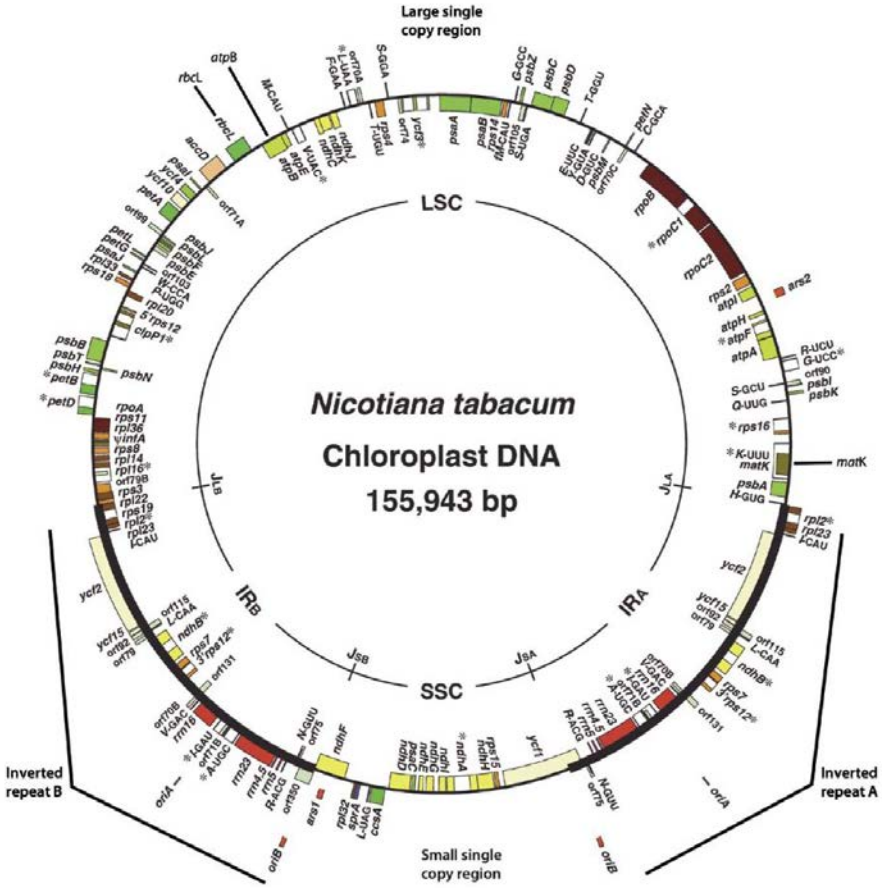


Figure 7 – Gene map of the *Nicotiana tabacum* plastome. Taken from Evert and Eichhorn 2013. *ycf* = hypothetical chloroplast reading frame.

Regarding the gene expression, plastid genomes show unique features combined with characteristics typical of prokaryotic genomes as well as eukaryotic genomes. This mixture of features derives from the adaptation of the prokaryotic plastid genome to interact with the eukaryotic nuclear genome. For instance, plastids use two RNA polymerases, NEP (nuclear-encoded polymerase) and PEP (plastid-encoded polymerase), each one using specific promoters (Toyoshima et al. 2005). Also, many genes are organized in operons like prokaryotic genes, but transcripts are post-transcriptionally processed as occurs in eukaryotes (Stern et al. 2010; Tillich et al. 2006). An intriguing aspect

of plastomes is that all the plastid types have an identical genome, being they undifferentiated proplastids or photosynthetic chloroplasts. In fact, the differences between plastid types depend on regulation mechanisms rather than gene content.

Recent studies on plants found that the nucleotide substitution rate between genomes of mitochondria, chloroplasts, and nuclei have a ratio of 1:3:10 (Drouin et al. 2008). This means that plastid gene sequences are more stable than those of nuclear genes but less than mitochondrial genes. Since plastomes and mitochondria reproduce asexually, they should accumulate deleterious mutations over time via Muller's ratchet (Muller 1964), consequently such a low mutation rate might appear strange. Nevertheless, studies suggest that the high polyploidy of plastomes along with correction processes by gene conversion may contribute to keep the mutation rate so low (Khakhlova and Bock 2006).

Plastomes show a low GC content (30-40%), particularly in the intergenic regions (Guisinger et al. 2011; Raubeson et al. 2007). The highest GC content is found in the IRs due to the presence of four rRNA genes which have the highest GC content than any other coding region. Among the functional categories of protein-coding genes, photosynthetic genes have the highest GC content (Guisinger et al. 2011). Conversely, A or T nucleotides are strongly preferred in the third position of the codons (Liu and Xue 2005).

1.3.2 *Gene content*

The average length of a plastome is 150 kbp but, despite the reduced size, plastid genomes have a high gene content, typically ranging from 80 to 200 genes (Bullerwell 2011). In particular, among the three genomes found in plant cells (in the nucleus, chloroplast, and mitochondrion) the plastome has the highest density of genes (Wakasugi et al. 2001). The chlorophycean *Floydiella terrestris* owns the largest chloroplast genome known so far, with a size of 521 kbp (Brouard et al. 2010). Conversely, plastids belonging to organisms that do not rely anymore on photosynthesis, like the holoparasite plant *Epifagus virginiana*, lost many genes and got a very condensed genomes of 70 kbp or less (dePamphilis and Palmer 1990). Noteworthy, the gene coding for rbcL (the Rubisco large subunit) is conserved also in many holoparasite plants, thus suggesting a role of rbcL in other important plastid functions such as the lipid biosynthesis (Schwender et al. 2004).

Table 4 - Gene content of higher plant plastomes. Adapted from table 4.2 in Bullerwell 2011.

Protein complex or functional category	Gene number	Gene name
GROUP I: Genetic system genes		
RNA polymerase	4	*rpo*
Intron maturase	1	*matK*
Ribosomal small subunit	14	*rps*
Ribosomal large subunit	11	*rpl*
Ribosomal RNAs	4	*rrn*
Transfer RNAs	30	*trn*
GROUP II: Photosynthesis and energy production		
Photosystem I	5	*psa*
Photosystem II	14	*psb*
Cytochrome b6f complex	6	*pet*
NAD(P)H dehydrogenase	11	*ndh*
ATPase	6	*atp*
Rubisco	1	*rbcL*
GROUP III: Conserved hypothetical reading frames and other genes		
Lipid metabolism	1	*accD*
Chaperone and protease	1	*clpP*
Hypothetical chloroplast reading frame	8	*ycf*

The majority of plastid genes are involved in gene expression (e.g. polymerases, maturases, rRNAs, tRNAs) and photosynthesis (e.g. photosystems, rbcL, ATP synthase, cytochrome complex, NAD(P)H dehydrogenase). The remaining genes have heterogeneous or unknown functions (Table 4).

All plastid genomes share a basic set of genes related to their main functions (Table 5). In general, plastomes contain genes encoding 16S, 23S, and

5S rRNA and some ribosomal proteins together with 27-31 tRNA genes. In addition, genes coding for components of photosystems I and II and the cytochrome complex are commonly present (Green 2011).

As organelles, plastids are dependent on the nuclear genome in order to have a complete proteome, indeed more than 95% of the proteins found in plastids are estimated to be imported from the cytosol (Hippler and Bock 2004). Nuclear-encoded proteins are essential for plastid functions, including gene expression and photosynthesis. For instance, in the green lineage the large subunit of Rubisco is encoded in the plastome but the small subunit is encoded in the nuclear genome.

The plastidial rRNA and tRNA molecules constitute a noteworthy exception since there is no evidence of import, and so they seem to derive exclusively from plastid genes (Lung et al. 2006).

Table 5 - Common plastid genes. Taken from table 1 of Green 2011. Common genes in all plastomes are written in bold.

Function	Genes	Comments
RNAs		
Ribosomal	***rns***, ***rnl***, ***rrn5***	4.5S rRNA in plants only
Transfer	*trnA(ugc), trnC(gca), trnD(guc), trnE(uuc), trnF(gaa), trnG(gcc), trnG(ucc), trnH(gug), trnI(cau), trnI(gau), trnK(uuu), trnL(caa), trnL(uaa), trnM(cau), trnN(guu), trnP(ugg), trnQ(uug), trnR(acg), trnR(ccg), trnR(ucu), trnS(gcu), trnS(uga), trnT(ugu), trnV(uac), trnW(cca), trnY(gua)*	
Others	*rnpB (ribonuclease P), ffs RNA (SRP), ssra(tmRNA)*	
Transcription	*cbbX, rbcR, rpoA,* ***rpoB***, ***rpoC1***, ***rpoC2***, *matK*	*matK* in greenline, *cbbX and rbcR* in redline
Translation	*tufA*	
Ribosomal proteins		
Small subunit	***rps2***, ***rps3***, ***rps4***, *rps5, rps6,* ***rps7***, ***rps8***, *rps9, rps10,* ***rps11***, ***rps12***, *rps13,* ***rps14***, *rps16, rps17,* ***rps18***, ***rps19***, *rps20*	All plastid in redline
Large subunit	*rpl1,* ***rpl2***, *rpl3, rpl4, rpl5, rpl6, rpl11, rpl12, rpl13,* ***rpl14***, ***rpl16***, *rpl18, rpl19,* ***rpl20***, *rpl21, rpl22, rpl23, rpl24, rpl27, rpl29, rpl31, rpl32, rpl33, rpl34, rpl35,* ***rpl36***	All plastid in redline and in some greens
Photosynthesis		
ATP synthase	***atpA***, ***atpB***, *atpD,* ***atpE***, ***atpF***, *atpG,* ***atpH***, *atpI*	All plastid in redline
Photosystem I	***psaA***, ***psaB***, ***psaC***, *psaD, psaE, psaF,* ***psaI***, ***psaJ***, ***psaL***, ***psaM***	*psaD, E, F* not in plants
Photosystem II	***psbA***, ***psbB***, ***psbC***, ***psbD***, ***psbE***, ***psbF***, ***psbH***, ***psbI***, ***psbJ***, ***psbK***, ***psbL***, ***psbN***, ***psbT***, ***psbV***, ***psbX***, *psbY, psbZ, psb28*	
Cytochrome complex	***petA***, ***petB***, ***petD***, *petF,* ***petG***, *petL(ycf7), petM (ycf31),* ***petN***(***ycf6***)	*petF* nuclear in plants and many algae
NADH dehydrogenase	*ndhA, ndhB, ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK*	Plants only, except some gymnosperms
Metabolism	*accD, acpP, chlB,* ***chlI***, *chlL, chlN, rbcL*, rbcS*, thiG, thiS, cysA*	*accD* plants only, **rbcL* and *rbcS* have different origins in red and green line (see text)
Protein quality control	***clpC***, *clpP, dnaB,* ***dnaK***, *ftsH(ycf25),* ***groEL***	
Assembly, membrane insertion	*ccs1, ccsA, secA, secG, secY, sufB, sufC, tatC*	Redline only

1.3.3 *Genome reduction*

As extensively described, plastid genomes derive from an ancestral cyanobacterial genome. Considering that the genomes of currently living

cyanobacteria are at least ten times larger than plastid genomes, a drastic genome reduction has occurred since the first endosymbiotic events.

Many genes moved to nuclear genome via a process called endosymbiotic gene transfer (EGT), a special form of lateral or horizontal gene transfer (HGT) (Timmis et al. 2004; Blanchard and Lynch 2000). However, EGT should not have been easy, considering that integrating prokaryotic genes into a eukaryotic genome requires compatible elements such as promoters, UTRs, upstream sequences (like stroma-targeting peptides), and so on.

It has been estimated that less than 10% of the ancestral cyanobacterial genome has been conserved in the plastid genome. This means that many genes have been lost or transferred to the nuclear genome by horizontal gene transfer. Evolutionary analysis showed that about 18% of the protein-coding genes in *Arabidopsis* nuclear genome derives from the cyanobacterial ancestor (Martin et al. 2002). Transferring genes to the nucleus could have been a way to escape from the Muller's ratchet (see the previous paragraphs) and so avoid high mutation rates.

An intriguing question arises from such a hypothesis: why some genes still reside in plastomes? One possible explanation is that the hydrophobicity of the proteins encoded by plastomes could interfere in the transfer through the plastid double membranes. Nonetheless, the large subunit of Rubisco (rbcL) is not a membrane protein with hydrophobic domains but it is still present in most plastomes. More recently, the CORR (COlocation for Redox Regulation) hypothesis has been formulated (Allen 2003). According to the CORR hypothesis, the proteins involved in the redox regulation processes are encoded and expressed directly in plastids in order to guarantee a rapid response to any perturbations in redox balance. Moreover the redox state would trigger the expression of proteins involved in such mechanisms (Bullerwell 2011). However, the import of nuclear encoded proteins is still essential to maintain the expression of proteins implied in CORR.

In addition, plastid genomes of parasite plants lost many genes as a consequence of parasitism. These plants use other plants as carbon source and, depending on their degree of parasitism, they partially or completely lost photosynthetic genes (Krause 2008). Among the parasitic plants that have retained some photosynthetic activity are the plants of the genus *Cuscuta*. On the other hand, *Epifagus virginiana* completely lost its photosynthetic activity and has one of the smallest plastome known so far, with a size of about 70kb.

1.3.4 *Plastid biotechnology*

The biotechnological applications on plants essentially aim to enhance their agronomic potential or produce bioproducts (Maliga 2014). Modifying plastomes rather than nuclear genomes could give important advantages by

reducing the risk of the spread of transgenes and increasing the levels of molecular farming (Maliga and Bock 2011). First, transgenes cannot diffuse by pollen because plastomes of most flowering plants are maternally inherited (Chandler and Dunwell 2008). Second, the high copy number of plastid genomes can ensure higher expression levels while the plastid organelles provide an ideal storage site.

Protocols are available for plastid transformation of a number of plants, including crops (e.g. tomato, potato, eggplant, lettuce, soybean, and cabbage), algae (*Chlamydomonas*), and bryophytes (Maliga 2014).

1.4 Transfer RNA and translation in plastids

1.4.1 *Genetic code and tRNA pool*

In plastids, gene expression includes transcription, RNA editing, RNA splicing, RNA cleavage and trimming, and finally RNA translation (Sugiura 2013). Gene expression is regulated at various levels but primarily during the post-transcriptional phase, especially during the translational process.

The protein-coding genes of plastids are translated into proteins using the standard genetic code, like Bacteria and Archaea (Table 6). This implies that 61 nucleotide triplets code for the standard 20 amino acids whereas the remaining 3 triplets specify termination codons. Since genetic code is degenerate, each amino acid can be coded by two to six synonymous codons, except for methionine and tryptophan.

Generally, synonymous codons are not used with comparable frequencies, so that variable degree of bias exists. Experiments made on *E. coli* showed that the translation efficiency of some codons is higher than others and correlates with their frequency (codon usage) and corresponding tRNA gene content (Tuller et al. 2010b; Plotkin and Kudla 2011). Studies on tobacco chloroplast showed that such correlation does not always exist but varies depending on codons (Nakamura and Sugiura 2007, 2011). It has been speculated that the use of codons with low translation efficiency could be a mechanism to control the translation (Tuller et al. 2010a).

Table 6 - Standard genetic code

First Base	Second Base: T	Second Base: C	Second Base: A	Second Base: G	Third Base
T	TTT F Phe	TCT S Ser	TAT Y Tyr	TGT C Cys	T
	TTC F Phe	TCC S Ser	TAC Y Tyr	TGC C Cys	C
	TTA L Leu	TCA S Ser	TAA * Ter	TGA * Ter	A
	TTG L Leu	TCG S Ser	TAG * Ter	TGG W Trp	G
C	CTT L Leu	CCT P Pro	CAT H His	CGT R Arg	T
	CTC L Leu	CCC P Pro	CAC H His	CGC R Arg	C
	CTA L Leu	CCA P Pro	CAA Q Gln	CGA R Arg	A
	CTG L Leu	CCG P Pro	CAG Q Gln	CGG R Arg	G
A	ATT I Ile	ACT T Thr	AAT N Asn	AGT S Ser	T
	ATC I Ile	ACC T Thr	AAC N Asn	AGC S Ser	C
	ATA I Ile	ACA T Thr	AAA K Lys	AGA R Arg	A
	ATG M Met	ACG T Thr	AAG K Lys	AGG R Arg	G
G	GTT V Val	GCT A Ala	GAT D Asp	GGT G Gly	T
	GTC V Val	GCC A Ala	GAC D Asp	GGC G Gly	C
	GTA V Val	GCA A Ala	GAA E Glu	GGA G Gly	A
	GTG V Val	GCG A Ala	GAG E Glu	GGG G Gly	G

A transfer RNA is described by IUPAC as “A single-stranded RNA molecule containing about 70-90 nucleotides, folded by intrastrand base pairing into a characteristic secondary ('cloverleaf') structure that carries a specific amino acid and matches it to its corresponding codon on an mRNA during protein synthesis” (IUPAC 1992).

Since the degeneracy of the genetic code, each amino acid is coded by specific codons which correspond to specific tRNAs. Therefore, 61 different codon sequences on mRNA match with 61 different anticodon sequences on tRNAs or “cognate” tRNAs. Those tRNAs that decode the same amino acid are defined “isoacceptors”. No cognate tRNAs correspond to the termination (or STOP) codons, which instead are recognised by proteins called release factors that mimic the tRNAs and block the translation (Moore and Steitz 2011).

The plastome of *N. tabacum* codes for a total of 37 tRNA genes, 23 are single copy genes located in the LSC and SSC regions whereas 7 are located in the IR regions and so they are duplicated (Shinozaki et al. 1986). According to the wobble pairing rules postulated by Francis Crick, not all the 64 possible tRNAs but at least 32 tRNA species are required to effectively decode all codons (Crick 1966). Interestingly, most plastomes encode for less than 32 tRNA species and there is no evidence of tRNA import into plastids (Lung et al. 2006; Rogalski et al. 2008), as it has been found in mitochondria (Salinas et al. 2008; Schneider 2011).

Usually, plant plastomes encode about 30 tRNA species but two lycophytes, *Selaginella moellendorfii* (Smith 2009) and *S. uncinata*, have lost many tRNA genes and encode only 12 tRNA species (Tsuji et al. 2007). Moreover, some parasitic organisms have lost many tRNA genes as a consequence of the extensive genome reduction: the parasitic green algae *Helicosporidium* sp. has retained 24 tRNA species whereas the parasitic flowering plant *E. Virginiana* owns only 18 tRNA species; some holoparasitic plants of the genus *Cuscuta* still retain the genes of 23 tRNA species; the apicomplexan parasite *Plasmodium* encode 24 tRNA species. The parasitic orchid *Rhizanthella gardneri* possesses the most reduced tRNA set found so far, with only 8 tRNA species (9 genes). In such extreme cases tRNA import seems to be mandatory, otherwise the translation could not be accomplished.

A number of experiments showed that 25 tRNA isoacceptors (Rogalski et al. 2008; Alkatib et al. 2012b; Tiller and Bock 2014) or even less (Koning and Keeling 2006) could read all codons by exploiting the so called “extended wobbling”. According to these rules a minimal set of 23 tRNA genes should be sufficient for plastids to autonomously decode all codons.

Finally, some tRNA does not have only a translational function and thus they are retained also in very reduced genomes like that of *R. gardneri*. For instance, tRNA-Glu is involved in the synthesis of d-aminolevulinic acid and consequently in the pyrroles biosynthesis which results in the production of a variety of compounds such as the pigments bilirubin and biliverdin. Some years ago it has been suggested that tRNA-Glu might be the only gene found in all plastid genomes (Barbrook et al. 2006). However, currently many plastome sequences available on GenBank seem to lack the genes coding for tRNA-Glu (e.g. NC_024286.1), but annotation errors cannot be excluded.

1.4.2 *Wobble and extended wobble rules*

Wobble pairings were postulated by Francis Crick (Crick 1966). He suggested that the first base on the anticodon has less spatial constraints than the other two bases, so non-standard base pairings are possible. In particular, the specificity of a tRNA is given mainly by the first and the second base of the

codon (or the second and the third base of the anticodon) which form strong Watson and Crick base pairings. Regarding the third base, Crick said that "in the base-pairing of the third base of the codon there is a certain amount of play, or wobble, such that more than one position of pairing is possible."

Among the various pairing possibilities, when the bases C or A are in wobble position on the tRNA anticodon, only the standard pairings are likely to occur and one tRNA recognizes one codon. Conversely, if the bases U or G are in wobble position on the tRNA anticodon, non-standard pairings can occur and one tRNA can recognize two codons.

Crick suggested also that tRNAs with an unmodified U in the first anticodon position can pair with A, U, G, or C nucleotides in the third codon position. This mechanism is called "four-way wobble", "hyperwobble" or "superwobble", and it implies that a single tRNA can decode up to four codons (Rogalski et al. 2008; Alkatib et al. 2012b; Tiller and Bock 2014). Superwobbling could explain the low number of tRNA species for Pro, Ala and Leu in the plastome of *N. tabacum*. In particular, tRNA(UGG)-Pro, tRNA(UGC)-Ala, and tRNA(UAG)-Leu are sufficient to read all the four codons of Pro, Ala and Leu, respectively. In chloroplast, it has been demonstrated that tRNA-Gly(UCC) can read all the four Glycine codons (Rogalski et al. 2008). Nevertheless, U:C and U:U pairings are unlikely, due to chemical and physical causes and also because superwobbling could cause translation errors in case of amino acids coded by two codons (Crick 1966). Probably, biochemical modifications of U in wobble position favor or limit the pairing capabilities accordingly to the coding potentiality of genomes (Rogalski et al. 2008).

An alternative hypothesis, called "two out of three" hypothesis, states that in some cases only the pairing of two of three nucleotides (first and second codon position) is essential for the interaction between anticodon and codon. The "two out of three" mechanism works best with codons having high GC content, which cause strong GC pairs with the anticodons. Evidence from *in vivo* experiments supports the "superwobble" hypothesis but not the "two out of three" hypothesis (Rogalski et al. 2008), whereas other authors support the essentiality of this rule only for decoding Arginine (CGN) codons (Karcher and Bock 2009; Delannoy et al. 2009a; Alkatib et al. 2012b; Zhou et al. 2013).

The *in vivo* experiments of Alkatib et al. showed that in all four-codon boxes, the tRNA species with U in the wobble position is essential. Probably, in such cases the two-out-of-three mechanism may not be possible because G-G and G-A base pairing in the wobble position is unlikely to occur (Alkatib et al. 2012b).

In most bacteria and in plastids the cytosine at the wobble position of tRNA(CAT)-Ile is post-transcriptionally modified to lysidine allowing the decoding of ATA codons (Alkatib et al. 2012a). A second type of tRNA editing that can occur in bacteria and plastids is the conversion of the adenosine to

inosine in tRNA(ACG)-Arg, which can decode also CGA and CGC codons (Karcher and Bock 2009; Zhou et al. 2013).

1.4.3 *RNA editing of gene transcripts*

In land plants, transcripts from plastid genome can be modified so that the mRNA sequence does not exactly correspond to the relative genomic sequence (Tillich et al. 2006; Sugiura 2008). In most cases, RNA editing consists in the conversion of C to U and sometime of U to C. In non-seed plants such as ferns, mosses and hornworts this process is much more frequent. The result of these modifications is beneficial due to the introduction of a start codon, the removal of a stop codon, or the restoration of a conserved amino acid. However, the exact function of such RNA editing is still under study, with some authors suggesting an error repair mechanism or also a kind of gene regulation (Stern et al. 2010).

In bioinformatics RNA editing may cause problems for phylogenetic analysis or codon usage calculation. Particular attention is needed when studying ferns, lycophytes and hornworts, which have high RNA editing rates. However, in seed plants the editing sites are significantly fewer, so the effect on the analyses should not be relevant. Furthermore, RNA editing sites can be predicted by different software such as PREPACT (plant RNA editing prediction and analysis computer tool), CURE (cytidine-to-uridine recognizing editor) and similar (Finster et al. 2012). ChloroplastDB lists a number of annotated sites targeted by RNA editing (Cui et al. 2006).

2 MATERIALS AND METHODS

2.1 Data Retrieval

2.1.1 *Genome sequences*

All the plastid genome sequences used in this thesis were retrieved from the NCBI Genome website, subsection “organelles” (NCBI Genome 2015). NCBI groups all plastids in taxa number 2759, so it is possible to retrieve all the GenBank accession codes from the following URL:

http://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=2759&opt=plastid.

Accession codes were used to download the corresponding GenBank full records from the NCBI Nucleotide database (NCBI GenBank 2015). In particular, only the reference sequences were considered (RefSeq database) and used to build a dataset of plastid genomes.

On the NBCI GenBank website, cyanobacteria are grouped under taxa number 1117, so in this case genome sequences were obtained by searching the NCBI GenBank database (NCBI GenBank 2015) for the following string:

txid1117[orgn] AND "complete genome"[title] AND "refseq"[Filter]

The genome sequence of *Escherichia coli* str. K-12 was used as reference for bacteria genomes. The genome sequence of E. coli has the accession number NC_000913, it can be downloaded from the following URL:

https://www.ncbi.nlm.nih.gov/nuccore/NC_000913

2.1.2 *Organisms classification*

The organisms analyzed in this thesis were grouped as illustrated in Table 7 (adapted from table 4.1 of Bullerwell 2011) according to taxonomic data provided by NCBI and the current knowledge of the evolutionary relationships among eukaryotes.

Table 7 - Phyla of eukaryotic organisms mentioned in this thesis.

Supergroups	Phyla		Plastid Origin
ARCHAEPLASTIDA	Streptophyta Chlorophyta Rhodophyta Glaucophyta or Glaucocystophyta		Primary Endosymbiosis
EXCAVATA	Euglenozoa		Secondary Endosymbiosis GREEN LINEAGE
RHIZARIA	Cercozoa/Chloroarachniophyta		Secondary Endosymbiosis GREEN LINEAGE
CHROMALVEOLATA	Chromista	Cryptophyta Heterokonta (or Stramenopiles) Haptophyta	Secondary Endosymbiosis RED LINEAGE
	Alveolata	Dinoflagellata Apicomplexa	

2.1.3 *GenBank annotations*

The plastid genome of *N. tabacum* (accession number NC_001879) is taken as an example to describe the GenBank annotations. The GenBank record is showed in Figure 8 and Figure 9. All the GenBank records begin with general information about the sequence such as title, length, date, organism taxonomy, authors, etc. The record continues with a section dedicated to the features that have been annotated for the given sequence. In particular, gene annotations include the position on the sequence, product type (e.g. CDS, rRNA, tRNA) and other specific information.

Figure 8 - GenBank record: general information.

```
gene            complement(6..80)
                /gene="trnH"
                /locus_tag="NitaCt113"
                /db_xref="GeneID:800436"
tRNA            complement(6..80)
                /gene="trnH"
                /locus_tag="NitaCt113"
                /product="tRNA-His"
                /db_xref="GeneID:800436"
gene            complement(536..1597)
                /gene="psbA"
                /locus_tag="NitaCp001"
                /db_xref="GeneID:800514"
CDS             complement(536..1597)
                /gene="psbA"
                /locus_tag="NitaCp001"
                /codon_start=1
                /transl_table=11
                /product="photosystem II protein D1"
                /protein_id="NP_054477.1"
                /db_xref="GI:11465935"
                /db_xref="UniProtKB/Swiss-Prot:P04848"
                /db_xref="GeneID:800514"
                /translation="MTAILERRESESLWGRFCNWITSTENRLYIGWFGVLMIPTLLTA
                TSVFIIAFIAAPPVDIDGIREPVSGSLLYGNNIISGAIIPTSAAIGLHFYPIWEAASV
                DEWLYNGGPYELIVLHFLLGVACYMGREWELSFRLGMRPWIAVAYSAPVAAATAVFLI
                YPIGQGSFSDGMPLGISGTFNFMIVFQAEHNILMHPFHMLGVAGVFGGSLFSAMHGSL
                VTSSLIRETTENESANEGYRFGQEEETYNIVAAHGYFGRLIFQYASFNNSRSLHFFLA
                AWPVVGIWFTALGISTMAFNLNGFNFNQSVVDSQGRVINTWADIINRANLGMEVMHER
                NAHNFPLDLAAIEAPSTNG"
gene            complement(1812..4409)
```

Figure 9 - GenBank record: features.

The pipeline recommended by NCBI for annotating genes is described in the NCBI Handbook (Ostell and McEntyre 2013), for both prokaryotic and eukaryotic genomes. Gene and product name are annotated using respectively, the /gene and /product tags. In case of tRNA genes, the gene name is given in the compact form (e.g. trnM) whereas the product name specifies the extended name (e.g. tRNA-Met). However, no exact guidelines are given about how annotating the tRNA anticodon sequence or alternatively the codon specificity.

The NCBI annotation pipeline uses tRNAscan-SE to identify tRNA genes (Lowe and Eddy 1997). As explained by NCBI, “to identify tRNA genes, the input genome sequence is split into 200 nucleotide (nt) windows with overlap of 100 nt and run through tRNAscan-SE program”. Furthermore, tRNAscan-SE can identify 99–100% of transfer RNA genes “with less than one false positive per 15 gigabases”. tRNAscan-SE provides several information about tRNA genes such as length, position in the genome, and the anticodon sequence (Figure 10). Unfortunately, several genomes lack any annotations of the tRNA anticodon.

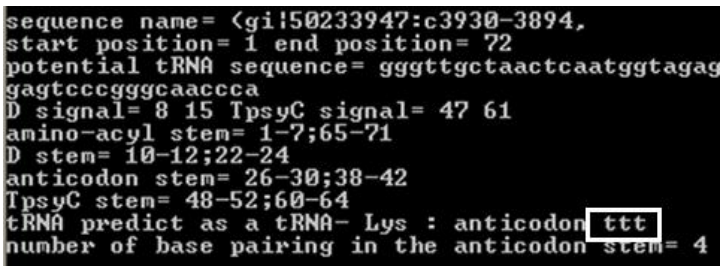


Figure 10 - tRNAscan: output example.

In GenBank, protein-coding genes are annotated as coding sequences (CDS). As already described for tRNA genes, gene name and product name are annotated together with other details, including the translated amino acid sequence (Figure 9).

2.2 Data Processing - tRNA tools

A dedicated set of computer programs, called "tRNA tools", was developed in order to analyse the tRNA gene copy number and the CDS codon frequency in plastid and bacteria genomes. The software includes three scripts named tRNA_finder, CODON_finder, and CORR_checker. The scripts were written in the Perl programming language and are freely available on a Git repository:

https://github.com/bioinfoplant/tRNA/releases/latest

2.2.1 *tRNA_finder*

"tRNA_finder" searches genomes for tRNA genes and retrieves their amino acid specificity and the corresponding anticodon sequence (triplet). tRNA-finder fetches each tRNA gene sequences (merging the exons when introns are present) and uses tRNAscan-SE 1.3.1 (Lowe and Eddy 1997) to identify the anticodon sequence.

tRNA_finder takes as input the GenBank flat files and analyses one record at once. First of all, genomes with no tRNA annotations are discarded, because the aim of the script is not to annotate the sequence from scratch. Once extracted the basic information (e.g. date, organism, accession number, etc.), the program starts to parse tRNA annotations. The tRNA specificity is written in the gene or product name, following the compact (one letter) or extended (three letters) notation. Pseudogenes are skipped as well as tRNA for non-standard amino acids.

Since the absence of standard annotation rules, extracting the tRNA anticodon sequence represents the most challenging task carried out by tRNA_finder. Indeed, the anticodon sequence can be found in /gene, as recommended by NCBI, but also as a note (/note) in the annotation. Sometimes a specific tag specifies the codon recognised, so it is possible to know the anticodon sequence from the reverse complement. Additional problems in the identification of the anticodon sequence are caused by different notation styles that use brackets or symbols. tRNA_finder uses a combination of regular expressions to extract all as much anticodon information as possible.

The worst scenario is the absence of any information about tRNA anticodons. In this case, tRNA_finder extracts the tRNA sequence and uses tRNAscan-SE to recover the missing data. The parameters of tRNAscan-SE are adjusted according to the genome analysed, whether it is an organellar, a prokaryotic or a eukaryotic genome. At the end of this recovering procedure, the results of tRNAscan-SE are compared with the existing annotations. The entire GenBank record is discarded if at least one anticodon sequence found by tRNAscan-SE does not correspond to the amino acid specified in the original annotation. The recovering procedure is very strict but assures the best correspondence between existing annotations and recovered data.

The output of tRNA_finder consists of tab-separated files, each one suitable for specific analyses. Table 8 shows the information extracted by tRNA_finder from the GenBank records.

Table 8 - Information retrieved by tRNA_finder

Field name	Description
NCBI ID	GenBank accession number
NAME	Name of the species [NCBI id]
DEFINITION	Sequence definition
DATE	Submission date
DIVISION	Subgroup name assigned by NCBI
CLASSIFICATION	Taxonomical classification
TOTAL tRNAs	Number of tRNA annotations found
tRNA species	Number of tRNA types found
tRNAs standard	Number of tRNA for standard amino acids
Unknown Anticodons	Number of missing anticodons
tRNA gene count	tRNA gene copy number

2.2.2 *CODON_finder*

"CODON-finder" extracts the CDS annotations from GenBank flat files, and then uses the program CodonW 1.4.4 (http://codonw.sourceforge.net/) to calculate codon usage statistics.

CODON_finder can be instructed to analyse the codon usage of either whole genomes or single genes. Stop codons are not considered in the analysis because they interact with release factors, not with tRNA molecules.

Along with the absolute codon number, CODON_finder can also retrieve information about the RSCU or Relative synonymous codon usage (Sharp et al. 1986). RSCU is the ratio of the observed frequency of a codon to its expected frequency, considering equal likelihood of usage among synonymous codons. RSCU scores equal to 1.0 indicate a lack of codon bias.

2.2.3 *CORR_checker*

The main purpose of the two scripts previously described, tRNA_finder and CODON_finder, is preparing the data to be further analysed with CORR_checker, the third script of the tRNA_tools bundle.

CORR_checker aims to calculate the correlation between codon usage and tRNA gene content in genes or whole genomes. So the first step is loading the data generated by tRNA_finder and CODON_finder, and then checking their correspondence. If an entry lacks data of tRNA gene content or codon usage, it will be discarded.

The second step is preparing the data in order to account for the wobble rules. Table 9 summarises the standard wobble pairings (A, G, T, C are the IUPAC symbols for the standard bases).

Table 9 - Standard wobble pairings

Anticodon (1st base)	Codon (3rd base)
A	T
G	C, T
T	A, G
C	G

In addition, two types of post-transcriptional tRNA editing were taken into account: the conversion of cytidine to lysidine (L) and the conversion of adenosine to inosine (I). The conversion of cytidine to lysidine occurs in CAT anticodons, changing the specificity of the tRNA from Methionine to Isoleucine. The conversion of adenosine to inosine allows the reading of three codons for Arginine with a single tRNA. Table 10 lists the tRNA editing and the special cases considered.

Table 10 – tRNA editing and special cases

tRNA	Anticodon	Target Codon	Description
tRNA-Arg	A(I)CG	CGT, CGC, CGA	Conversion of Adenine to Inosine
tRNA-Ile	C(L)AT	ATA	Conversion of Cytidine to Lysidine
STOP	TCA	TGA	Does not target TGG (Trp)
tRNA-Ile	TAT	ATA	Does not target ATG (Met) codons to avoid mistranslation

CORR_checker uses an additive method to implement wobble rules, tRNA editing, and special cases in evaluating the correlation between codon usage and tRNA gene content. Essentially the method takes into account how many codons can be read by each tRNA anticodon. This means that the final codon number corresponding to each tRNA corresponds to how many codons it can actually decode.

Table 11 shows an example to better understand how the algorithm works. The amino acid Alanine can be coded by four tRNAs, therefore both the wobble and the superwobble rules could be applied. Wobble rules indicates that tRNAs having T (U) or G as the first anticodon base can pair with two codons. Hence, tRNA(TGC) can pair with GCA but also with GCG, so the number of codons that can be read by such tRNA is the sum of the number of GCA and GCG codons. In the example the total number is 600 codons (500+100). The same procedure is applied to tRNA(GGC), which can pair with GCC and GCT. According to the superwobble rules, tRNAs having T (U) as the first anticodon base can pair with four codons. On the basis of Crick hypothesis and the studies made on *N. tabacum*, the algorithm of CORR_checker applies this rule only to the four-codon families of amino acids such as Alanine, Valine, Glycine, etc. In

the example tRNA(TGC) can pair with GCA, GCC, GCG, and GCT, hence the corresponding number of codons is 1250 (500+50+100+600).

Table 11 – Example of wobble rules implementation

	tRNA(TGC)	tRNA(GGC)	tRNA(CGC)	tRNA(AGC)
Amino acid	Ala	Ala	Ala	Ala
Codon Triplet	GCA	GCC	GCG	GCT
N. Codons – Real	n1	n2	n3	n4
N. Codons – Wobbling	n1+n3	n2+n4	n3	n4
N. Codons – Superwobbling	n1+n2+n3+n4	n2+n4	n3	n4

After the implementation of the wobble rules, CORR_checker moves to the next step, that is estimating the correlation between codon number and tRNA gene number.

Since tRNA gene copy number and codon usage are not normally distributed, the correlation is estimated by using the non-parametric Spearman's statistics (Spearman 1904). The null hypothesis to test is that the correlation coefficient is greater than zero, that is a one-tailed test. Different methods can be used to estimate the significance of Spearman's statistics: Fisher transformation, Student's t approximation, or random permutation test (Field et al. 2012). CORR_checker can use the Student's t approximation and the random permutations.

The first method calculates the Spearman's rank coefficients using the Statistics::RankCorrelation perl module. Using equation (1), the Spearman's correlation coefficient *rho* is converted to a *t*-value distributed approximately as a Student's *t*-distribution. Next, the p-value can be found from the critical values of the Student's distribution.

$$t = rho\sqrt{\frac{n-2}{1-rho^2}} \quad (1)$$

The second method relies on random permutations of data. CORR_checker is able to directly generate random permutations using the "shuffle" function included in the List::Util Perl module. Afterwards, the Spearman's correlation is calculated between tRNA gene number and both the original and the permuted codon data. The script counts how many times the Spearman's correlation of permuted codon data versus tRNA data results equal or greater (one-tailed test) than the Spearman's correlation between original codon data and tRNA data. Finally, the permutational probability value (P_{perm}) is calculated using equation (2).

$$P_{perm} = \frac{\sum_{n=1}^{N_{all}} rho_n \geq rho^*}{N_{all}+1} \quad (2)$$

The previously described implementation has the advantage to use directly the Perl modules, so it works on all the computer systems having the Perl interpreter installed. The downside of using Perl to generate the permutations is the very low speed of execution. To overcame this issue CORR_checker includes a second method of random permutation which is based on the R software (R Development Core Team 2013). Indeed, Perl scripts can get access to the R engine by using the Statistics::R Perl module. Inside R, the package "coin" provides the function "spearman_test" which calculates the Spearman's statistics many times faster than Perl and uses Monte-Carlo resampling (useful in case of a low number of samples) (Zeileis et al. 2008).

2.3 Additional statistics analyses

The statistical analyses were carried out using the R software version 3 (R Development Core Team 2013). In order to accomplish specific tasks, the following packages were added to the base distribution: gplots (Warnes et al. 2014), vegan (Oksanen et al. 2013), MASS (Ripley et al. 2014).

2.3.1 *Explorative statistics*

A heatmap graph was generated in order to have a general view of the median content of tRNA genes within each supergroup of plastid genomes. A

heatmap is a matrix of values represented with levels of colours and clustered according a criterion of similarity (Wilkinson and Friendly 2009). Rows represent plastids from every supergroup whereas columns represent the 61 species of tRNA, so that each cell shows the median value of the gene copy number of a given tRNA for each supergroup. The distance matrix for the clustering was calculated with the “vegdist” function included in the Vegan package, using “jaccard” method for rows and “euclidean” method for columns. The clustering was performed using the “hclust” function included in the R base distribution. Rows were clustered with “complete” method whereas columns were clustered with “ward.d2” method. Finally, the heatmap graph was drawn with the “heatmap.2” function included in the “gplots” package.

2.3.2 *Multivariate statistics*

Multivariate statistics was carried out by using the functions included in the packages Vegan and MASS. The tRNA gene copy number of all the analysed organisms was imported in R as a dataframe. A distance matrix was computed using the “vegdist” function and selecting “jaccard” as method. The quantitative form of the Jaccard distance implemented in Vegan is actually the Ruzicka index and was preferred over the Euclidean distance for its better performances in presence of species containing missing tRNA type (i.e. excludes joint absences). Moreover the Jaccard index compared to the Bray-Curtis distance is fully metric (Anderson 2006; Anderson et al. 2011)

In order to obtain a graphical representation of the distances between every genome sequence an unconstrained ordination was performed by computing a Non-metric Multidimensional Scaling (NMDS). The NMDS helps to visualize distances between samples by producing a low-dimensional spatial map. Moreover the non-metric scaling methods are able to handle non-Euclidean distances such as the Bray-Curtis and the Jaccard distances. The function included in Vegan to perform NMDS with random start is called “metaMDS”. The default options of metaMDS are optimized for ecological data but this issue can be overcome by setting “autotransform=FALSE” and “noshare=FALSE” (Oksanen et al. 2013).

We estimated the quality of the non-metric multidimensional scaling using the Shepard diagram and the “stress” value, a normalized version of the sum of squared errors considered acceptable if lower than 0.2 (Sturrock and Rocha 2000).

The statistical significance of the differences in tRNA gene content among supergroups was estimated using the “adonis" function included in the Vegan package. The “adonis" function performs a Permutational Multivariate Analysis of Variance (PERMANOVA) avoiding the use of the Bonferroni correction (Anderson 2001). The function “betadisper” in combination with the

"permutest" function gives an estimation of the homogeneity of variance (dispersion between groups) based on the PERMDISP2 method (Anderson 2006).

Finally, the function "envfit" included in the Vegan package was used to assess the influence of each variable (tRNA gene copy number) on the position of plastid genomes on the NMDS map. In addition, the parameter "permutations" was set to 999 in order to calculate the significance of each fitting using a permutational approach.

3 RESULTS

3.1 tRNA tools

The first result obtained from the work behind this thesis has been the development of "tRNA tools", a suite of bioinformatic software designed to analyse the tRNA gene content in plastid genomes. All the software was written with the Perl programming language, thus it can be executed on all the computers having the Perl interpreter, whether they be Windows, Mac, or UNIX systems. Moreover, the entire project is open source and freely available on the public repository GitHub.

As extensively described in chapter 2, tRNA tools are composed by three computer programs each one designed to accomplish specific tasks. tRNA_finder explores the plastid genomes annotated on NCBI GenBank in order to find tRNA genes and define their type, number, and anticodon sequence. Additionally, tRNA_finder is able to recover missing information about the anticodon sequence. The other two programs included in tRNA tools are CODON_finder and CORR_checker. CODON_finder processes the protein-coding genes to calculate the codon usage statistics. The third program, CORR_checker, uses the data provided by tRNA_finder and CODON_finder, to evaluate the relationship between tRNA gene content and codon usage.

tRNA_finder, CODON_finder and CORR_checker can be used in sequence as parts of a pipeline or independently of each other. Moreover, the output consists of flat text files in tab-separated format, which can be easily imported into other software such as R, SPSS, Excel, and many others.

3.2 Data enrichment

As already described in chapter 1, so far the NCBI Genome database lists nearly 1,000 complete plastome sequences (Table 3). It is clear that the database is extremely rich of plastome sequences belonging to Streptophyta plants, mostly Embryophyta. Conversely, other group of organisms are poorly represented, especially Rhizaria and Glaucocystophyceae.

In addition to the biased number among groups, nearly 40% of the plastome sequences submitted to NCBI GenBank has tRNA gene annotations with no information about the anticodon sequence, and consequently about the codon recognised (Figure 11). Noteworthy, among the "defective" plastomes are those of *C. paradoxa* and *P. chromatophora*.

tRNA_finder, the first program included in tRNA tools, is specifically designed to analyse tRNA annotations and recover the anticodon sequences. Processing the data with tRNA_finder allowed recovering about the 70% of the

missing anticodon sequences (27% out of 38%). For only the 11% of plastomes was not possible to recover anticodon data in agreement with the information provided by the existing annotations (Figure 11).

The purpose of this thesis was to analyze the tRNA annotations collected and annotated by NCBI, not to create new annotations or correct the existing ones, so the plastomes with missing anticodon information was discarded.

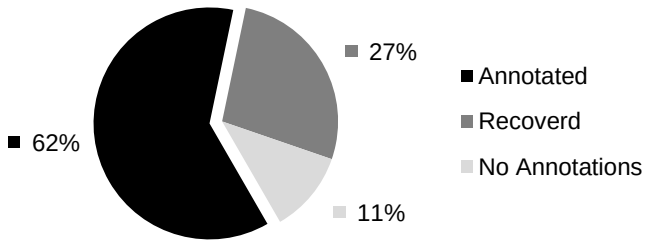


Figure 11 – tRNA annotations

As shown in Figure 12, tRNA_finder recovered the anticodon information for plastomes belonging to all groups. Above all, tRNA_finder succeeded in recovering the tRNA anticodon sequences from the annotations of the plastomes of *C. paradoxa* and *P. chromatophora*. Since the importance of these two organisms in the study of plastid evolution, it has been extremely important to recover their annotations.

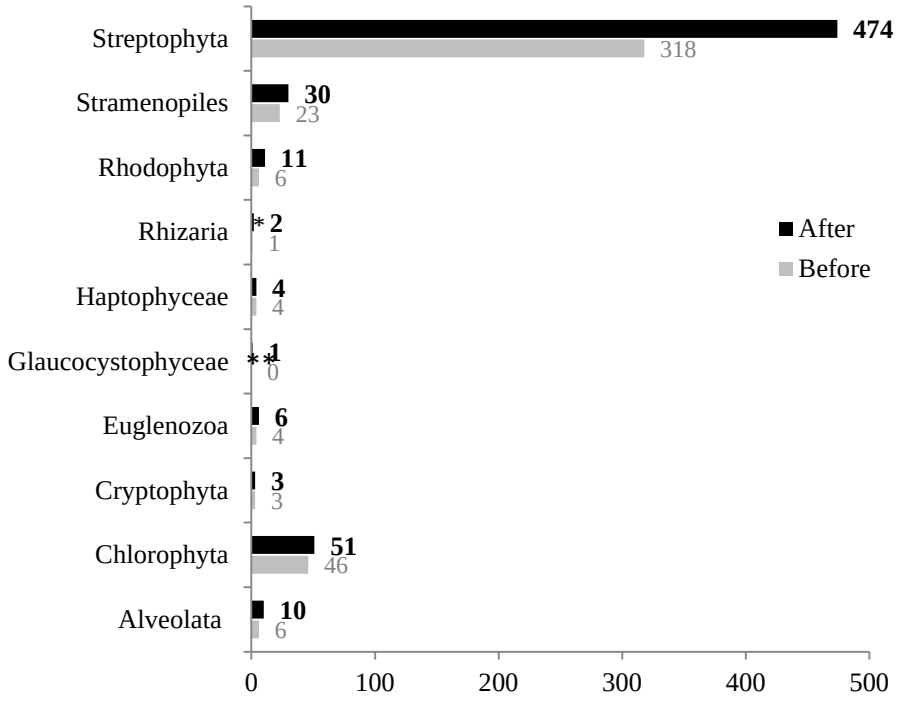


Figure 12 - Data enrichment within groups. * *P. chromatophora*. ***C. paradoxa* the only Glaucophyta sequenced so far.

3.3 tRNA genes

3.3.1 *tRNA gene content*

The output data of tRNA_finder shows several aspects of the plastid genomes and their tRNA gene content. First of all, it is important to observe the number of tRNA gene among the different groups of plastids. The boxplots in Figure 13 and Figure 14 provide a graphical summary of the distributions of tRNA genes and tRNA isoacceptor species, respectively.

The median value of the annotated tRNA genes among all plastid genomes is 37 (95% CI = 36.7- 37.3). Most of the genomes belong to Streptophytes, which have indeed a median number of tRNA genes equal to 37 (95% CI = 36.9- 37.1), so the global median is strongly influenced by this group. Noteworthy, the tRNA gene distribution in Rhizaria have the largest standard deviation (SD = 9.2) because in this group there are two very diverse

plastids: the cyanelle of *P. chromatophora* and the plastid of *Bigelowiella natans*.

Cyanobacteria genomes contain a median of 43 tRNA genes (95% CI = 42.2- 43.9), higher than plastid genomes. However, certain cyanobacteria, mostly of the genera *Prochlorococcus*, have only 37 tRNA genes like the median number of tRNA genes in plastid genomes. At the opposite, a number of cyanobacteria such as *Nostoc punctiforme* have more than 70 tRNA genes, almost twice the amount of tRNA genes in plastid genomes.

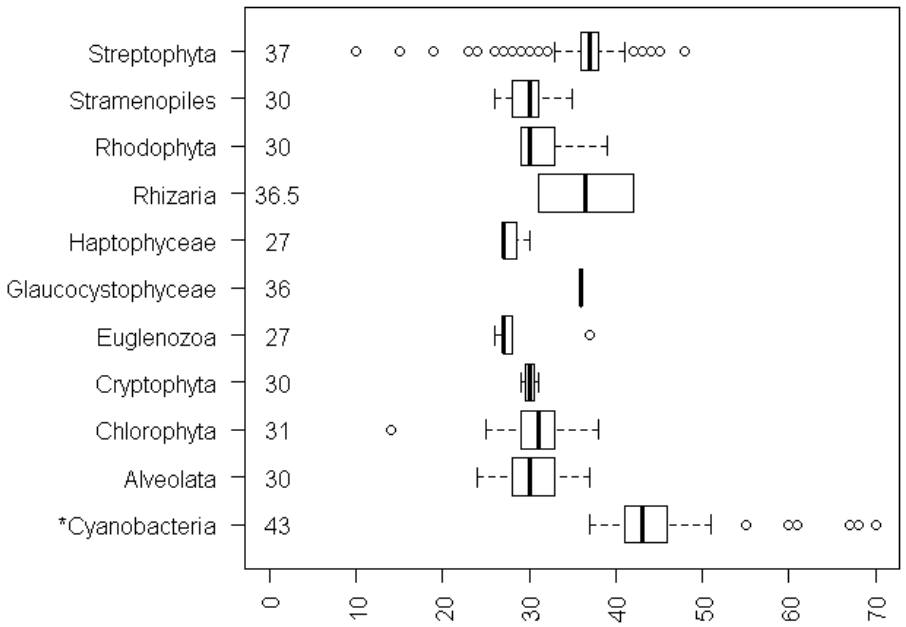


Figure 13 – Distribution of the number of tRNA genes

Figure 14 shows the distribution of the various tRNA isoacceptor species, which result from the number of the different anticodon sequences detected by tRNA_finder. The overall median of tRNA species is 29 (95% CI = 28.9 – 29.1), a number significantly lower than the 37 tRNA genes. Again, the results are clearly influenced by the plastids of Streptophytes, which have 7 tRNA genes located in the inverted repeats (IRs) that are duplicated. As a result, there are 30 different tRNA isoacceptor species in Streptophytes despite the 37 tRNA genes. The tRNA species are 29 because tRNA_finder skips the annotations for tRNA-fMet (formyl-Methionine). With regard to cyanobacteria,

the median number of different tRNA isoacceptors is 39 (95% CI = 38.7 – 39.4), ranging from a minimum of 34 to a maximum of 44 tRNA isoacceptors.

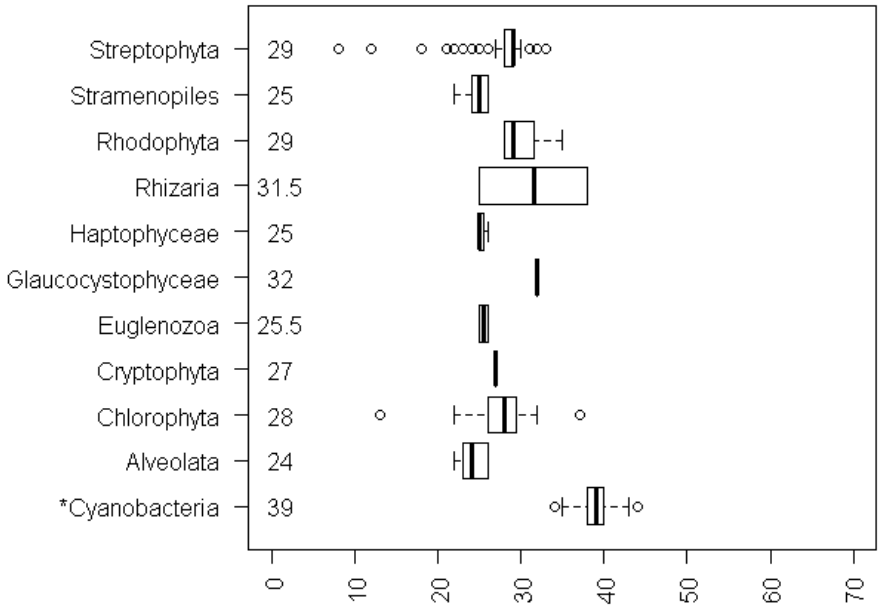


Figure 14 - Distribution of the number of tRNA isoacceptor species

On the basis of the existing annotations, 73 plastome sequences lack tRNA genes for at least one standard amino acid. Among these only 17 plastomes have no tRNA genes for two amino acids or more. However, 5 species show a consistent deficiency in their tRNA variety, losing the ability to decode from 5 to 12 standard amino acids (Table 12).

Table 12 - Plastomes with a considerable deficiency of tRNAs

Name	Notes	Division	tRNA species	N. of AAs without any tRNAs
Rhizanthella gardneri NC_014874.1	Parasite	Streptophyta	8	12
Bathycoccus prasinos NC_024811.1	Ambiguous sequence data	Chlorophyta	13	8
Selaginella moellendorffii NC_013086.1	Lycophytes	Streptophyta	12	8
Conopholis americana NC_023131.1	Parasite	Streptophyta	18	5
Epifagus virginiana NC_001568.1	Parasite	Streptophyta	18	5

Some plastid genomes such as that of *Cuscuta gronovii* encode at least 23 tRNA isoacceptors, the minimum amount of tRNA species still able to decode all codons by exploiting the superwobble rules. Though, by looking at which tRNA genes are available, the number of tRNA species does not always assure a complete translational ability (Table 13). Indeed, despite the fact that a number of plastomes encode 23 or more tRNA species, some amino acid lacks all its tRNA isoacceptors.

Table 13 - Plastomes lacking tRNA for two AAs.

Name	Division	tRNA species	N. of AAs without any tRNAs
Boulardia latisquama NC_025641.1	Streptophyta	21	2
Cuscuta gronovii NC_009765.1	Streptophyta	23	2
Cuscuta obtusiflora NC_009949.1	Streptophyta	23	2
Klebsormidium flaccidum NC_024167.1	Streptophyta	25	2
Micromonas sp. RCC299 NC_012575.1	Chlorophyta	22	2
Nageia nagi NC_023120.1	Streptophyta	29	2
Neottia nidus-avis NC_016471.1	Streptophyta	25	2
Orobanche purpurea NC_023132.1	Streptophyta	22	2
Orobanche ramosa NC_023465.1	Streptophyta	22	2
Pharus latifolius NC_021372.1	Streptophyta	24	2
Picea sitchensis NC_011152.3	Streptophyta	25	2
Triticum urartu NC_021762.1	Streptophyta	26	2

Only 14 plastomes (Table 14), including the cyanelle of *P. chromatophora*, have a number of tRNA isoacceptors equal to 32 or higher, suggesting that the standard wobble rules alone are not sufficient to decode all the codons.

Table 14 - Plastomes coding for a high number of tRNA isoacceptors

Name	Division	tRNA species
Paulinella chromatophora NC_011087.1	Rhizaria	38
Bryopsis hypnoides NC_013359.1	Chlorophyta	37
Galdieria sulphuraria NC_024665.1	Rhodophyta	35
Chlorokybus atmophyticus NC_008822.1	Streptophyta	33
Pyropia haitanensis NC_021189.1	Rhodophyta	33
Pyropia perforata NC_024050.1	Rhodophyta	33
Cyanophora paradoxa NC_001675.1	Glaucocystophyceae	32
Equisetum arvense NC_014699.1	Streptophyta	32
Equisetum hyemale NC_020146.1	Streptophyta	32
Mesostigma viride NC_002186.1	Streptophyta	32
Nephroselmis astigmatica NC_024829.1	Chlorophyta	32
Osmundastrum cinnamomeum NC_024157.1	Streptophyta	32
Psilotum nudum NC_003386.1	Streptophyta	32
Zygnema circumcarinatum NC_008117.1	Streptophyta	32

3.3.2 *tRNA isoacceptors among different groups*

As described in chapter 1, plastid genomes have lost many genes during evolution and tRNA genes have met the same fate. Figure 15 show a heatmap that summarises the median number of tRNA isoacceptors for every codon. At least 23 different types of tRNAs are generally conserved in all plastomes, among which Alveolata have exactly a median number of 23 tRNA species.

tRNA isoacceptors can be divided in three groups: conserved (group I), differentially conserved (group II), missing or lost (group III) (Table 15).

The first group (group I) includes 23 tRNA isoacceptors that have been conserved by all genomes. Near all the 23 conserved tRNAs have G or T in the first anticodon position, namely they can exploit wobble or superwobble pairings. Exceptions exist for the amino acids decoded by only one codon as occurs for tRNA(CAT)-Met and tRNA(CCA)-Trp.

Conversely, 29 tRNA species (group III) are missing in all plastomes, while the larger genomes of cyanobacteria (including the symbiont of *P. chromatophora*) lack 21 tRNA species. The missing tRNAs have A or C in the first anticodon position, so they cannot wobble. The only exceptions are tRNA(GCG)-Arg and tRNA(TCG)-Arg, which probably are not essential due to the fact that the high GC content may facilitate the use of the "two out of three" rule.

The last group (group II) comprises 17 tRNA isoacceptors that can be functionally replaced using the wobble or superwobble rules but that are still conserved in some genomes. For example, tRNA(GGA)-Ser may be not necessary since tRNA(TGA)-Ser can read all the TCN codons exploiting the superwobbling ("four way"). The same explanation is valid for all the amino acids that have four-codon boxes.

Interesting is the case of tRNA-Arg isoacceptors, since tRNA(ACG)-Arg is the only species encoded in plastomes of Alveolata, Euglenozoa, and Streptophyta it should exploit the "two out of three" rule to read CGN codons.

Table 15 – Genetic code and tRNA isoacceptors in plastid genomes.

First Base	Second Base: T	Second Base: C	Second Base: A	Second Base: G	Third Base
T	TTT F Phe	TCT S Ser	TAT Y Tyr	TGT C Cys	T
	TTC F Phe	TCC S Ser	**TAC Y Tyr**	**TGC C Cys**	C
	TTA L Leu	**TCA S Ser**	TAA * Ter	TGA * Ter	A
	TTG L Leu	TCG S Ser	TAG * Ter	**TGG W Trp**	G
C	CTT L Leu	CCT P Pro	CAT H His	**CGT R Arg**	T
	CTC L Leu	CCC P Pro	**CAC H His**	CGC R Arg	C
	CTA L Leu	**CCA P Pro**	**CAA Q Gln**	CGA R Arg	A
	CTG L Leu	CCG P Pro	CAG Q Gln	CGG R Arg	G
A	ATT I Ile	ACT T Thr	AAT N Asn	AGT S Ser	T
	ATC I Ile	ACC T Thr	**AAC N Asn**	**AGC S Ser**	C
	ATA I Ile	**ACA T Thr**	**AAA K Lys**	**AGA R Arg**	A
	ATG M Met	ACG T Thr	AAG K Lys	AGG R Arg	G
G	GTT V Val	GCT A Ala	GAT D Asp	GGT G Gly	T
	GTC V Val	GCC A Ala	**GAC D Asp**	GGC G Gly *	C
	GTA V Val	**GCA A Ala**	**GAA E Glu**	**GGA G Gly**	A
	GTG V Val	GCG A Ala	GAG E Glu	GGG G Gly	G

Bold: conserved tRNA(n. 23); Grey: missing/lost tRNA (n.29); Normal:differentially conserved(n. 9).
* Lost only in Alveolata.

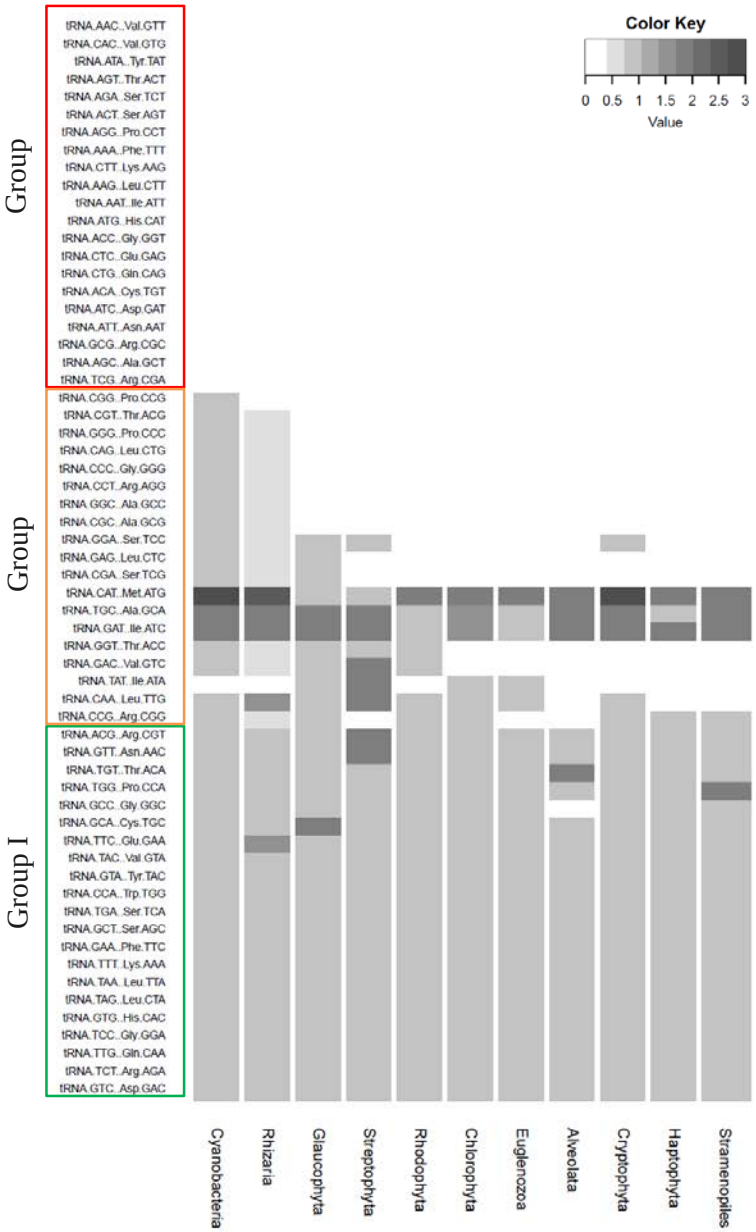


Figure 15 – Heatmap representation of the median numbero of all tRNA isoacceptors for every group.

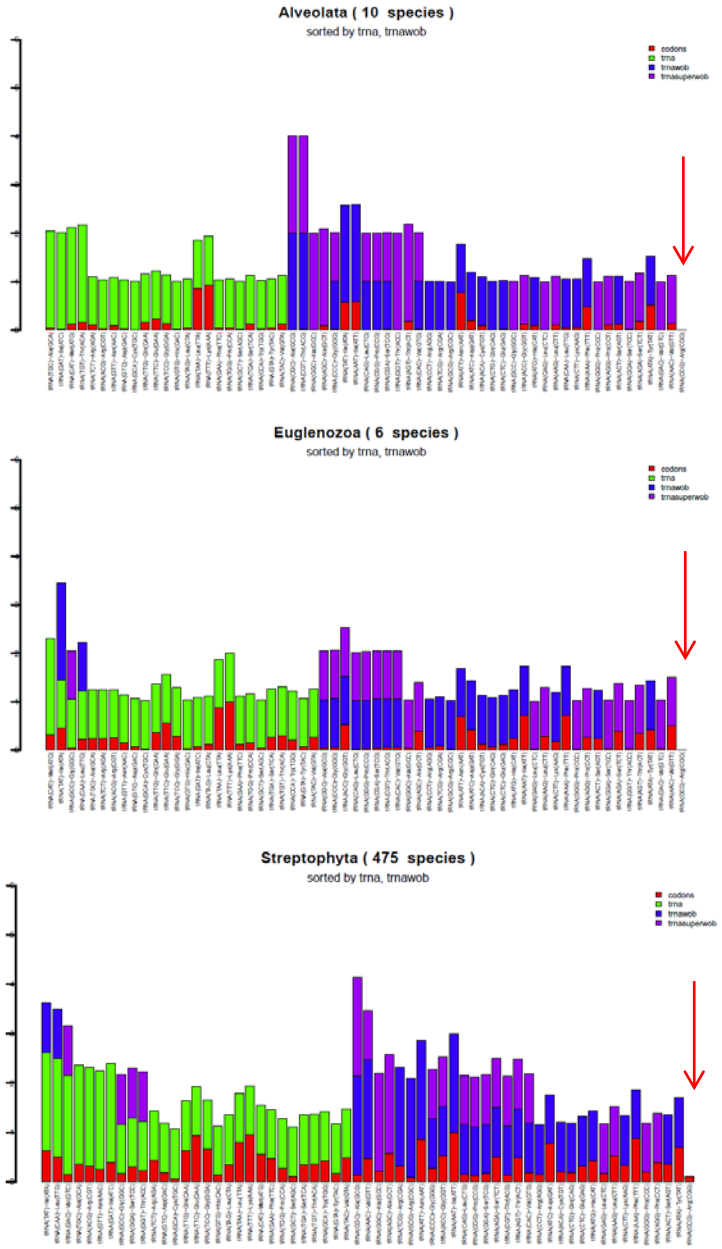


Figure 16 - Codon coverage in Alveolata, Euglenozoa, and Streptophyta. The arrows indicate the Arginine CGG codon, which lacks the corresponding tRNA.

These results prove that even with only 23 tRNA species all the 20 standard amino acids can be decoded, exploiting all the wobble and superwobble rules, in addition to the RNA editing. Nonetheless, it does not mean that all the plastomes encoding a minimum of 23 tRNA species are able to decode all codons. For example, as previously seen in Table 13, *Cuscuta gronovii* and *Cuscuta obtusiflora* own plastomes that encode 23 tRNA species but are unable to decode any codons of two amino acids (Alanine and Lysine) and several other codons (Figure 17).

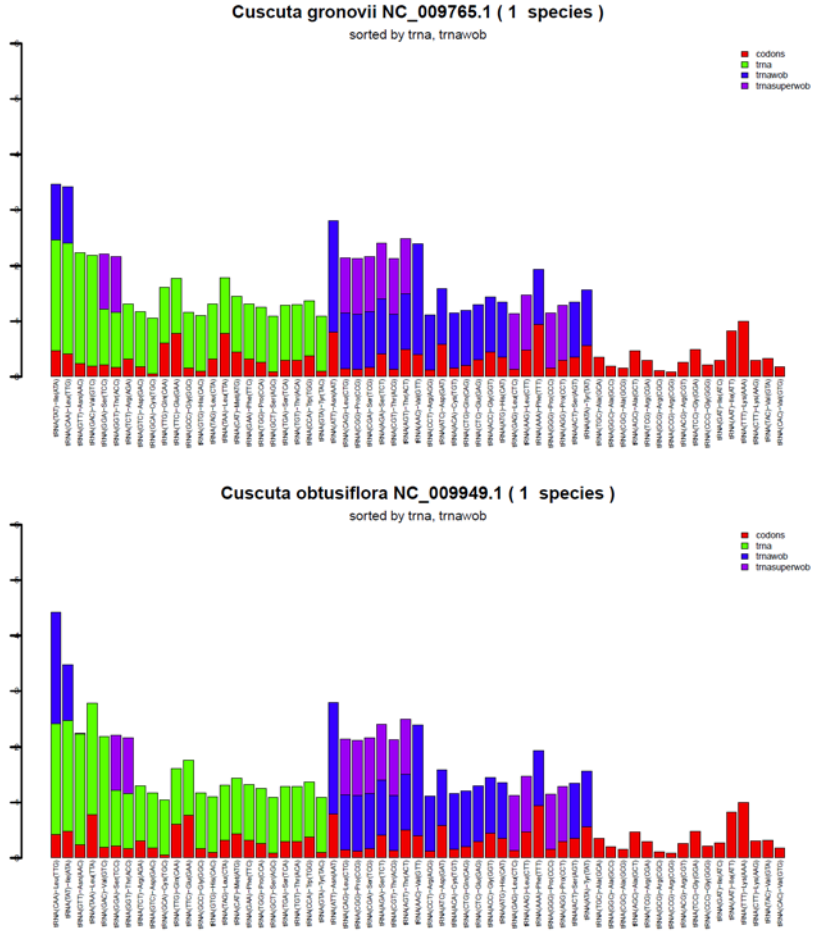


Figure 17 – Codon coverage of *Cuscuta gronovii* and *Cuscuta obtusiflora.*

With regard to the plastomes which lost many tRNA genes such as those in *Selaginella moellendorfii*, *Epifagus virginiana*, and *Rhizanthella gardneri*, they can decode only a limited number of codons even exploiting all the known wobble and superwobble rules. As shown in Figure 18, for such plastids the import of tRNAs from cytosol seems to be mandatory.

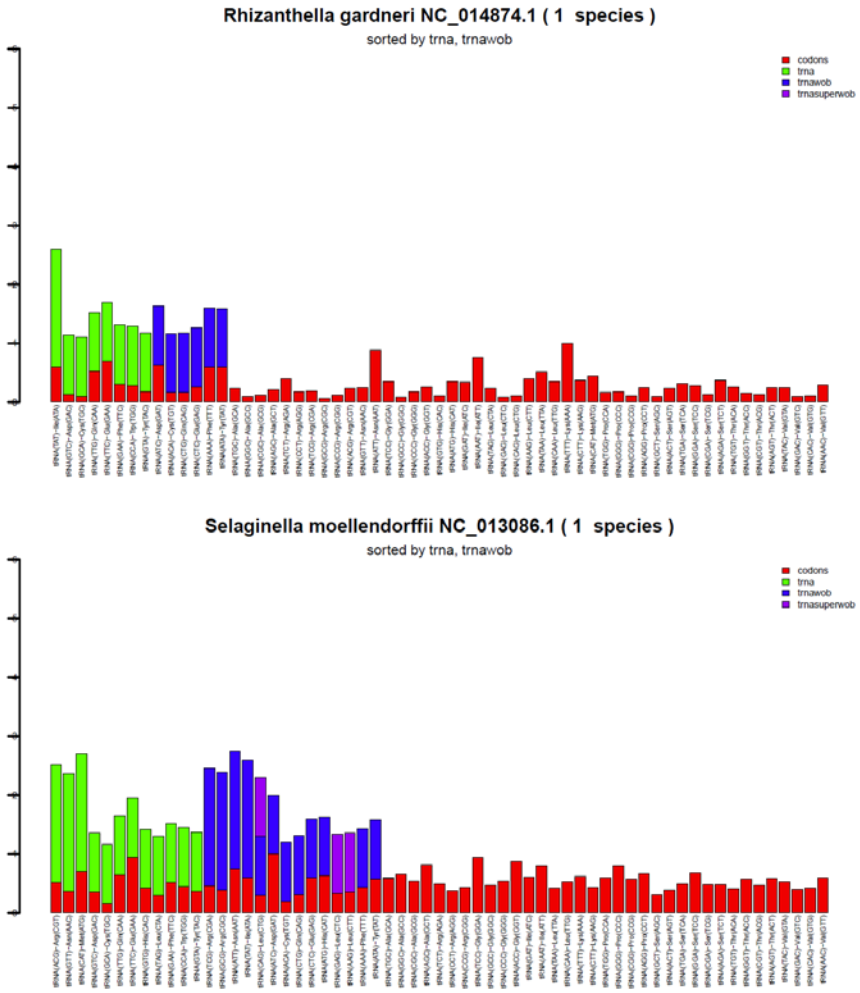


Figure 18 - Codon coverage of *Rhizanthella gardneri* and *Selaginella moellendorfii.*

Interestingly, both the reduced plastomes of *Rhizanthella gardneri* and *Selaginella moellendorfii* have conserved two tRNA-Glu, probably due to its role in pyrrole biosynthesis.

3.3.3 *Phylogenetic relationships and tRNA isoacceptors.*

As previously seen, different groups of organisms seems to have retained specific tRNA isoacceptors and with a specific copy number.

The difference in the copy number of every tRNA isoacceptors can be estimated by calculating a distance matrix. It is possible to have a graphical view of the distances between genomes by using a multivariate analysis called Non-metric Multi-Dimensional Scaling, described in chapter 2. The result of the NMDS is shown in Figure 19.

The NMDS obtained a stress value of 0.133 with two dimensions, which is very low considering the high number of points and the bi-dimensionality of the scaling (Sturrock and Rocha 2000). Significant differences ($p < 0.0001$) between the groups were confirmed by the PERMANOVA test, after having confirmed that there was no difference in dispersion between groups using the PERMDISP2 test ($p > 0.05$).

On the basis of their tRNA gene content, cyanobacteria appear as a distinct group on the NMDS map. Since the cyanelle of *P. chromatophora* is a cyanobacterial symbiont, its position within the cyanobacteria on the map is quite reasonable.

The only Glaucophyta for which the plastid genome has been annotated is *C. paradoxa.* On the NMDS map the point representing this unique plastome is barely visible but it is located close to the edge between cyanobacteria and plastid genomes. As previously shown in Figure 15, the plastome of *C. paradoxa* codes for 32 tRNA isoacceptors, which is a very high number compared to other plastid genomes.

Red and green algae together with other protists form a heterogeneous group of overlapping points. Probably, the difference among the plastomes of these groups is more subtle and then cannot be clearly visualised. However, some Rhodophyta are very close to Cyanobacteria in term of tRNA gene content. At the opposite, a number of Alveolata are very distant because of their reduced tRNA gene content.

The plastids of Streptophyta (i.e. chloroplasts) form a large group which is quite distinct from the plastids of protists. A few plastomes are very isolated or pretty distant from the other. Their position on the NMDS map reflects a particularly limited set of tRNA isoacceptors, indeed they correspond to the organisms listed in Table 12.

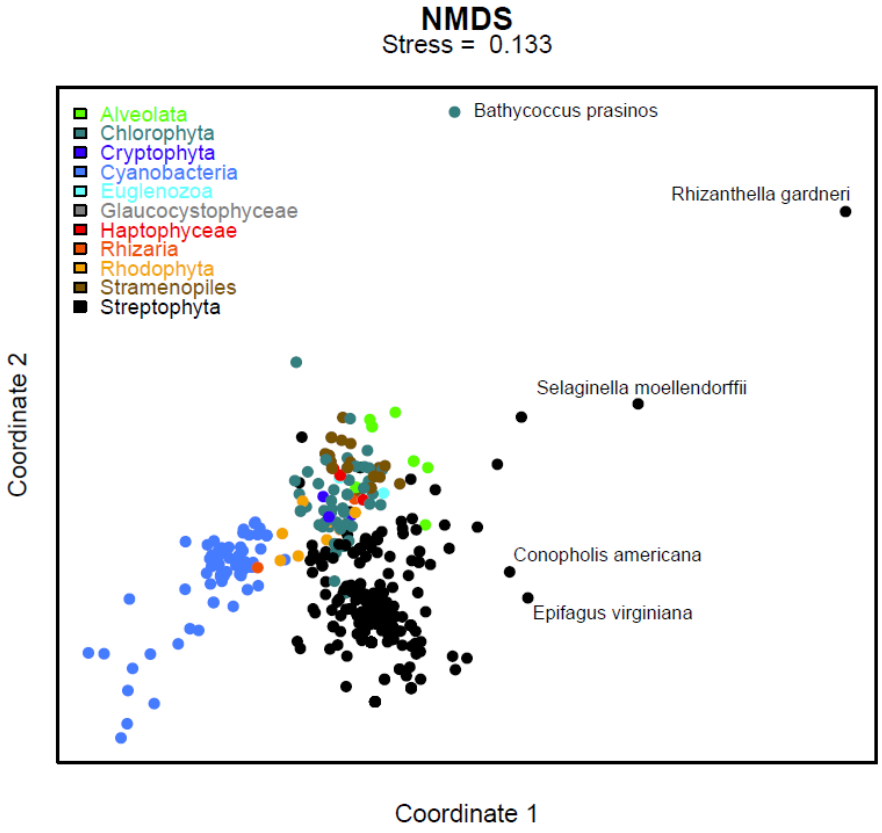


Figure 19 - NMDS based on differences in tRNA isoacceptors

The fitting of the variables onto the NMDS ordination provides an estimation of which one (i.e. tRNA gene copy number) is responsible of the main differences between the genomes. Table 16 reports the tRNA isoacceptor genes having the highest correlation (r^2>0.5) with the NMDS ordination and thus highly influencing the relative position of every plastid genome on the NMDS map.

Table 16 – tRNA with high correlation with the NMDS ordination

tRNA Anticodon	Amino acid Codon	r2	Pr(>r)
tRNA CAA	Leu TTG	0.703	0.0001
tRNA GAC	Val GTC	0.693	0.0001
tRNA GTT	Asn AAC	0.684	0.0001
tRNA CGT	Thr ACG	0.646	0.0001
tRNA TAT	Ile ATA	0.644	0.0001
tRNA GGC	Ala GCC	0.635	0.0001
tRNA ACG	Arg CGT	0.632	0.0001
tRNA CAG	Leu CTG	0.609	0.0001
tRNA CGG	Pro CCG	0.560	0.0001
tRNA CGC	Ala GCG	0.560	0.0001
tRNA CGA	Ser TCG	0.555	0.0001
tRNA CCG	Arg CGG	0.551	0.0001
tRNA CCC	Gly GGG	0.525	0.0001

tRNAs existing in all genomes but differing in their gene copy number are highlighted in grey.

As shown in Table 16, 13 tRNA obtained a correlation coefficient higher than 0.5, and 8 tRNA had values higher than 0.6. Most of the tRNA listed in Table 16 belongs to group II, which means that they are conserved and specific of only a few genomes.

On the contrary, two tRNAs, tRNA(GTT)-Asn and tRNA(ACG)-Arg, belong to group I, namely they are conserved in all genomes but in this case they have a different copy number. Since tRNA(GTT)-Asn and tRNA(ACG)-Arg are located on IRs, they are generally doubled in Streptophyta plastomes but, as described in the Introduction chapter, it is not the same for other groups.

3.4 tRNA gene copy number and codon usage

3.4.1 *Correlation at whole genome level*

In order to evaluate the relationship between tRNA gene copy number and mRNA codon usage, CORR_checker estimates the tRNA coverage for each amino acid codon. Such data allows visualizing how each codon can be decoded by tRNAs, whether it has a cognate tRNA or requires wobble and superwobble rules.

Figure 20 shows an example of data provided by CORR_checker. The codon usage is scaled between 0 and 1, from the least frequent to the most frequent codon. The gene copy number of tRNAs is represented by different colours: green for cognate tRNAs, blue for wobbling tRNAs, and purple for superwobbling tRNAs. Wobbling and superwobbling tRNAs are meant as those tRNA isoacceptors that can read codons other than their cognate codon by using wobble or superwobble rules. If there was a correlation between tRNA gene copy number and codon usage, tRNAs with a high copy number would correspond to the abundant (i.e. preferred) codons.

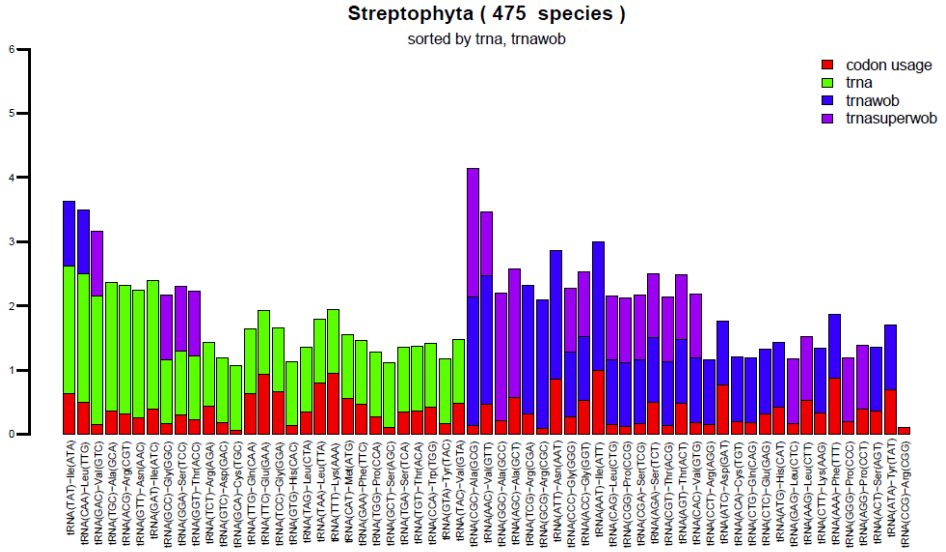


Figure 20 - Codon coverage and codon usage, example of Streptophyta.

Figure 20 helps to visualize a lot of information about tRNA genes and mRNA codons but, in order to mathematically assess the correlation between them, CORR_checker calculates the Spearman's correlation coefficient (*rho*) for each cyanobacterial and plastidial genome. Figure 21 shows the distribution of the correlation coefficients among the different groups of eukaryotic organisms. For each group the correlation was considered significant when the median of the Spearman's coefficients *rho* resulted greater than 0.215 (threshold of significance).

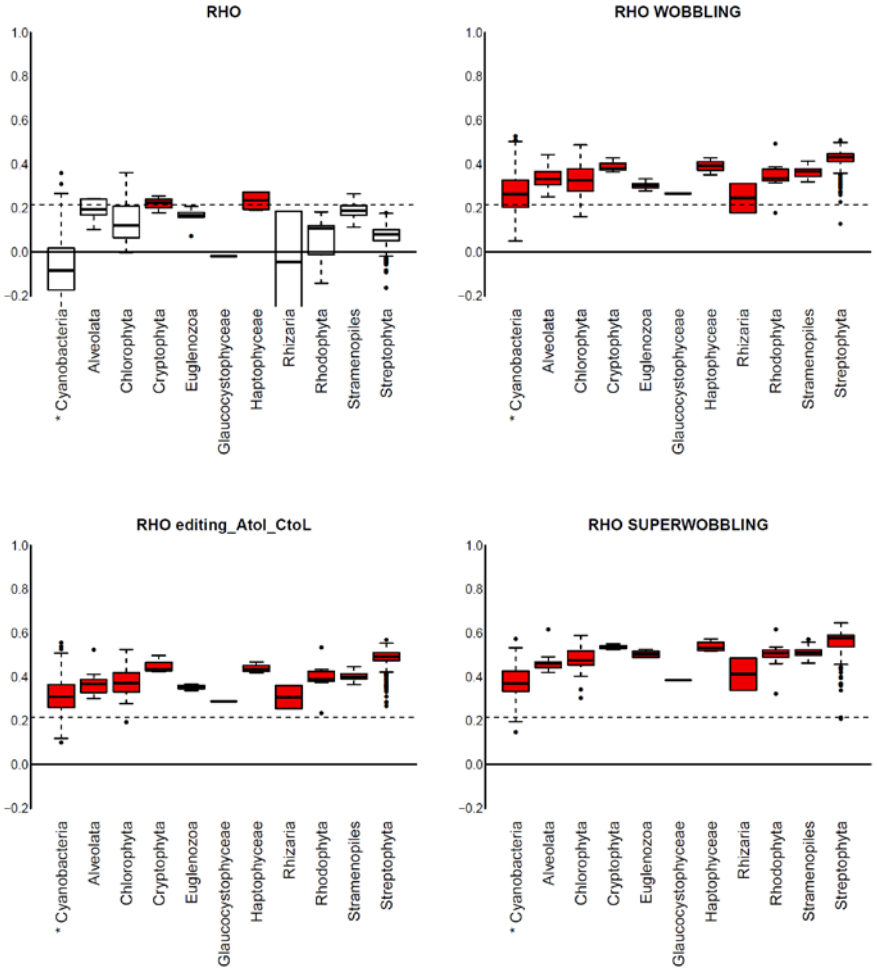


Figure 21 – Boxplots of the Spearman's correlation coefficients related to tRNA gene copy number and codon usage among different groups of eukaryotic organisms. Dashed line represents

the threshold (*rho* = 0.215) above which the correlation is considered significant (*p-value* < 0.05).

The first panel of Figure 21 shows the correlation, between the gene copy number of cognate tRNAs and the codon usage. The calculation of such correlation does not consider the effect of wobbling and superwobbling rules and, in this case, only the plastid genomes of Cryptophyta and Haptophyta show significant *rho* values. Nevertheless, the correlation is still weak and very close to the threshold of significance.

The results drastically change when considering the wobbling and superwobbling mechanisms, showing from weak to moderate correlation among all groups. Streptophyta shows the highest correlation coefficients when the wobble rules are taken into account, despite no correlation was observed by considering only the standard pairings.

Contrary to what was previously seen, plastomes with a reduced tRNA set such as *Rhizanthella gardneri* and *Selaginella moellendorfii* do not show significant correlation between codon usage and tRNA gene copy number when considering the superwobble rules (Figure 22).

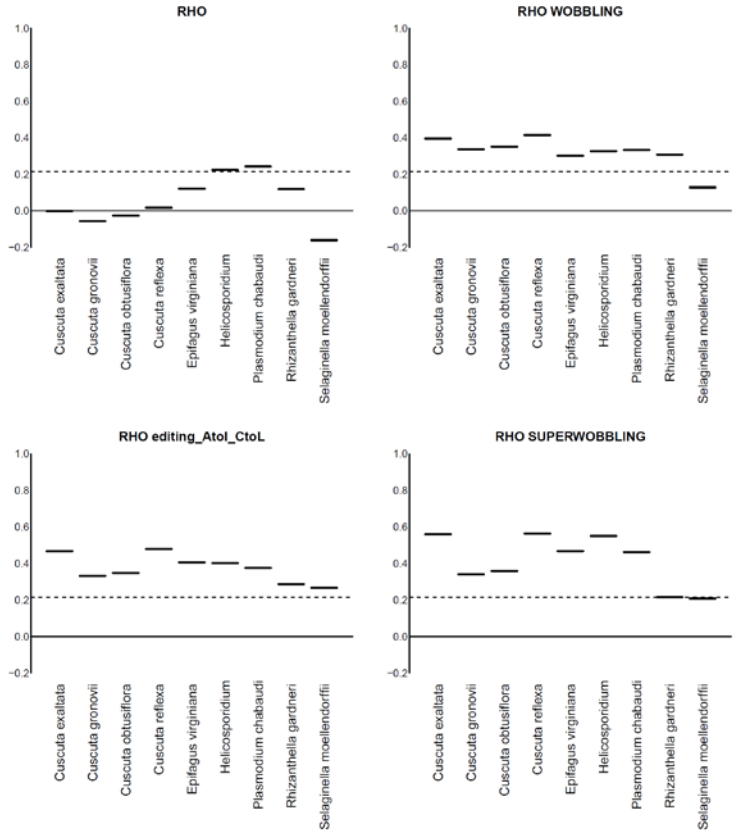


Figure 22 - Spearman's correlation coefficients related to tRNA gene copy number and codon usage in plastomes of particularly interesting organisms. Dashed line represents the threshold (*rho* = 0.215) above which the correlation is considered significant (*p-value* < 0.05).

3.4.2 *Correlation at single gene level*

The correlation between tRNA gene copy number and codon usage was estimated also at single-gene level. According to Green 2011, nearly 60 genes (Table 5) are almost universal in all plastomes, except in the most reduced ones. CODON_checker was used to calculate the codon abundance in every coding sequence, and then the correlation with the tRNA gene content was estimated with CORR_checker.

As it was done with the correlations at whole-genome level, the distribution of the Spearman's correlation coefficients was summarized by using boxplots (Figures 20-30).

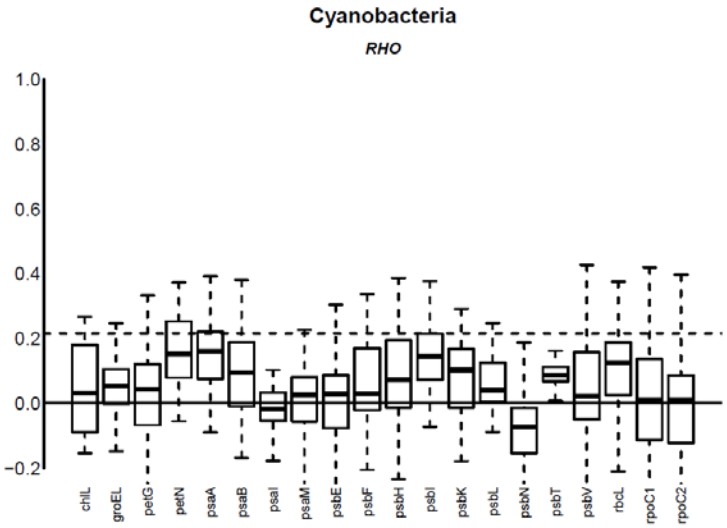


Figure 23 – Cyanobacteria, boxplots of the Spearman's correlation coefficients at gene-level (standard pairings). On average no significant correlations.

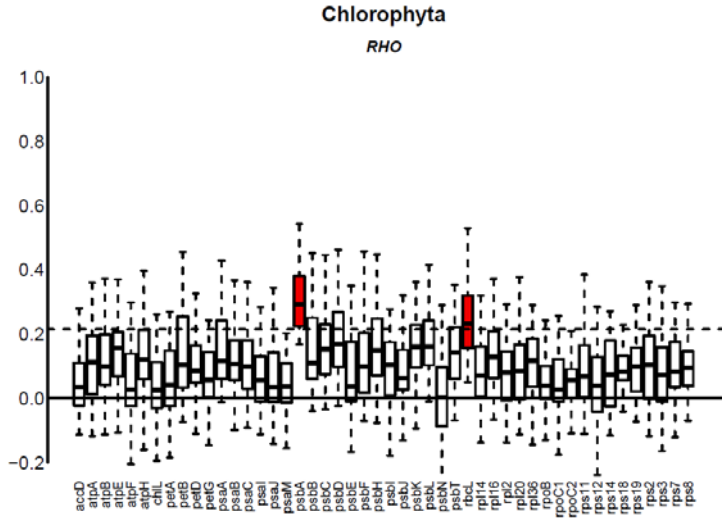


Figure 24 – Chlorophyta, boxplots of the Spearman's correlation coefficients at gene-level (standard pairings). Significant correlation for: *psbA* and *rbcL*.

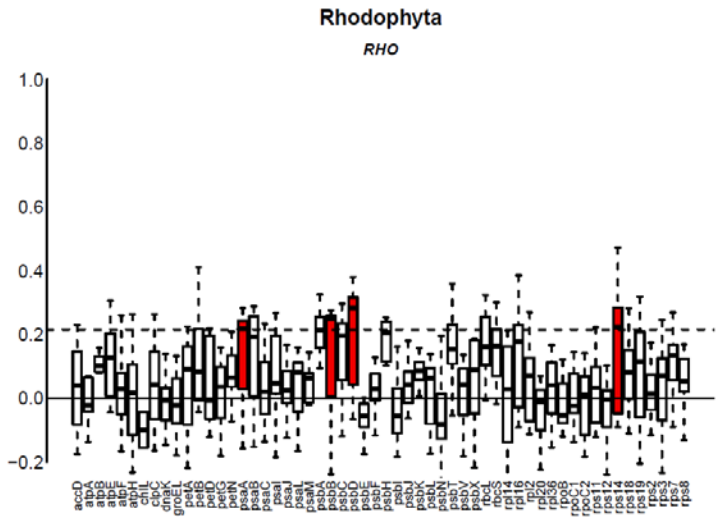


Figure 25- Rhodophyta, boxplots of the Spearman's correlation coefficients at gene-level (standard pairings). Significant correlation for: *psaA*, *psbB*, *psbD*, and *rps14*.

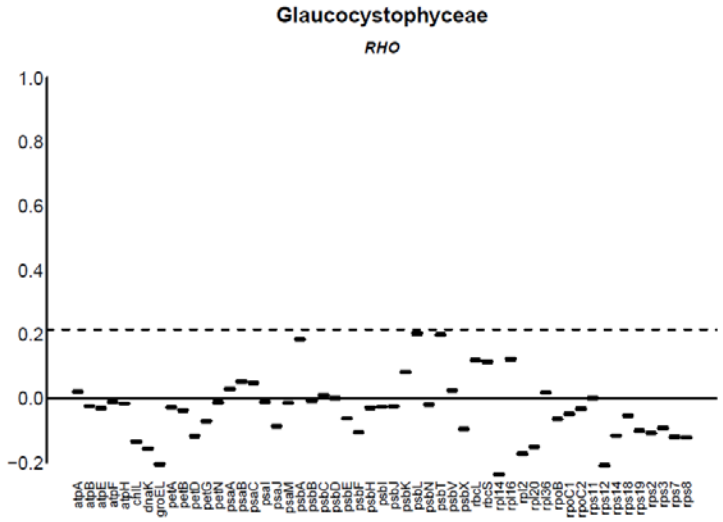


Figure 26 – Glaucophyta, Spearman's correlation coefficients at gene-level (standard pairings). No significant correlations.

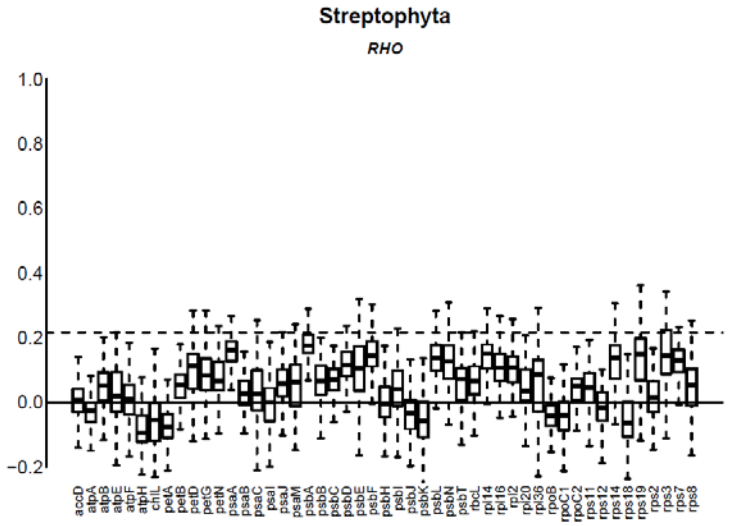


Figure 27 – Streptophyta, boxplots of the Spearman's correlation coefficients at gene-level (standard pairings). On average, no significant correlations.

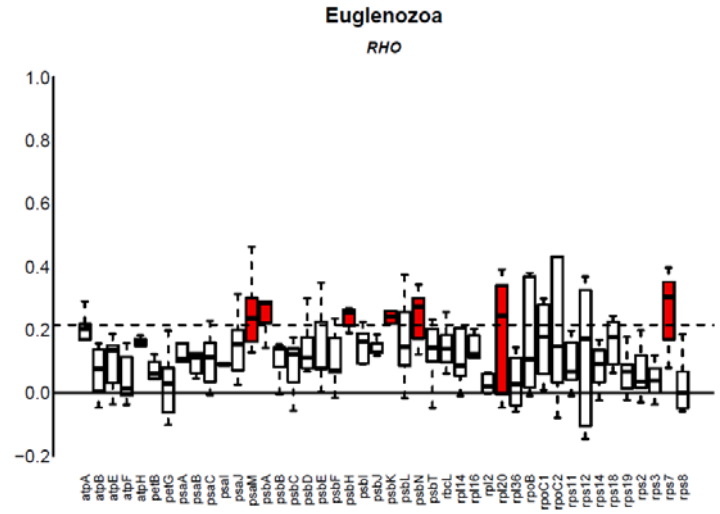


Figure 28 – Euglenozoa, boxplots of the Spearman's correlation coefficients at gene-level (standard pairings). Significant correlation for: *psaM*, *psbA*, *psbH*, *psbK*, *psbN*, *rpl20*, and *rps7*.

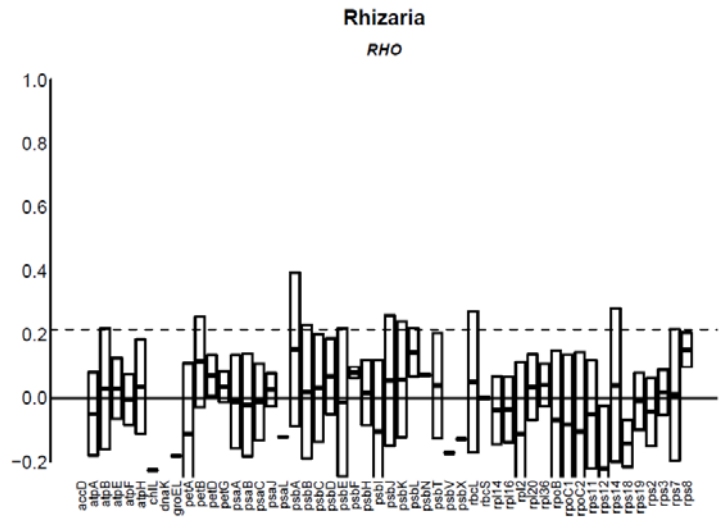


Figure 29 – Rhizaria, boxplots of the Spearman's correlation coefficients at gene-level (standard pairings). On average, no significant correlations.

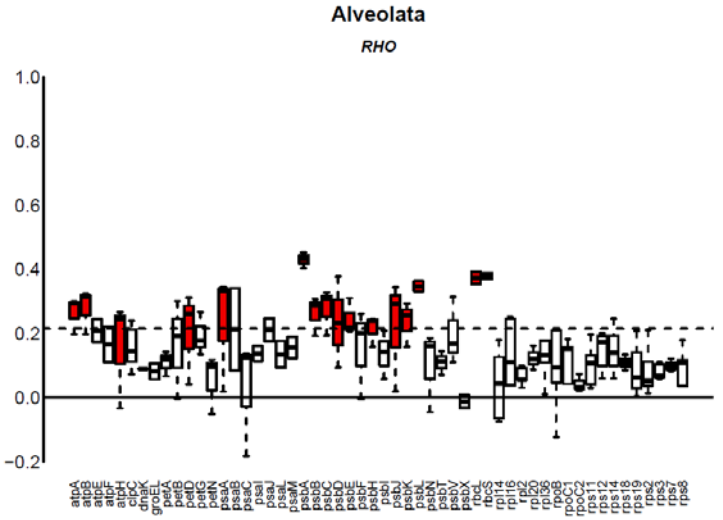


Figure 30 – Alveolata, boxplots of the Spearman's correlation coefficients at gene-level (standard pairings). Significant correlation for: *atpA*, *atpB*, *atpH*, *petD*, *psaA*, *psbA-E*, *psbJ-L*, *rbcL*, and *rbcS*.

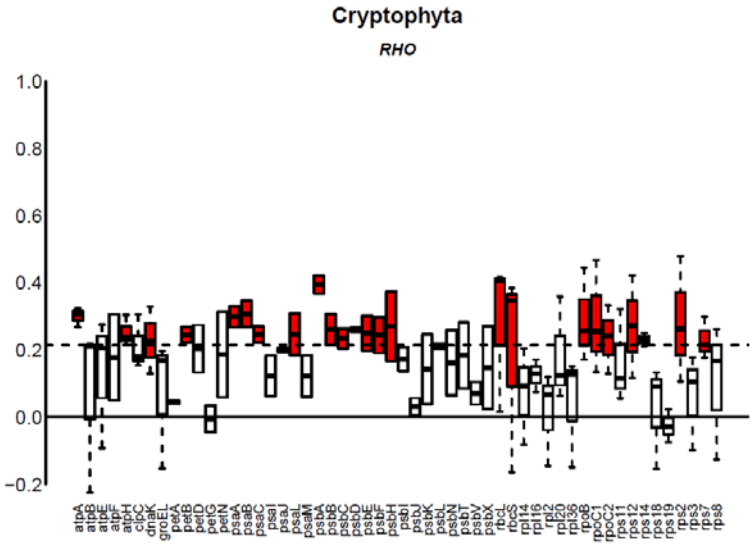


Figure 31 - Cryptophyta, boxplots of the Spearman's correlation coefficients at gene-level (standard pairings). Significant correlations for: *atpA*, *atpH*, *dnaK*, *petB*, *psaA*, *psaB*, *psaC psaL*, *psbA-H*, *rbcL*, *rbcS*, *rpoB*, *rpoC1*, *rpoC2*, *rps12*, *rps14*, *rps2*, and *rps7*.

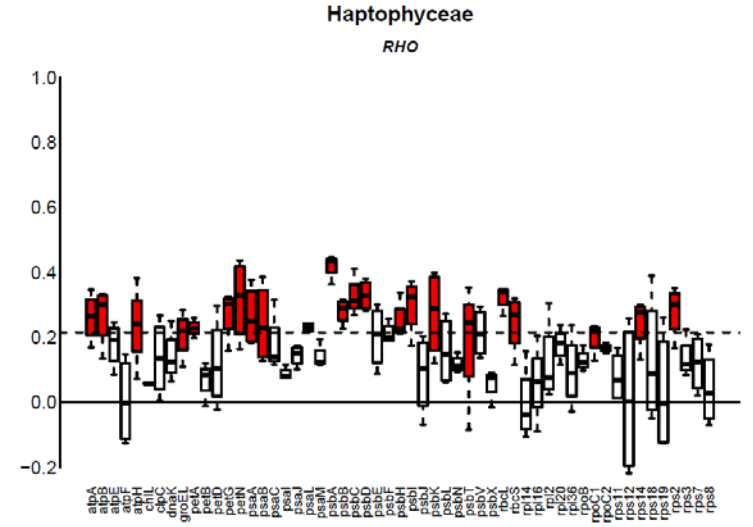


Figure 32 – Haptophyta, boxplots of the Spearman's correlation coefficients at gene-level (standard pairings). Significant

correlations for: *atpA, atpB, atpH, groEL, petA, petG, petN, psaA, psaB, psaL, psbA-E, psbH, psbI, psbT, rbcL, rbcS, rpoC1, rps14,* and *rps2.*

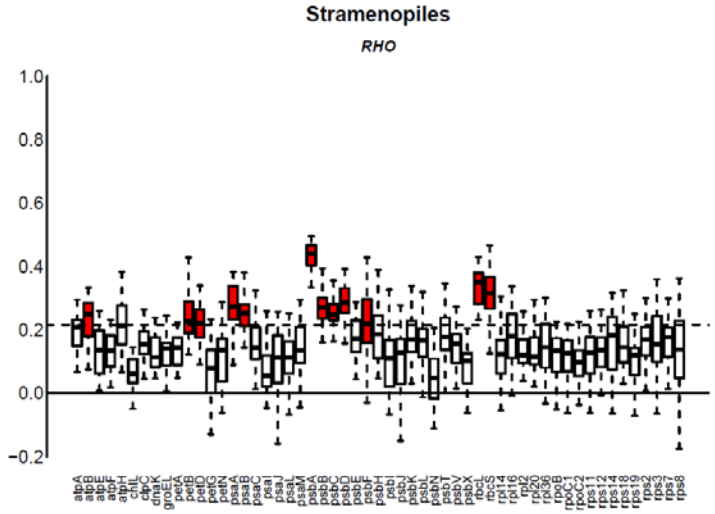


Figure 33 – Stramenopiles, boxplots of the Spearman's correlation coefficients at gene-level (standard pairings). Significant correlations for: *atpB, petB, petD, psaA, psaB, psbA-D, psbF, rbcL,* and *rbcS.*

In a few groups, the codon preference of a number of genes (coding sequences) is weakly or moderately correlated with the tRNA gene copy number. Apparently, this is true especially for the genomes of plastids originated by secondary endosymbiosis (Figures 25-30). The group of Rhizaria is composed by only two organisms due to the absence of other sequences in NCBI GenBank. Moreover, as already described, the plastid of *P. chromatophora* is not a "normal" plastid but a cyanobacterial symbiont. Therefore, the boxplots of Figure 29 show a median value which is not very informative.

Generally, the genes having a codon preference correlated with the tRNA gene copy number express proteins that play a central role in the main functions of plastids.

4 DISCUSSION

4.1 Plastids and their genomes

4.1.1 *Importance of plastome sequencing projects*

Plastids are a family of organelles found mainly in the cells of plants, algae and other protists. Chloroplast is the most popular type of plastid, as well as photosynthesis is the most popular function of plastids, despite not all plastids carry out photosynthesis.

According to the endosymbiotic theory, plastids originated 1.6 billion years ago from the engulfment of cyanobacterium-like organisms by eukaryotic organisms (Yoon et al. 2004). These events of primary endosymbiosis gave rise to the main groups of plastids: namely Glaucophyta, Rhodophyta, and Chlorophyta (McFadden and van Dooren 2004). Subsequent events of secondary endosymbiosis led to the formation of the plastids found in Euglenozoa, Cercozoa, Alveolata, and Chromista (McFadden 2001).

From their primordial bacterial ancestor, plastids inherited a genome known as “plastome”, which was first isolated in 1960s. The development of DNA sequencing technologies in 1980s allowed the beginning of plastome sequencing projects and nowadays the NCBI Genome database lists nearly 1,000 complete plastome sequences.

During the evolution, plastids adapted their genomes in order to better interact with their host. As a consequence, many genes were lost and transferred to the nuclear genome by horizontal gene transfer, so that the size of plastomes consistently shrank. The genes still present in plastomes are involved mainly in essential processes such as gene expression and photosynthesis.

Having information on several plastome sequences is extremely important firstly from a purely biological perspective. Indeed, plastomes represent minimal genomes which conserve and effectively express genes fundamental for the plastid functions. Plastid genome might be considered as a branch of the nuclear genome and understanding why some genes have been conserved while others have been transferred into the nucleus is extremely interesting. Furthermore, the plastid genome is a former prokaryotic genome adapted to be functional in a eukaryotic environment, a very intriguing example of integration. Another peculiar aspect of plastomes is their nucleotide substitution rate, which places in the middle between nuclear genome and mitochondria genome. Finally, the information stored in plastid genes is fundamental for studying the evolution of plastids and their endosymbiotic origin.

From a biotechnological point of view, plastids and their genomes are a valuable resource to improve the agronomic potential of plants or produce biomolecules (Maliga 2014). The first purpose relies on the reduced risk of spreading transgenes by modifying plastid genomes instead of nuclear genome. On the other hand, a number of plastid features are convenient for producing and harvesting bioproducts, among which the high copy number of plastomes and the confined space of the organelle.

4.1.2 *GenBank annotations and tRNA genes*

While it has been estimated that more than 95% of the plastid proteome come from outside of the organelle, there is no evidence of import of rRNA and tRNA molecules into plastids, therefore they are supposed to be transcribed exclusively from plastid genes (Lung et al. 2006).

Knowing exactly which tRNA genes and how many copies of them are encoded in plastomes becomes essential to assess if plastomes are able to decode every amino acid codon. Unfortunately, the absence of a standard format for annotating the codon specificity in the GenBank records generates serious problems when such data are to be used in bioinformatic analyses.

Furthermore, the annotation guidelines provided by NCBI GenBank are not sufficiently strict and several tRNA annotations lack information about the codon specificity. Since tRNAscan-SE has been widely used to annotate the tRNA annotations existing on GenBank, it would be quite easy for the authors to include also the anticodon sequence in the sequence submission. As described in the Results chapter, to date nearly the 40% of tRNA annotations in plastome sequence records do not have such information. To fix this issue, it is necessary to run again tRNAscan-SE on the annotated tRNA genes and retrieve the anticodon sequence.

tRNA_finder, included in the tRNA tools developed for this thesis, uses a dedicated set of regular expressions in order to read all the variety of tRNA gene annotations found in NCBI GenBank and, in addition, retrieves the missing anticodon sequences through tRNAscan-SE. Developing the software required a lot of effort and several hundreds of lines of code written in the Perl language.

In conclusion, a critical revision of the annotation procedure for tRNA genes seems mandatory, especially when the primary purpose of GenBank is providing “curated sequence data and related information for the community to use as a standard”.

4.2 tRNA tools and its contribution to the scientific research

The large amount of data stored in online and freely available databases represents a precious resource for every scientist but, on the other hand, retrieving specific information and performing statistical analyses became quite hard or even impossible without the help of bioinformatic tools.

In this thesis, studying the genomes of cyanobacteria and plastids in term of their tRNA gene content and codon usage required the use of bioinformatic tools which had not been developed yet. Indeed, most of the studies previously made on the tRNA gene content of genome sequences relied on dedicated databases such as GtRNAdb (Chan and Lowe 2009), which does not include organellar genomes. Moreover, the sequencing rate of organellar genomes has increased constantly over time, so that it became more reliable to retrieve data directly from GenBank instead of using derived databases.

tRNA tools can use GenBank flat files as input for finding tRNA gene copy number and codon usage in all kinds of genome, including plastid and mitochondria genomes. Additionally, the Perl programming language is multiplatform, as long as the Perl interpreted is installed, and the hardware requirements are minimal.

tRNA_finder provides a lot of information about tRNA genes such as the total number of tRNA genes, the tRNA gene copy number, and the codon specificity. CODON_finder is able to compute codon usage statistics on whole genomes, by taking all the protein coding sequences, or on a chosen group of genes as well as single genes. Such information is extremely useful for a variety of studies, for example to investigate tRNA gene content and phylogenetic relationships.

For the purposes of this thesis, CORR_checker represents the central software of tRNA tools, as it estimates the correlation between the tRNA gene copy number in a given genome and the codon usage in its protein coding genes. So far, many methods have been created to estimate the codon usage bias, such as the widely used CAI (Codon Adaptation Index), but no other computer program directly correlates the tRNA gene copy number with the codon abundance. This avoids choosing any reference gene sets and do not rely on *a priori* knowledges.

4.3 tRNA genes

4.3.1 *tRNA gene content*

tRNA tools allowed to explore the tRNA gene content of nearly 80 cyanobacteria genomes and 600 plastid genomes.

First of all, it is important to highlight that considering only the total number of tRNA genes would be misleading in evaluating the tRNA gene content of a genome. In fact, the median value of the annotated tRNA genes is 37 (95% CI = 36.7- 37.3) for plastid genomes and 43 (95% CI = 42.2- 43.9) for cyanobacteria genomes, but the median number of encoded tRNA species are 29 (95% CI = 28.9 – 29.1) and 39 (95% CI = 38.7 – 39.4), respectively. From this perspective, it appears evident that plastid genomes must use the extended wobble rules to translate mRNA, since the minimum number of tRNA species required by the standard wobble rules is 32.

On the basis of the existing annotations, 10% of the plastome sequences lack enough tRNA genes for translating all the 20 standard amino acids. In particular, the plastomes of *Rhizanthella gardneri*, *Conopholis Americana*, *Epifagus virginiana*, *Selaginella moellendorfii*, and *Bathycoccus prasinos* show a consistent deficiency in the variety of their tRNA isoacceptors, so that they could not decode several standard amino acids.

A consistent lack of tRNA genes may be related to a loss of plastid functionality, as occurs in parasite organisms such as *R. gardneri*, *C. Americana*, or *E. virginiana*. On the other hand, the plastome sequence of *B. prasinos* contains a large block of ambiguous nucleotides (n) that could have caused missing annotations. Other cases of defective translational apparatus, as that in the plastids of *Selaginella moellendorfii*, are well known in literature but still not well clear

Furthermore, although a number of plastid genomes such as that of *Cuscuta gronovii*, *Cuscuta obtusiflora*, *Nageia nagi*, and others, encodes at least 23 tRNA species (theoretically the minimum number to decode all codons by exploiting the superwobble rules), they are still not enough for decoding all the standard amino acids.

In conclusion, in case of an insufficient number of tRNA species, two explanations are given: a loss of plastid functionality or the existence of tRNA import from the cytosol. The first hypothesis could be possible only for a limited number of organisms such as the parasites which rely on their host. The tRNA import has never been observed in plastids but it is commonly accepted that in mitochondria this phenomenon occurs. Since the wide range of functions ascribed to plastids, it seems reasonable to think that in most cases their functionality has not been completely lost and their defective tRNA gene content may be compensated in some way.

Finally, a number of errors in the GenBank annotation of tRNA genes were found during the work of this thesis and probably many others still remains unknown. Problems could have occurred at different levels including the sequencing and the annotation procedures as well as the final submission of the data. Certainly, more detailed indications for annotating tRNA genes and a critical revision of the existing ones could drastically improve the reliability of GenBank.

4.3.2 *tRNA isoacceptors*

tRNA_finder allowed estimating the tRNA gene copy number in several cyanobacteria and hundreds of plastid genomes belonging to different eukaryotic supergroups. By clustering this data, it was possible to identify three typical categories of tRNA genes: generally conserved, differentially conserved and generally lost.

Firstly, only a few plastomes have a number of tRNA isoacceptors equal to 32 or higher, confirming the fact that the standard wobble pairings alone could not assure the decoding of all codons.

On average, 23 tRNA genes resulted conserved in all plastomes, as well as cyanobacteria. Noteworthy, near all the conserved tRNA genes have anticodons with G or T bases in the first position of the triplet in order to exploit wobble or superwobble pairings, except for the amino acids coded by single codons (e. g. Methionine). Therefore, it is no wonder that several not essential tRNA genes have been lost in favour of conserving a minimal but still effective tRNA set able to carry out the translation process.

On the other hand, the category of generally lost tRNA genes includes tRNA that cannot wobble, since they have A or C bases in the wobble position. The only exceptions are tRNA(GCG)-Arg and tRNA(TCG)-Arg, which are generally considered not essential since that tRNA(ACG)-Arg can exploit the "two out of three" rule to read all the CGN codons (Delannoy et al. 2009b). tRNA(ACG)-Arg is the only species encoded in plastomes of Alveolata, Euglenozoa, and Streptophyta. The post-transcriptional editing of Adenosine to Inosine allows the reading of only three codons (CGT, CGC, and CGA). Theoretically, the fourth codon, CGG, could not be read by tRNA(ICG)-Arg and wobble and superwobble rules do not apply to this case (Figure 16). A given explanation is that tRNA(ACG)-Arg can exploit the "two out of three" rule, favoured by the GC pairings occurring between codon and anticodon (Delannoy et al. 2009; Alkatib et al. 2012b).

On the basis of these results, it was possible to identify a set of 23 tRNA genes which seems to be universally shared by most of the plastid genomes. *In vivo* experiments made on plastids of *Nicotiana tabacum* led to

similar conclusions (Alkatib et al. 2012b; Rogalski et al. 2008), which can be here extended to all plastid genomes.

However, it does not mean that all the plastomes encoding a set of 23 tRNA species are able to decode all codons. Indeed, even if the plastomes of *Cuscuta gronovii* and *Cuscuta obtusiflora* encode 23 tRNA species, no tRNAs seems to be able to decode any codons of Alanine and Lysine and several other codons of different amino acids (Figure 17).

Finally, it is well known that some tRNA does not have only a translational function. For instance, tRNA-Glu is involved in the pyrrole biosynthesis. Interestingly, both the reduced plastomes of *Rhizanthella gardneri* and *Selaginella moellendorfii* have conserved two genes of tRNA-Glu.

Some years ago it has been suggested that tRNA-Glu might be the only gene found in all plastid genomes (Barbrook et al. 2006). However, currently many plastome sequences available on GenBank seem to lack genes coding for tRNA-Glu (e.g. NC_024286.1). Although annotation errors cannot be excluded, the essentiality of tRNA-Glu in plastids could again suggest the occurrence of tRNA import mechanisms.

4.3.3 *Phylogenetic relationships and tRNA isoacceptors*

Previous studies showed that a different tRNA gene composition in nuclear genomes characterises each of the three kingdoms of life, namely Bacteria, Eukarya, and Archaea (Novoa and Pavon-Eternod 2012). The work of this thesis provides a further advance in the use of the tRNA gene composition as phylogenetic tool.

The initial purpose of calculating the distance between the tRNA gene content of different genomes was to study and identify peculiar traits of the plastids originated from the main lineages (glaucophytes, red algae, and green algae). Analytical techniques generally used in ecology, as the Bray-Curtis distances and the multidimensional scaling, were applied to genetic data to assess and visualise existing differences.

As expected, since their high number of tRNA genes, cyanobacteria genomes form a distinct group from plastid genomes. The only exception is the genome of the cyanelle of *Paulinella chromatophora*, which is actually a cyanobacterial symbiont. Another defined group is formed by the genomes of red algae, green algae, and other protists. The heterogeneity of this group is probably caused by the subtle difference in term of tRNA gene content, suggesting a convergent conservation of specific tRNA species.

The plastomes belonging to Streptophyta appeared to have distinctive traits from the others, since that they form a pretty well-defined group in the NMDS map. An explanation may be found in the fact that Streptophyta are

characterised by 7 double-copy tRNA genes located on the IRs, as observed in *Nicotiana tabacum*. Since plastomes may have different types of IR or do not have IRs at all, their gene content appears to be a distinctive feature of Streptophyta plastid genomes.

The results of the NMDS ordination was further analysed in order to find which tRNA genes mostly influenced the differences among the groups of organisms. The loss or the conservation of certain tRNA genes are a characterizing feature the plastomes from different supergroups but also the copy number of other tRNA genes, especially tRNA(GTT)-Asn and tRNA(ACG)-Arg. In particular, tRNA(ACG)-Arg is thought to be able to read all the CGN codons for Arginine.

4.3.4 *Correlation between tRNA gene copy number and codon preference*

A certain degree of correlation between codon usage and tRNA abundances have been shown in several studies (Ikemura 1985; Novoa and Ribas de Pouplana 2012; Novoa and Pavon-Eternod 2012). Since the copy number of tRNA genes could influence the abundance of tRNA molecules, it appears reasonable to think that tRNA gene copy number and codon usage are correlated (Tuller et al. 2010a, 2010b).

On the other hand, the wobble base pairings are known to reduce the translation efficiency when compared to the standard base pairings (Rogalski et al. 2008). Therefore, the best translation efficiency should occur when the number of preferred codons is correlated with the gene copy number of their cognate tRNA. The computer program named CORR_checker, included in tRNA tools, was developed in order to test for this correlation.

At a global genome-level, the results obtained in this thesis showed that the tRNA gene copy number and the codon usage generally do not correlate when only the standard pairings are taken into account. A significant but weak correlation was observed only in Cryptophyta and Haptophyta, which own plastids derived from a secondary endosymbiosis (Lee 2008) and have been recognized as closely related sister groups, although haptophytes lost their nucleomorph (Patron et al. 2007). The distinctive features of these two groups of organisms, such as the presence of the nucleomorph and the number of membranes, might reduce the capability to import nucleus-derived molecules and hence the need to increase the translation efficiency.

The results drastically change when the wobbling and superwobbling mechanisms are considered, showing from weak to moderate correlation among all groups. Streptophyta shows the highest correlation coefficients when the

wobble rules are taken into account, despite no correlation is observed by considering only the standard pairings.

Among Streptophyta, plastomes with a reduced tRNA gene set such as *Rhizanthella gardneri* and *Selaginella moellendorfii* do not show significant correlation between codon usage and tRNA gene copy number when considering the superwobble rules. Reduced genomes are supposed to use extensively the wobble rules in order to compensate the lack of tRNA genes, so the case of *R. gardneri* and *S. moellendorfii* could appear strange. Noteworthy, such plastomes underwent a great tRNA gene loss, so that a codon usage optimization may be of no use in this case.

Lastly, plastid genomes of Cyanobacteria and Rhizaria showed low correlation scores either with or without taking into account the wobble rules. This two groups include genomes with a high number of tRNA and protein-coding genes, indeed Rhizaria comprises only two plastid genomes and one of them belong to the cyanelle of *Paulinella Chromatophora.* Considering the variety of expression levels corresponding to such a high number of genes, an absent or weak correlation at genome-level could be expected.

At single-gene level, it was possible to have a deeper insight into the correlation levels in each group of plastid genomes. As occurred at global genome level, none of the genes analysed the plastomes of Cyanobacteria, Rhizaria, Glaucophyta, and Streptophyta, showed significant correlation between codon usage and tRNA gene content considering only the standard pairings.

On the other hand, a number of genes in the plastomes of Chlorophyta, Rhodophyta, Alveolata, Excavata, and Stramenopiles showed significant correlation coefficients, despite an absence of correlation at global genome-level. Cryptophyta and Haptophyta showed the highest number of genes which resulted to be correlated with the tRNA gene content, as already shown at genome-level.

Commonly, the genes with significant correlation encode subunits of photosystem I, photosystem II, RuBisCO enzyme, ATPase, or Ribosome. It is important to highlight that the correlation resulted by considering only the standard pairing rules, which are the most efficient for translating the codons. Noteworthy, among all the genes tested in Chlorophyta, only *psbA* and *rbcL* showed a codon composition significantly correlated with the corresponding tRNA gene content. In addition, significant correlations were observed only for a few genes in Rhodophyta and for no genes in Glaucophyta.

Such results led to the conclusion that secondary plastids, deriving from either the green or the red lineages, have conserved a codon composition and a tRNA gene content which correlates each other more than how it happens in primary plastids. Moreover, *psbA* was showed to be the gene with the most optimized codon composition, which should result in high translation efficiency. The D1 reaction center of photosystem II encoded by *psbA* is known

to undergo oxidation damage during light absorption, hence its high turnover rate. Therefore, the correlation observed between codon composition and tRNA gene content is in accordance with the known high expression level of *psbA*, together with its high Codon Adaptation Index (Hecht et al. 2012).

In conclusion, it appears reasonable that a significant degree of correlation between the synonymous codon composition and tRNA gene composition has been conserved for a number of genes that have to be translated with high efficiency. Conversely, genes with a lower rate of expression are not optimized in this sense and rely more on the wobble and superwobble rules to be effectively translated.

Conclusions

The development of bioinformatic tools able to handle thousands of tRNA annotations from plastid genomes, allowed studying the tRNA gene content, the codon usage of protein-coding genes, and the relationship between them. Assuming that translationally optimal codons are related to the most abundant tRNA isoacceptors and that tRNA abundance is affected by the gene copy number, it was assessed if the codon composition of highly expressed genes is related to the tRNA gene content.

After having analysed about 600 plastid genomes and 80 cyanobacteria genomes it was shown that despite the reduced number of tRNA genes, most plastid genomes can effectively translate all the 20 standard amino acids by exploiting the wobble and the extended wobble rules. On the other hand, a number of plastid genomes have a limited repertoire of tRNA genes, so that the existence of tRNA import into plastids cannot be excluded.

The tRNA gene content was used also to study evolutionary relationships through the application of multivariate statistics to genomic data. Distinctive traits among different groups of eukaryotic organisms were observed, suggesting that this new approach has great potentiality. Noteworthy, a set of 23 tRNA genes was recognised as universally shared by most of the plastid genomes.

In addition, tRNA(ACG)-Arg was found to be the only tRNA available to decode Arginine codons in Alveolata, Excavata, and Streptophyta, thus confirming the essentiality of the “two out of three” rule. Besides, a number of plastome sequences lack the genes coding for tRNA-Glu which is involved in the pyrrole biosynthesis. Although annotation errors cannot be excluded, the essentiality of tRNA-Glu in plastids could again suggest the occurrence of tRNA import mechanisms.

With regard to the hypothesis that the use of specific synonymous codons could be related to the most abundant tRNA isoacceptors, it was shown that at genome-level tRNA gene copy number and the codon usage generally do not correlate when only the standard pairings are taken into account. At the opposite, the correlation become significant when the wobbling and superwobbling mechanisms are considered, except for plastomes with a reduced tRNA gene set such as *Rhizanthella gardneri* and *Selaginella moellendorfii*.

At singe-gene level, the genes encoding for the subunits of photosystem I, photosystem II, RuBisCO enzyme, ATPase, or Ribosome, obtained significant correlation coefficients. Among all the groups of plastids, the *psbA* gene resulted to have the most optimized codon composition in term of correlation with the gene copy number of cognate tRNAs, which should increase its translation efficiecy. Conversely, genes with lower rates of

expression seem to be not optimized in this sense and rely more on the wobble and superwobble rules to be effectively translated.

References

Adams DG, Duggan PS, Jackson O. 2012. *Ecology of Cyanobacteria II, Their Diversity in Space and Time*. Springer Science & Business Media.

Alkatib S, Fleischmann TT, Scharff LB, Bock R. 2012a. Evolutionary constraints on the plastid tRNA set decoding methionine and isoleucine. *Nucleic Acids Res* 40: 6713–6724.

Alkatib S, Scharff L, Rogalski M. 2012b. The Contributions of Wobbling and Superwobbling to the Reading of the Genetic Code. *PLoS Genet* 8: e1003076.

Allen JF. 2003. Why chloroplasts and mitochondria contain genomes. *Comp Funct Genomics* 4: 31–36.

Anderson MJ. 2001. A new method for non-parametric multivariate analysis of variance. *Austral Ecol* 26: 32–46.

Anderson MJ. 2006. Distance-Based Tests for Homogeneity of Multivariate Dispersions. *Biometrics* 62: 245–253.

Anderson MJ, Crist TO, Chase JM, Vellend M, Inouye BD, Freestone AL, Sanders NJ, Cornell H V, Comita LS, Davies KF, et al. 2011. Navigating the multiple meanings of β diversity: a roadmap for the practicing ecologist. *Ecol Lett* 14: 19–28.

Archibald JM, Keeling PJ. 2002. Recycled plastids: a “green movement” in eukaryotic evolution. *Trends Genet* 18: 577–584.

Barbrook AC, Howe CJ, Purton S. 2006. Why are plastid genomes retained in non-photosynthetic organisms? *Trends Plant Sci* 11: 101–108.

Biller SJ, Berube PM, Lindell D, Chisholm SW. 2014. Prochlorococcus: the structure and function of collective diversity. *Nat Rev Microbiol* 13: 13–27.

Blanchard JL, Lynch M. 2000. Organellar genes: why do they end up in the nucleus? *Trends Genet* 16: 315–320.

Blankenship RE. 2010. Early Evolution of Photosynthesis. *PLANT Physiol* 154: 434–438.

Bocchi S, Malgioglio A. 2010. Azolla-Anabaena as a Biofertilizer for Rice Paddy Fields in the Po Valley, a Temperate Rice Area in Northern Italy. *Int J Agron* 2010: 1–5.

Bock R. 2007a. *Cell and molecular biology of plastids*. Springer Science &

Business Media.
Bock R. 2007b. Structure, function, and inheritance of plastid genomes. In *Cell and molecular biology of plastids*, pp. 29–63, Springer.
Boffey SA, Leech RM. 1982. Chloroplast DNA levels and the control of chloroplast division in light-grown wheat leaves. *Plant Physiol* 69: 1387–1391.
Brouard J-S, Otis C, Lemieux C, Turmel M. 2010. The exceptionally large chloroplast genome of the green alga Floydiella terrestris illuminates the evolutionary history of the Chlorophyceae. *Genome Biol Evol* 2: 240–56.
Bullerwell CE. 2011. *Organelle Genetics: Evolution of Organelle Genomes and Gene Expression*. Springer-Verlag.
Chan PP, Lowe TM. 2009. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res* 37: D93–7.
Chandler S, Dunwell JM. 2008. Gene flow, risk assessment and the environmental release of transgenic plants. *CRC Crit Rev Plant Sci* 27: 25–49.
Coghlan A, Wolfe KH. 2000. Relationship of codon bias to mRNA concentration and protein length in Saccharomyces cerevisiae. *Yeast* 16: 1131–1145.
Crick FH. 1966. Codon--anticodon pairing: the wobble hypothesis. *J Mol Biol* 19: 548–555.
Cui L, Veeraraghavan N, Richter A, Wall K, Jansen RK, Leebens-Mack J, Makalowska I, dePamphilis CW. 2006. ChloroplastDB: the Chloroplast Genome Database. *Nucleic Acids Res* 34: D692–D696.
De Marais DJ. 2000. Evolution. When did photosynthesis emerge on Earth? *Science* 289: 1703–1705.
Delannoy E, Le Ret M, Faivre-Nitschke E, Estavillo GM, Bergdoll M, Taylor NL, Pogson BJ, Small I, Imbault P, Gualberto JM. 2009a. Arabidopsis tRNA Adenosine Deaminase Arginine Edits the Wobble Nucleotide of Chloroplast tRNAArg(ACG) and Is Essential for Efficient Chloroplast Translation. *Plant Cell* 21: 2058–2071.
Delannoy E, Le Ret M, Faivre-Nitschke E, Estavillo GM, Bergdoll M, Taylor NL, Pogson BJ, Small I, Imbault P, Gualberto JMJM. 2009b. Arabidopsis tRNA Adenosine Deaminase Arginine Edits the Wobble Nucleotide of Chloroplast tRNAArg(ACG) and Is Essential for Efficient Chloroplast Translation. *Plant Cell* 21: 2058–2071.

dePamphilis CW, Palmer JD. 1990. Loss of photosynthetic and chlororespiratory genes from the plastid genome of a parasitic flowering plant. *Nature* 348: 337–9.

Dittmar KA, Mobley EM, Radek AJ, Pan T. 2004. Exploring the Regulation of tRNA Distribution on the Genomic Scale. *J Mol Biol* 337: 31–47.

Domanskii V, Rassadina V, Gus-Mayer S, Wanner G, Schoch S, Rüdiger W. 2003. Characterization of two phases of chlorophyll formation during greening of etiolated barley leaves. *Planta* 216: 475–483.

Drouin G, Daoud H, Xia J. 2008. Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants. *Mol Phylogenet Evol* 49: 827–831.

Dvořák P, Poulíčková A, Hašler P, Belli M, Casamatta DA, Papini A. 2015. Species concepts and speciation factors in cyanobacteria, with connection to the problems of diversity and classification. *Biodivers Conserv* 24: 739–757.

Evert RF, Eichhorn SE. 2013. Raven Biology of Plants. In *Biology of Plants*, pp. 1–880, W.H. Freeman & Company.

Field A, Miles J, Field Z. 2012. *Discovering Statistics Using R*. SAGE Publications.

Finster S, Legen J, Qu Y, Schmitz-Linneweber C. 2012. Land plant RNA editing or: don't be fooled by plant organellar DNA sequences. In *Genomics of Chloroplasts and Mitochondria*, pp. 293–321, Springer.

Flombaum P, Gallegos JL, Gordillo RA, Rincón J, Zabala LL, Jiao N, Karl DM, Li WKW, Lomas MW, Veneziano D, et al. 2013. Present and future global distributions of the marine Cyanobacteria Prochlorococcus and Synechococcus. *Proc Natl Acad Sci U S A* 110: 9824–9.

Gao L, SU Y, Wang T. 2010. Plastid genome sequencing, comparative genomics, and phylogenomics: current status and prospects. *J Syst Evol* 48: 77–93.

Gilson PR, Su V, Slamovits CH, Reith ME, Keeling PJ, McFadden GI. 2006. Complete nucleotide sequence of the chlorarachniophyte nucleomorph: nature's smallest nucleus. *Proc Natl Acad Sci* 103: 9566–9571.

Gingold H, Pilpel Y. 2011. Determinants of translation efficiency and accuracy. *Mol Syst Biol* 7.

Giordano M, Beardall J, Raven JA. 2005. CO 2 concentrating mechanisms in algae: mechanisms, environmental modulation, and evolution. *Annu Rev*

Plant Biol 56: 99–131.
Green BR. 2011. Chloroplast genomes of photosynthetic eukaryotes. *Plant J* 66: 34–44.
Guisinger MM, Kuehl J V, Boore JL, Jansen RK. 2011. Extreme reconfiguration of plastid genomes in the angiosperm family Geraniaceae: rearrangements, repeats, and codon usage. *Mol Biol Evol* 28: 583–600.
Hecht MK, MacIntyre RJ, Clegg MT. 2012. *Evolutionary Biology*. Springer US.
Hedges SB, Chen H, Kumar S, Wang DYC, Thompson AS, Watanabe H. 2001. A genomic timescale for the origin of eukaryotes. *BMC Evol Biol* 1: 4.
Hippler PDDM, Bock R. 2004. Extranuclear Inheritance: Chloroplast Proteomics. In *Progress in Botany*, pp. 90–105, Springer.
Hoffmann L, Komárek J, Kaštovský J. 2005. System of cyanoprokaryotes (cyanobacteria) – state in 2004. *Arch Hydrobiol Suppl Algol Stud* 117: 95–115.
Hoiczyk E, Hansel A. 2000. Cyanobacterial Cell Walls: News from an Unusual Prokaryotic Envelope. *J Bacteriol* 182: 1191–1199.
Holland HD. 2006. The oxygenation of the atmosphere and oceans. *Philos Trans R Soc B Biol Sci* 361: 903–915.
Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol* 2: 13–34.
IUPAC. 1992. Iupac Gold Book. *Gloss Chem terms used Biotechnol (IUPAC Recomm 1992* 148. http://goldbook.iupac.org/B00663.html.
Jarvis P, López-Juez E. 2013. Biogenesis and homeostasis of chloroplasts and other plastids. *Nat Rev Mol Cell Biol* 14: 787–802.
Kaplan-Levy RN, Hadas O, Summers ML, Rücker J, Sukenik A. 2010. *Dormancy and Resistance in Harsh Environments*. eds. E. Lubzens, J. Cerda, and M. Clark. Springer Berlin Heidelberg, Berlin, Heidelberg.
Karcher D, Bock R. 2009. Identification of the chloroplast adenosine-to-inosine tRNA editing enzyme. *RNA* 15: 1251–1257.
Kauff F, Büdel B. 2011. Phylogeny of cyanobacteria: an overview. In *Progress in Botany 72*, pp. 209–224, Springer.
Keeling PJ. 2004. Diversity and evolutionary history of plastids and their hosts. *Am J Bot* 91: 1481–1493.
Khakhlova O, Bock R. 2006. Elimination of deleterious mutations in plastid genomes by gene conversion. *Plant J* 46: 85–94.
Kim E, Archibald JM. 2009. Diversity and evolution of plastids and their

genomes. In *The Chloroplast*, pp. 1–39, Springer.

Knoll AH. 2003. The geological consequences of evolution. *Geobiology* 1: 3–14.

Komárek J., Kaštovský J. MJ& JJR. 2014. Taxonomic classification of cyanoprokaryotes (cyanobacterial genera) 2014, using a polyphasic approach. *Preslia* 86: 295–335.

Koning AP de, Keeling PJ. 2006. The complete plastid genome sequence of the parasitic green alga Helicosporidium sp. is highly reduced and structured. *BMC Biol* 4: 12.

Krause K. 2008. From chloroplasts to "cryptic" plastids: evolution of plastid genomes in parasitic plants. *Curr Genet* 54: 111–121.

Krupinska K, Melonek J, Krause K. 2013. New insights into plastid nucleoid structure and functionality. *Planta* 237: 653–664.

Kugrens P, Clay BL, Meyer CJ, Lee RE. 1999. Ultrastructure and description of Cyanophora biloba, sp. nov., with additional observations on C. paradoxa (Glaucophyta). *J Phycol* 35: 844–854.

Lagerkvist U. 1978. "Two out of three": an alternative method for codon reading. *Proc Natl Acad Sci U S A* 75: 1759–62.

Larkum AWD, Lockhart PJ, Howe CJ. 2007. Shopping for plastids. *Trends Plant Sci* 12: 189–195.

Lau AOT, McElwain TF, Brayton KA, Knowles DP, Roalson EH. 2009. Babesia bovis: a comprehensive phylogenetic analysis of plastid-encoded genes supports green algal origin of apicoplasts. *Exp Parasitol* 123: 236–43.

Lee RE. 2008. *Phycology*. Cambridge University Press.

Lewin RA. 2002. Prochlorophyta - a matter of class distinctions. *Photosynth Res* 73: 59–61.

Lim L, McFadden GI. 2010. The evolution, metabolism and functions of the apicoplast. *Philos Trans R Soc B Biol Sci* 365: 749–763.

Liu Q, Xue Q. 2005. Comparative studies on codon usage pattern of chloroplasts and their host nuclear genes in four plant species. *J Genet* 84: 55–62.

Lowe TM, Eddy SR. 1997. tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. *Nucleic Acids Res* 25: 0955–964.

Lung B, Zemann A, Madej MJ, Schuelke M, Techritz S, Ruf S, Bock R,

Hüttenhofer A. 2006. Identification of small non-coding RNAs from mitochondria and chloroplasts. *Nucleic Acids Res* 34: 3842–3852.

Maliga P. 2014. Chloroplast Biotechnology. *Methods Mol Biol* 1132.

Maliga P, Bock R. 2011. Plastid biotechnology: food, fuel, and medicine for the 21st century. *Plant Physiol* 155: 1501–10.

Margulis L. 1970. *Origin of eukaryotic cells: evidence and research implications for a theory of the origin and evolution of microbial, plant, and animal cells on the Precambrian earth*. Yale University Press New Haven.

Marinos NG. 1967. Multifunctional plastids in the meristematic region of potato tuber buds. *J Ultrastruct Res* 17: 91–113.

Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, Lins T, Leister D, Stoebe B, Hasegawa M, Penny D. 2002. Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc Natl Acad Sci* 99: 12246–12251.

McFadden G. 2001. Primary and secondary endosymbiosis and the origin of plastids. *J Phycol* 37: 951–959.

McFadden GI, van Dooren GG. 2004. Evolution: red algal genome affirms a common origin of all plastids. *Curr Biol* 14: R514–R516.

Mereschkowsky C. 1905. Über Natur und Ursprung der Chromatophoren im Pflanzenreiche. *Biol Zentrablatt* 25: 593–604.

Mereschowsky K. 1910. Theorie der zwei Plasmaarten als Grundlage der Symbiogenesis, einer neuen Lehre von der Ent-stehung der Organismen. *Biol Cent* 30: 353–367.

Moore PB, Steitz TA. 2011. The roles of RNA in the synthesis of protein. *Cold Spring Harb Perspect Biol* 3: a003780.

Morden CW, Sherwood AR. 2002. Continued evolutionary surprises among dinoflagellates. *Proc Natl Acad Sci* 99: 11558–11560.

Muller HJ. 1964. The relation of recombination to mutational advance. *Mutat Res Mol Mech Mutagen* 1: 2–9.

Mullineaux CW. 2005. Function and evolution of grana. *Trends Plant Sci* 10: 521–5.

Nabout JC, da Silva Rocha B, Carneiro FM, Sant'Anna CL. 2013. How many species of Cyanobacteria are there? Using a discovery curve to predict the species number. *Biodivers Conserv* 22: 2907–2918.

Nakamura M, Sugiura M. 2007. Translation efficiencies of synonymous codons are not always correlated with codon usage in tobacco chloroplasts. *Plant J* 49: 128–34.

Nakamura M, Sugiura M. 2011. Translation efficiencies of synonymous codons for arginine differ dramatically and are not correlated with codon usage in chloroplasts. *Gene* 472: 50–4.

NCBI GenBank. 2015. NCBI GenBank website. https://www.ncbi.nlm.nih.gov/nuccore/.

NCBI Genome. 2015. NCBI Organelle Genome Resources website. https://www.ncbi.nlm.nih.gov/genome/organelle/.

Novoa E, Pavon-Eternod M. 2012. A role for tRNA modifications in genome structure and codon usage. *Cell* 149: 202–213.

Novoa EM, Ribas de Pouplana L. 2012. Speeding with control: Codon usage, tRNAs, and ribosomes. *Trends Genet* 28: 574–581.

Nowack ECM, Melkonian M, Glöckner G. 2008. Chromatophore Genome Sequence of Paulinella Sheds Light on Acquisition of Photosynthesis by Eukaryotes. *Curr Biol* 18: 410–418.

Nürnberg DJ, Mariscal V, Parker J, Mastroianni G, Flores E, Mullineaux CW. 2014. Branching and intercellular communication in the Section V cyanobacterium Mastigocladus laminosus, a complex multicellular prokaryote. *Mol Microbiol* 91: 935–49.

Ohyama K, Fukuzawa H, Kohchi T, Shirai H, Sano T, Sano S, Umesono K, Shiki Y, Takeuchi M, Chang Z, et al. 1986. Chloroplast gene organization deduced from complete sequence of liverwort Marchantia polymorpha chloroplast DNA. *Nature* 322: 572–574.

Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Wagner H. 2013. vegan: Community Ecology Package.

Ostell J, McEntyre J. 2013. The NCBI Handbook. *NCBI Bookshelf* 1–8.

Paerl HW, Huisman J. 2009. Climate change: a catalyst for global expansion of harmful cyanobacterial blooms. *Environ Microbiol Rep* 1: 27–37.

Palmer JD. 1985. Comparative organization of chloroplast genomes. *Annu Rev Genet* 19: 325–354.

Palmer JD, Thompson WF. 1982. Chloroplast DNA rearrangements are more frequent when a large inverted repeat sequence is lost. *Cell* 29: 537–550.

Patron NJ, Inagaki Y, Keeling PJ. 2007. Multiple gene phylogenies support the

monophyly of cryptomonad and haptophyte host lineages. *Curr Biol* 17: 887–891.

Pfanzagl B, Zenker A, Pittenauer E, Allmaier G, Martinez-Torrecuadrada J, Schmid ER, De Pedro MA, Löffelhardt W. 1996. Primary structure of cyanelle peptidoglycan of Cyanophora paradoxa: a prokaryotic cell wall as part of an organelle envelope. *J Bacteriol* 178: 332–9.

Plotkin JB, Kudla G. 2011. Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet* 12: 32–42.

Prechtl J, Kneip C, Lockhart P, Wenderoth K, Maier U-G. 2004. Intracellular spheroid bodies of Rhopalodia gibba have nitrogen-fixing apparatus of cyanobacterial origin. *Mol Biol Evol* 21: 1477–1481.

Price DC, Chan CX, Yoon HS, Yang EC, Qiu H, Weber APM, Schwacke R, Gross J, Blouin NA, Lane C, et al. 2012. Cyanophora paradoxa Genome Elucidates Origin of Photosynthesis in Algae and Plants. *Science (80-)* 335: 843–847.

Pyke K. 2007. Plastid biogenesis and differentiation. In *Cell and molecular biology of plastids*, pp. 1–28, Springer.

R Core Team. 2014. R: A Language and Environment for Statistical Computing.

R Development Core Team. 2013. R Software. *R A Lang Environ Stat Comput*.

Ran W, Higgs PG. 2010. The Influence of Anticodon–Codon Interactions and Modified Bases on Codon Usage Bias in Bacteria. *Mol Biol Evol* 27: 2129–2140.

Raubeson LA, Peery R, Chumley TW, Dziubek C, Fourcade HM, Boore JL, Jansen RK. 2007. Comparative chloroplast genomics: analyses including new sequences from the angiosperms Nuphar advena and Ranunculus macranthus. *BMC Genomics* 8: 174.

Reis M dos, Savva R, Wernisch L. 2004. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res* 32: 5036–5044.

Ripley B, Venables B, Bates DM, 1998) KH (partial port ca, 1998) AG (partial port ca, Firth D. 2014. MASS: Support Functions and Datasets for Venables and Ripley's MASS. *R Packag version*.

Rippka R, Deruelles J, Waterbury JB, Herdman M, Stanier RY. 1979. Generic Assignments, Strain Histories and Properties of Pure Cultures of Cyanobacteria. *J Gen Microbiol* 111: 1–61.

Rogalski M, Karcher D, Bock R. 2008. Superwobbling facilitates translation with reduced tRNA sets. *Nat Struct Mol Biol* 15: 192–198.

Sagan L. 1967. On the origin of mitosing cells. *J Theor Biol* 14: 255–274.

Sager R, Ishida MR. 1963. Chloroplast DNA in Chlamydomonas. *Proc Natl Acad Sci U S A* 50: 725.

Salinas T, Duchêne A-M, Maréchal-Drouard L. 2008. Recent advances in tRNA mitochondrial import. *Trends Biochem Sci* 33: 320–9.

Sarma TA. 2012. *Handbook of Cyanobacteria*. CRC Press.

Schirrmeister BE, Antonelli A, Bagheri HC. 2011. The origin of multicellularity in cyanobacteria. *BMC Evol Biol* 11: 45.

Schirrmeister BE, Gugger M, Donoghue PCJ. 2015. Cyanobacteria and the Great Oxidation Event: evidence from genes and fossils ed. A. Smith. *Palaeontology* 58: n/a–n/a.

Schneider A. 2011. Mitochondrial tRNA import and its consequences for mitochondrial translation. *Annu Rev Biochem* 80: 1033–1053.

Schopf JW. 2002. The Fossil Record: Tracing the Roots of the Cyanobacterial Lineage. In *The Ecology of Cyanobacteria* (eds. B.A. Whitton and M. Potts), Kluwer Academic Publishers, Dordrecht.

Schwab BDM& E. 2012. Symbiosis: “Living together” in chaos. *Stud Hist Biol* 4: 7–25.

Schwender J, Goffman F, Ohlrogge JB, Shachar-Hill Y. 2004. Rubisco without the Calvin cycle improves the carbon efficiency of developing green seeds. *Nature* 432: 779–782.

Sharkey TD, Bock R, Planck M, Plant M. 2012. *Photosyntesis: Plastid Biology, Energy Conversion and Carbon Assimilation*. Springer Science & Business Media.

Sharp PM, Tuohy TMF, Mosurski KR. 1986. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res* 14: 5125–5143.

Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayashida N, Matsubayashi T, Zaita N, Chunwongse J, Obokata J, Yamaguchi-Shinozaki K, et al. 1986. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. *EMBO J* 5: 2043–2049.

Slepecky R a., Hemphill HE. 2006. *The Prokaryotes: Vol. 4: Bacteria: Firmicutes, Cyanobacteria*. Springer Science & Business Media.

Smith DR. 2009. Unparalleled GC content in the plastid DNA of Selaginella.

Plant Mol Biol 71: 627–39.
Spearman C. 1904. The proof and measurement of association between two things. *Am J Psychol* 15: 72–101.
Stern DB, Goldschmidt-Clermont M, Hanson MR. 2010. Chloroplast RNA metabolism. *Annu Rev Plant Biol* 61: 125–155.
Stirewalt VL, Michalowski CB, Löffelhardt W, Bohnert HJ, Bryant DA. 1995. Nucleotide sequence of the cyanelle genome fromCyanophora paradoxa. *Plant Mol Biol Report* 13: 327–332.
Sturrock K, Rocha J. 2000. A Multidimensional Scaling Stress Evaluation Table. *Field methods* 12: 49–60.
Sugiura M. 2008. RNA editing in chloroplasts. In *RNA editing*, pp. 123–142, Springer.
Sugiura M. 2013. Translation in Chloroplasts of Flowering Plants. In *Translation in Mitochondria and Other Organelles*, pp. 207–224, Springer.
Tiller N, Bock R. 2014. The Translational Apparatus of Plastids and Its Role in Plant Development. *Mol Plant* 7: 1105–1120.
Tillich M, Lehwark P, Morton BR, Maier UG. 2006. The evolution of chloroplast RNA editing. *Mol Biol Evol* 23: 1912–1921.
Timmis JN, Ayliffe MA, Huang CY, Martin W. 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat Rev Genet* 5: 123–135.
Tomitani A, Okada K, Miyashita H, Matthijs HCP, Ohno T, Tanaka A. 1999. Chlorophyll b and phycobilins in the common ancestor of cyanobacteria and chloroplasts. 400: 159–162.
Toyoshima Y, Onda Y, Shiina T, Nakahira Y. 2005. Plastid transcription in higher plants. *CRC Crit Rev Plant Sci* 24: 59–81.
Tsuji S, Ueda K, Nishiyama T, Hasebe M, Yoshikawa S, Konagaya A, Nishiuchi T, Yamaguchi K. 2007. The chloroplast genome from a lycophyte (microphyllophyte), Selaginella uncinata, has a unique inversion, transpositions and many gene losses. *J Plant Res* 120: 281–90.
Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, Pan T, Dahan O, Furman I, Pilpel Y. 2010a. An Evolutionarily Conserved Mechanism for Controlling the Efficiency of Protein Translation. *Cell* 141: 344–354.
Tuller T, Waldman YY, Kupiec M, Ruppin E. 2010b. Translation efficiency is determined by both codon bias and folding energy. *Proc Natl Acad Sci*

107: 3645–3650.

Wakasugi T, Tsudzuki T, Sugiura M. 2001. The genomics of land plant chloroplasts: gene content and alteration of genomic information by RNA editing. *Photosynth Res* 70: 107–118.

Warnes GR, Bolker B, Bonebakker L, Gentleman R, Huber W, Liaw A, Lumley T, Maechler M, Magnusson A, Moeller S, et al. 2014. gplots: Various R programming tools for plotting data. 2013. *R Packag version* 2.

Whitton B a, Potts M. 2000. *The Ecology of Cyanobacteria: Their Diversity in Time and Space*. Springer Science & Business Media.

Wilkinson L, Friendly M. 2009. The history of the cluster heat map. *Am Stat* 63.

Wise RR, Hoober JK. 2007. *The structure and function of plastids*. Springer Science & Business Media.

Wolf PG, Rowe CA, Sinclair RB, Hasebe M. 2003. Complete nucleotide sequence of the chloroplast genome from a leptosporangiate fern, Adiantum capillus-veneris L. *DNA Res* 10: 59–65.

Xu C, Cai X, Chen Q, Zhou H, Cai Y, Ben A. 2011. Factors Affecting Synonymous Codon Usage Bias in Chloroplast Genome of Oncidium Gower Ramsey. *Evol Bioinform Online* 7: 271–278.

Yoon HS, Hackett JD, Ciniglia C, Pinto G, Bhattacharya D. 2004. A molecular timeline for the origin of photosynthetic eukaryotes. *Mol Biol Evol* 21: 809–18.

Yoon HS, Reyes-Prieto A, Melkonian M, Bhattacharya D. 2006. Minimal plastid genome evolution in the Paulinella endosymbiont. *Curr Biol* 16: R670–R672.

Zeileis A, Wiel MA, Hornik K, Hothorn T. 2008. Implementing a class of permutation tests: The coin package. *J Stat Softw* 28: 1–23.

Zhang Z, Green BR, Cavalier-Smith T. 1999. Single gene circles in dinoflagellate chloroplast genomes. *Nature* 400: 155–159.

Zhou W, Karcher D, Bock R. 2013. Importance of adenosine-to-inosine editing adjacent to the anticodon in an Arabidopsis alanine tRNA under environmental stress. *Nucleic Acids Res* 41: 3362–72.

Appendix

Brief description of the algae mentioned in this thesis (Box 2 of Archibald and Keeling 2002).

Green algae (and plants) (e.g.*Chlamydomonas*)	Plants evolved from green algae and are very similar to them in many respects. Both are extremely abundant, morphologically diverse, successful lineages.
Red algae (e.g.*Porphyra*)	Very abundant, diverse group ranging from microscopic unicellular forms to large multicellular seaweeds. Some of the large seaweeds are used to produce carbohydrates such as carrageenan or to make the nori used to wrap sushi.
Glaucocystophytes (e.g.*Cyanophora*)	A little-studied group of algae with a primary plastid. Most remarkable because their plastid is the only one that has retained the peptidoglycan wall between its two membranes.
Chlorarachniophytes (e.g.*Chlorarachnion*)	Relatively rare marine amoeboflagellate algae with green secondary plastids. Best known because the secondary endosymbiont has retained its nucleus (called a nucleomorph) and a miniature genome.
Euglenids (e.g.*Euglena*)	Common algae in marine and freshwater environments with a green secondary plastid. Known for their peculiar movement and as close relatives of the parasitic trypanosomes.
Cryptomonads (e.g.*Guillardia*)	Common algae with a red secondary plastid. Best known because, along with chlorarachniophytes, they have retained a nucleomorph. The complete sequence of a cryptomonad nucleomorph genome is now known and is a model of reduction and compaction.
Haptophytes (e.g.*Emiliania*)	Common, ecologically important algae with a red secondary plastid. Many haptophytes are covered in elaborate calcareous scales called coccoliths, which are a primary component of chalk sediments such as the white cliffs of Dover.
Heterokonts (e.g.*Laminaria*,*Phytophthora*)	A very diverse group that includes many photosynthetic forms (e.g. kelps and diatoms) and non-photosynthetic forms (e.g. oomycetes

	such as the potato-late-blight agent). Photosynthetic types have a red secondary plastid, and evidence now suggests that the entire group is derived from a photosynthetic ancestor
Dinoflagellates (e.g.*Amphidinium*)	Very common group with a red secondary plastid. Best known for causing 'red tides' and toxic shellfish poisoning, but also very important ecologically.
Apicomplexa (e.g.*Plasmodium,Cryptosporidium*)	A very diverse group, all of which are obligate intracellular parasites. They cause many medically and commercially significant diseases, notably malaria. Recently found to contain a plastid, now known to be a red secondary plastid.

Summary of the supergroups of eukaryotic organisms used in this thesis.

Supergroups	**Phyla**		**Plastid Origin**
ARCHAEPLASTIDA	Streptophyta Chlorophyta Rhodophyta Glaucophyta (or Glaucocystophyta)		Primary Endosymbiosis
EXCAVATA	Euglenozoa		Secondary Endosymbiosis GREEN LINEAGE
RHIZARIA	Cercozoa/Chloroarachniophyta		Secondary Endosymbiosis GREEN LINEAGE
CHROMALVEOLATA	Chromista	Cryptophyta Heterokonta (or Stramenopiles) Haptophyta	Secondary Endosymbiosis RED LINEAGE
	Alveolata	Dinoflagellata Apicomplexa	

Publications

REVIEW ARTICLE
Biodivers Conserv (2015) 24:739–757

I SPECIES CONCEPTS AND SPECIATION FACTORS IN CYANOBACTERIA, WITH CONNECTION TO THE PROBLEMS OF DIVERSITY AND CLASSIFICATION.

Petr Dvořák, Aloisie Poulíčková, Petr Hašler, Mattia Belli, Dale A. Casamatta, Alessio Papini.

Abstract

The cyanobacteria are the most important prokaryotic primary producers on Earth, inhabiting a great diversity of aquatic and terrestrial environments exposed to light. However, the evolutionary forces leading to their divergence and speciation remain largely enigmatic compared to macroorganisms due to their prokaryotic nature, including vast population sizes, and largely asexual reproduction. The advent of modern molecular techniques has facilitated an understanding of the important factors shaping cyanobacterial evolution, including horizontal gene transfer and homologous recombination. We review the forces shaping the evolution of cyanobacteria and discuss the role of cohesive forces on speciation. Further, while myriad species concepts and definitions are currently used, only a limited subset might be applied to cyanobacteria due to their asexual reproduction. Additionally, concepts based solely on phenotypes provide insufficient resolution. A monophyletic species concept which is universal may be ideal for cyanobacteria. Actual identification of the cyanobacteria is difficult due to cryptic diversity, lack of morpho- logical variability, and frequent convergent evolutionary events. Thus, applied molecular techniques such as DNA barcoding will be useful for identifications of environmental samples. Lastly, we show that the real biodiversity of the cyanobacteria is widely under-estimated, due in part to low sampling efforts, sensitivity to the molecular markers employed, and the species definitions employed by researchers. In conclusion, we an- ticipate a rapid increase in cyanobacterial taxa described and large revisions of the system in the future as scientists adopt a common approach to cyanobacterial systematics.

Introduction

The Cyanobacteria (also known as the Cyanophyceae, Cyanophyta, cyanoprokaryota, blue- green algae or blue-green bacteria) are prokaryotes possessing oxygenic photosynthesis, while sharing similar habitats to eukaryotic algae (Kauff and Bu¨del 2011). Moreover, cyanobacteria can live in some of the most extreme habitats on earth (Seckbach 2007). On the basis of fossil records, Schopf (2000) estimated that cyanobacteria may have evolved

3.5 BYA, making them the oldest oxygen producing photosynthetic microbes, and sig- nificant contributors to the sudden increase in atmospheric oxygen during the Great Oxidation Event (Bekker et al. 2004; Kauff and Bu¨del 2011).

The cyanobacteria exhibit remarkable variability in morphology and ultrastructure, from unicellular to filamentous forms (Figs. 1, 2). They may also possess intercellular connections or microplasmodesmata, considered a sign of multicellularity (Nu¨rnberg et al. 2014). Moreover, some genera exhibit morphological and functional cell differentiation such as heterocytes (adapted to nitrogen fixation) and akinetes (resting stage cells) (Whitton and Potts 2000).

The purpose of this paper is three-fold. First, we will review the most recent literature relating to the evolutionary processes forming bacterial (and cyanobacterial in particular) species. Second, we will evaluate their application in cyanobacterial taxonomy, distribu- tion, species concepts and species definitions. Third, we will discuss some practical aspects of cyanobacterial taxonomy and systematics.

Species concept in (cyano)bacteria

Some authors postulate that all prokaryotes are species-less or fuzzy (e.g. Hanage et al. 2005; Konstantinidis and Tiedje 2005; Hanage 2013), because they lack ecologically or genetically coherent groups. Such ‘‘fuzziness’’ might be apparent in ambiguous ecological boundaries among species, which was suggested by Cohan and Perry (2007) and Kopac et al. (2014). For cyanobacteria, the most obvious phenotypic features (i.e. cell morphology) may at times be phylogenetically uninformative when compared to phylogenies generated by 16S rRNA gene data, which is the currently accepted ‘‘Gold-standard’’ in bacterial systematics (e.g. Honda et al. 1999; Robertson et al. 2001; Kim et al. 2014).

Moreover, species identifications might be complicated by the analyses used, with some traditional methods lacking species-level resolution. For example, Hanage et al. (2005) showed that multilocus sequence analysis is required to distinguish highly recombinant species of the human inhabiting bacterium Neisseria spp. Similarly, marine picoplanktic Synechococcus is

composed of several ecological and geographical lineages, which may be recognized only based upon multilocus sequence analysis (Mazard et al. 2012).

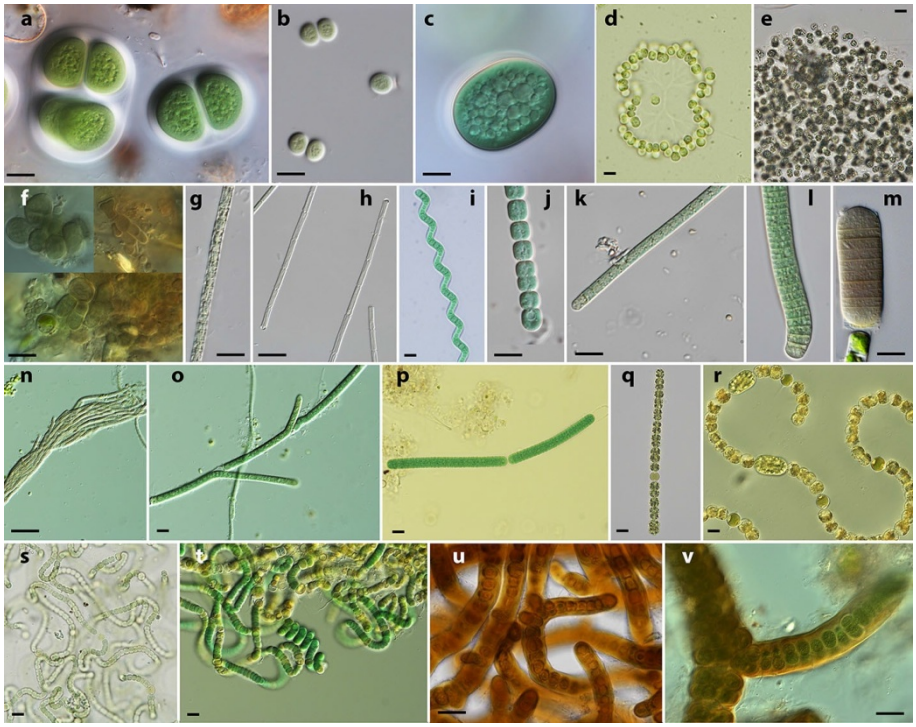


Fig. 1 Illustration of morphological diversity in cyanobacteria. Groups (orders) follow Rippka et al. (1979).
I. Chroococcales: a Chroococcus subnudus, b Ch. limneticus, c Cyanothece aeruginosa, d Snowella litoralis, e Microcystis aeruginosa. II. Pleurocapsales: f Pleurocapsa minor. III. Oscillatoriales: g Planktothrix agardhii, h Limnothrix redekei, i Arthrospira jenneri, j Johanseninema constricum, k Phormidium sp., l, m Oscillatoria sp., n Schizothrix sp., o Tolypothrix sp., p Katagnymene accurata., IV. Nostocales: q Dolichospermum planctonicum, r Dolichospermum sp., s Nostoc sp., t Nodularia moravica.
V. Stigonematales: u, v Stigonema sp. Scale bar a–u = 10 lm, v = 20 lm. (Color figure online)

On the other hand, since the frequency of horizontal gene transfer (HGT) and ho- mologous recombination (HR) decreases with the genetic distance suggests coherence within evolutionary lineages and thus the existence

of prokaryotic species. However, this assertion might be complicated by the methods used or by the stage of speciation. Models of speciation presented by Polz et al. (2013) and Dvorˇa´k et al. (2014b) showed mixed phylogenetic signals based on different loci at the beginning of speciation caused by HGT and HR. A stronger phylogenetic signal comes later during speciation and is balanced until coherent species units are evident (Shapiro et al. 2012). Conversely, Cohan (2011) argues that cohesion is not maintained by barriers of recombination, but rather concerned with ecological diversification, which precedes barriers of recombination (Wiedenbeck and Cohan 2011).

Cohesion might not necessarily be the key factor for the existence of species de- limitations, and Kopac et al. (2014) proposed ecological differences among ecotypes as key features. They suggest that ecotypes exist indefinitely, but lineages within ecotypes are changing. In conclusion, regardless mechanisms of coherence, there seem to be coherent evolutionary lineages in cyanobacteria, which might be called species.

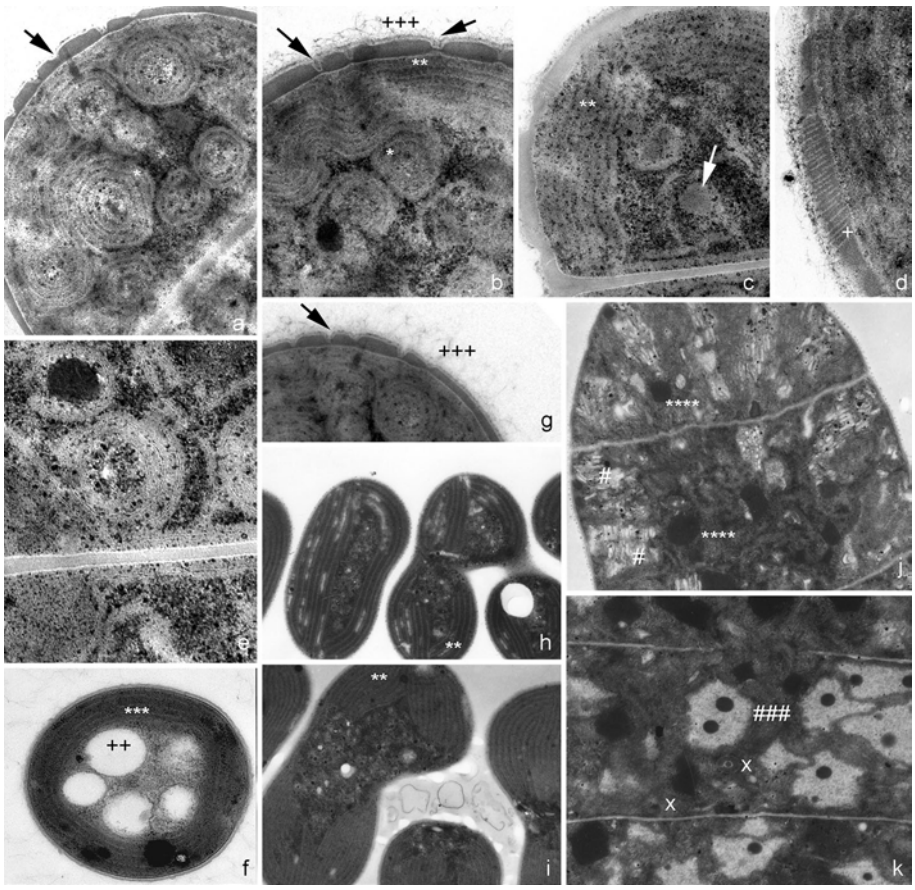


Fig. 2 Transmission electron images of cyanobacteria—illustration of diversity in ultrastructure; a–e, g Hormoscilla pringsheimii; f Neosynechococcus sphagnicola; h, i Spirulina; j, k Arthrospira; a The cell wall depressions (wall pores, black arrows) are the passages through which mucilage crosses the wall. A large part of the cytoplasm is occupied by thylakoids often coiled to form roundish structures formed by circular thylakoids (asterisk) b detail of a. The black arrows indicate the cell wall depressions (mucilage pores). Abundant sheath fibrillar mucilage **(???)** is evident along the wall. Some wavy thylakoids (**) run along the cytoplasmic membrane as single lamellae, while more internally they coil to form roundish bodies (single asterisk). Roundish electron dense bodies (****) represent polyphosphate granules. c Grey spherical bodies (white arrow) in the cytoplasm represent cyanophycin bodies. d Junction pores (?) through the cell wall appear as channels orthogonal to the cytoplasmic membrane surface. e Detail of a polyphosphate granule (****). Many ribosomes can be observed in the cytoplasm,

particularly close to the thylakoids. f In unicellular species, thylakoids are typically arranged parietally (***), along the cytoplasmic membrane. Polyhydroxybutyrate bodies (PHB, ??) are visible. g Detail of the cell wall depressions (black arrow) and their relationships with the fibrillar component of mucilage. h Image of the spirally arranged filament of Spirulina, whence the frequent observation of double flanked cells. The wavy thylakoids are clustered in bundles of lamellae. i Detail of h. A large part of the cytoplasm is occupied by wavy thylakoids, while the "free" cytoplasm appears lectron dense and containing many different bodies at very variable level of electron density. j In a filament of Arthrospira the apical cell appears to have a different shape with respect to the other cells. Many heterogeneous cytoplasmic structures are visible, among which polyphosphate bodies (****) and gas vesicles (aerotopes, #). k Detail of j. Apparently even spaces possibly enclosed by membranes ### and containing electron dense bodies and fibrillar material can be observed, such bodies are interpreted as assembling carboxysomes. Cylindrical bodies (x) can be observed in the cytoplasm. Material can pass through cell wall pores (microplasmodesmata) from one cell to another in filamentous genera

Before we begin a discussion of species concepts in bacteria, we would like to em- phasize the differences between a species concept and species definition, which is often confused. A species concept is a theoretical demarcation of the species, which would be ideally applicable to all organisms. Conversely, species definitions are a set of rules used for practical identification of species (Hanage 2013). For instance, most bacteriologist use distance among genes or genomes (DNA–DNA hybridization, average nucleotide identity; Richter and Rossello´-Mo´ra 2009) as a species definition while not considering an actual species concept. It does not take into account a phase of speciation or phylogenetic po- sition, and thus it does not show the true evolutionary history of the species.

A possible concept of a bacterial species may be a ''genomically and phenomically cohesive cluster'' to which a possible concept of species may be applied (Doolittle and Zhaxybayeva 2009). However, the same authors pointed out that there would be ''no principled way in which questions about prokaryotic species, such as how many there are, how large their populations are, or how globally they are distributed, can be answered''. Thus, the question remains: how to evaluate biodiversity among prokaryotes?

Is there a quantitative threshold of genetic difference sufficient to describe a prokaryotic species in order that eukaryotes-centered biological species concept might work (sensu Mayr 1942, 1946)?

The recognition of the prokaryotic species problem eventually led to a partial consensus about species delimitations (Gevers et al. 2005, 2006; Staley

2006). According to these authors, a prokaryotic species should be recognized primarily on the basis of genotypic similarity and hence mainly on genetic distances. Stakebrandt et al. (2002) proposed that two isolates may be assigned to the same species in case of a value higher than 70 % in a standardized DNA–DNA hybridization experiment. Other distances based on the small subunit (SSU, or 16S) rRNA, could be used to exclude the belonging to the same species in case of a [97.5 % similarity (Fox et al. 1992; Stackebrandt and Goebel 1994). Another threshold range 98.7–99 % has been proposed by Strackerbrandt and Ebers (2006). Goris et al. (2007), and Richter and Rosello´-Mo´ra (2009) proposed 95–96 % average nucleotide identity (ANI) of homologous genomic regions as a gold standard for species delimitation and also as an alternative to DNA–DNA hybridization. Most recently, Kim et al. (2014) combined previously mentioned approaches and proposed 98.65 % similarity in 16S rRNA as a threshold for species delimitation. Unfortunately, these are all similarity based criteria, and not in line with modern systematics approaches which emphasize broader tools of reconstruction of evolutionary relationships (Castenholz and Norris 2005; Johansen and Casamatta 2005; Koma´rek 2010 and many others).

The main problem with bacterial species concepts is that they do not fit well into the requirements of the classical species concept used for eukaryotes. Staley (2006) proposed the genomic-phylogenetic species concept, while Achtman and Wagner (2008) adapted the de Queiroz (2005, 2007) general lineage concept to a prokaryote-limited metapopulation lineages concept, requiring only that ‘‘members’’ of a species (lineage) evolve separately from other lineages. Such separation would provide the cohesive force that eventually forms a species. However, they observed that such a concept does not provide sufficient detection and quantification of cohesive forces.

One of the main issues in bacterial systematics is whether or not lineages necessarily represent a genetic continuum (Konstantinidis et al. 2006). For example, a simple com- putational model of randomly replicating lineages will produce groups of genetically related individuals separated by genetic gaps (Zhaxybayeva and Gogarten 2004; Mes 2008; Doolittle and Zhaxybayeva 2009). Hence a ‘‘good’’ species should have deeper gaps with respect to what happens with a random model. The possibility of a failed recognition of intermediate forms may also arise due to sampling or difficult cultivability of many bac- terial strains, because most bacterial species are unculturable (reviewed in Stewart 2012).

Recent papers have employed multi-locus DNA sequences analyses (MLSA) for species definition, which often yield results that fit with traditionally delimited species (Gevers et al. 2005; Hanage et al. 2005, 2006). MLSA has been developed originally for identi- fication of pathogenic strains of bacteria due to lack of resolution of traditional genetic markers, mainly the 16S rRNA gene (Maiden et al. 1998). For example, Melendrez et al. (2011)

used three protein coding genes and found 4–14 times more ecotypes in the ther- mophilic Synechococcus sp. inhabiting Mushroom Spring in Yellowstone National Park than based solely on 16S rRNA and 16S-23S ITS sequence. Thus MLSA provides sig- nificantly higher resolution. Similar considerations may be obtained from the investiga- tions on the marine planktic genera Synechococcus and Prochlorococcus (Johnson et al. 2006) or fine-scale distribution of marine Vibrionaceae (Preheim et al. 2011). On the other hand, Kopac et al. (2014) analyzed all orthologous genes within Bacillus subtilis and showed that MLSA was insufficient to distinguish ecotypes, which are considered as species in this paper.

Recently, DNA barcoding has been proposed as a possibility for cyanobacteria. Eckert et al. (2015) tested barcoding gaps in cyanobacteria and found that barcoding gaps among species were identified in a half of investigated cases. Thus, this approach has to be further investigated before it takes place in practical identification of species.

Speciation factors in (cyano)bacteria

Bacteria and Archaea are evolutionarily intriguing as they are asexual, and possess ex- tensive populations with relatively short generation times (for review see Cohan 2001, 2002). Cohesive or disruptive forces shaping bacterial species have remained enigmatic for a long time. However, the development of modern molecular methods has shown sig- nificant differences between prokaryotic and eukaryotic evolutionary trajectories. For example, some of the non-consistent phylogenetic signals of different gene families within the same bacterial species have been explained by HGT and HR (e.g. Hanage et al. 2005; Lodders et al. 2005; David and Alm 2011). It has been suggested that a part of the bacterial genome usually referred to as the core genome is more stable with less evolutionary changes. The core genome is usually defined as a portion of genes shared by some group of bacteria coding for essential metabolic pathways (Daubin et al. 2002; Shi and Falkowski 2008; Polz et al. 2013). The shell or flexible genome refers to a less stable part of bacterial genome which undergoes substantial evolutionary changes including HGT (Hess 2011). It often contains genes specific to some environment with a large portion of unannotated gene families without any known function (e.g. Shi and Falkowski 2008), which putatively plays an important role in rapidly changing environments (Rodriguez-Valera et al. 2009) and niche partitioning among close relatives (Kopac et al. 2014). The shell genome genes do not seem to be randomly dispersed over the chromosome, but rather concentrated within genomic islands with frequent HGT and HR events (Hacker and Carniel 2001; Rodriguez-Valera et al. 2009). However, Narechania et al. (2012) showed that many core genes have an identical phylogenetic signal as shell genes, which denotes their common evolutionary history. Narechania et al. (2012) defined core genome as

orthologs with the same phylo- genetic tree topology and the shell genome as composed of the rest of orthologs. Core genes may also exhibit evidence of HGT events, which might be identified by comparing scenarios of gene phylogenies with individual species trees (David and Alm 2011; Nakhleh 2013). The question remains, though, do these changes provide enough force to diverge evolutionary lineages with subsequent cohesion to form an analogue of the eukaryotic ‘‘sexual’’ species?

The most extensively studied HGT events are concerned with the human microbiome

(e.g. Smillie et al. 2011) and marine picoplankton, mostly of the genera Synechococcus and Prochlorococcus (e.g. Marston et al. 2012). In terms of cyanobacteria, it has been further suggested that most HGT are mediated by phages (cyanophages) (e.g. Sullivan et al. 2010; Sabehi et al. 2012). These phages often contain genes important in photosynthesis (Zheng et al. 2014). While HGT events may occur between phylogenetically divergent lineages, they are most frequent among individual species within the same environment and de- crease with the overall genetic distance of genomes (Popa et al. 2011). A very similar phenomenon has been observed in HR (Smillie et al. 2011). Fraser et al. (2007) modeled HR within bacteria, showing that if HR exceeds mutation rate, a species evolves in a similar manner as sexually reproducing eukaryotes, and with low HR the populations are clonal. Polz et al. (2013) suggested in their synthesis that rather than genetic isolation of emerging lineages, there exist local genetic innovative gene pools (i.e. local metagenomes), which are constantly changing by HGT within a pool and by input of incoming genotypes.

Besides genetic isolation resulting from genome differentiation, which takes place in population without geographical isolation (sympatric speciation) often observed in bacteria

(e.g. Friedman et al. 2013; Koeppel et al. 2013), there are geographical and ecological

factors affecting bacterial speciation. The speciation of macroorganisms is often driven by geographical isolation (allopatry) due to their limited dispersal capabilities. This has also been in, e.g., asexual rotifers, but on a larger geographical scale (Fontaneto et al. 2008). However, how these processes relate to microorganisms is still subject to broad debate (see Martiny et al. 2006; Ramette and Tiedje 2007 for review). Baas Becking (1934) postulated that all microbes can spread everywhere and only the specific local environmental conditions would select actual species composition. However, recent analyses of different molecular markers reveal an ambiguous signal. For example, thermophilic cyanobacteria Mastigocladus laminosus and Synechococcus spp. showed geographical difference based on 16S rRNA analysis (Papke et al. 2003; Miller et al. 2007). However, it should be noted that in the case of Synechococcus (Papke et al. 2003), the clusters were genetically very distant,

which might be because they belong to different taxa (even genera) and there is not sufficient variation within a species to elucidate meaningful patterns. Dvorˇa´k et al. (2012) showed that episodic genetic isolation of the mat-forming cyanobacterium Microcoleus vaginatus may have led to the speciation events. On the other hand, 16S-23S ITS phy- logenies of the freshwater, planktic cyanobacterium Microcystis aeruginosa revealed no connection between geographic position and a placement in phylogeny (van Gremberghe et al. 2011). Further, no geographical patterning has been observed in polar cyanobacteria based on 16S rRNA (Jungblut et al. 2010). Taken together, the role of geographical isolation as it relates to the speciation of microbes should be further investigated using whole genome data or using more variable genome regions (Ramette and Tiedje 2007). A whole genome approach has been used in thermophilic archeon Sulfolobus islandicus, which has shown clear geographical patterning (Reno et al. 2009). This may reveal very recent events of genetic exchange leading to speciation as in case of marine picoplanktic Synechococcus (e.g. Mazard et al. 2012) because the geographical isolation may be im- portant in a very short time frame (Ramette and Tiedje 2007; Dvorˇa´k et al. 2012) rather than in relatively long times as observed in macroorganisms.

Cohan (2001) advocates a bacterial speciation model by ecological diversification. When a new niche is introduced the stable ecotype is periodically overgrown by new, ‘‘fitter’’ ecotypes, which are able to effectively exploit the new niche. Periodic selection events decrease the overall genetic diversity and after some time a new ecotype can be sufficiently diversified to form a new stable ecotype, which may be non-competing with the parental genotype. This also results with the coherence within evolutionary linages.

The ecotype model of prokaryotic speciation proposed by Cohan (2001, 2002, 2006) and Cohan and Perry (2007) treat bacteria as asexual clones, where homologous recom- bination rates are low. Thus, many crucial questions still remain as unanswered concerning the mechanism of bacterial speciation.

Particular problems of species definitions and concepts in cyanobacteria

The previously mentioned concepts and definitions of species are also applicable to cyanobacteria, but in the following paragraphs, we will emphasize some important con- siderations pertaining to cyanobacteria.

A classic, phenetic species concept using only morphological or ecological data has been shown to be insufficient to describe the real biodiversity within cyanobacteria. Morphology alone in cyanobacteria often lacks resolution on the species level, while completely ignoring cryptic species (e.g. Johansen and Casamatta 2005; Hasˇler et al. 2012, and many others, see further). Cyanobacterial species have traditionally been distinguished based on

the similarity of morphological markers, which might be very subjective. Moreover, some morphological characters, such as sheath formation or presence of hete- rocytes, may be lost in cultures and environmentally plastic. For example, Microcoleus vaginatus, which is usually found in soil crusts, puddles and other aerophytic habitats, has multiple filaments enclosed in common sheath. However, strains isolated from epipelon (fine lake sediment) produce no sheath in nature or culture. An analysis of morphology, 16S rRNA-based phylogeny and 16S-23S ITS secondary structure revealed very close relations with soil crust M. vaginatus strains (Hasˇler et al. 2012). 16S rRNA of all strains also contained an 11 bp insert typical for this species (Boyer et al. 2002). Phenotypic characters (i.e. cell dimension, division type, color) provided insufficient resolution for discerning these lineages. The employment of new characters (mostly 16S-23S ITS region) have allowed researchers to recognize finer differences among taxa with coherent mor- phology, leading to the idea of cryptic speciation (Boyer et al. 2001; Siegesmund et al. 2008; Koma´rek 2010, 2011; Hasˇler et al. 2012). Cryptic taxa are unrecognizable using solely morphological characters. Cryptic species have been identified or suggested in almost all traditional genera (Koma´rek 2010) such as with the mat-forming cyanobacteria Microcoleus (Siegesmund et al. 2008), Oculatella (Osorio-Santos et al. 2014), Tri- chocoleus (Mu¨hlsteinova´ et al. 2014) and Phormidium (Casamatta et al. 2003; Hasˇler et al. 2012). This topic is discussed in great extent elsewhere (e.g. Johansen and Casamatta 2005; Koma´rek 2010). It should be noted that genus Oculatella consists of 7 cryptic species (Osorio-Santos et al. 2014), which were able to be resolved based on 16S-23S ITS se- quence, which has higher resolution under the genus level.

A majority of recent taxonomic revisions and descriptions use a combination of mor- phological, ecological, and genetic observations, referred to as a polyphasic approach (Castenholz 1992; Castenholz and Norris 2005; Koma´rek 2003, 2010; Koma´rek et al. 2014). It has already been employed to recognize separate evolutionary lineages and for description of new species. A polyphasic approach is commonly used in taxonomic works in combination with a monophyletic species concept sensu Johansen and Casamatta (2005), if phylogenetic analyses of 16S rRNA or other genes are used.

16S rRNA sequencing and progress in phylogenetic reconstruction have allowed re- searchers to employ the evolutionary species concept sensu Simpson (1953). This has fa- cilitated further derived species concepts like the monophyletic species concept. Johansen and Casamatta (2005) used this to define a species as the smallest monophyletic group with recognizable autapomorphy (a trait unique only for particular taxon). They also designed a concrete species definition based on the mentioned concept, which is suitable for cyanobacteria and may be used under the International Code of Botanical Nomenclature. It is probably the most widely accepted concept with cyanobacteria (according to the number of taxonomic papers using the concept

under the Botanical Code), although sometimes not precisely followed (Siegesmund et al. 2008; Perkerson et al. 2011; Dvorˇa´k et al. 2014a; Hasˇler et al. 2012, 2014a, b; Osorio-Santos et al. 2014 and many others). The most important advantage of the monophyletic species concept is the general applicability to asexual or- ganisms. However, it might be problematic when a monophyletic lineage lacks sufficient morphological, ecological or physiological differentiation. Moreover, monophyletic species concept is not accepted in the International Code for Nomenclature of Bacteria, which uses species concepts mentioned above.

Synechococcus sensu lato is a group of cyanobacteria with cosmopolitan distribution inhabiting almost all environments (Koma´rek and Anagnostidis 1998), including thermal and aerophytic habitats (Honda et al. 1999; Robertson et al. 2001). Although some cyanobacteria, such as Synechococcus sensu lato, lack phenotypic variability, great eco- logical and genetic diversity suggest that polyphyletic complexes of cryptic taxa might exist (Honda et al. 1999; Robertson et al. 2001; Dvorˇa´k et al. 2014a, b). Many traditional cyanobacterial genera (Geitler 1932) are polyphyletic (Koma´rek 2010; Engene et al. 2011; Hasˇler et al. 2012; Engene et al. 2013; Dvorˇa´k et al. 2014a, b; Hasˇler et al. 2014a) and need extensive revisions, which will be very difficult particularly in the case of Synehococcus sensu lato due to extreme polyphyly within this genus. We suggest that this extreme polyphyly (cryptogenera sensu Koma´rek et al. 2014) should be distinguished from poly- phyly in the original sense, since in extreme polyphyly a large number of lineages derived over very long time period (over 3 billion years), as shown in Dvorˇa´k et al. (2014b).

Such extreme polyphyletic groups with little morphological distinction, in which similar morphotypes may belong to polyphyletic lineages and hence different genera, suggest that taxonomic revisions based solely on morphological data must be performed with great caution. Thus, stable molecular markers such as 16S rRNA should be used for taxonomic revisions (see Koma´rek 2010 for a review). However, even revisions of genera without molecular support for all studied species have been recently proposed. For instance, after recent revisions of polyphyletic genera with molecular markers, some authors have added new species based on morphological similarity as new combinations (Strunecky´ et al. 2014). However, these species might be polyphyletic taxa in a manner similar to Syne- chococcus. Therefore, we recommend the use of molecular data in all cases to increase the certainty of taxonomic revisions.

A growing number of polyphyletic genera recently identified might be connected with frequent convergent evolutionary events in cyanobacteria. Convergent evolution is a phenomenon that occurs when similar features have evolved in independent lineages. It seems to be very frequent in cyanobacteria and is evidenced by several phenotypic traits (e.g. Shishido et al. 2013;

Dvorˇa´k et al. 2014b). We have chosen the example of the prochlorophytes to show another case of convergence in cyanobacteria.

Prochlorococcus, Prochlorothrix, and Prochloron are cyanobacteria that additionally pro- duce chlorophyll b (the typical pigment of green algae and land plants) and lack phycobilisomes (Giddings et al. 1980; Burger-Wiersma et al. 1986; Miller et al. 1988; Chisholm et al. 1992; Hess et al. 1996; Pinevich et al. 1997; Kauff and Bu¨del 2011). On this basis and due to a psbA gene based phylogenetic analysis, they were considered strictly associated with the chloroplast of green algae and terrestrial plants (Morden and Golden 1989). Successive analyses (e.g., Litvaitis 2002) showed that prochlorophytes actually nested within cyanobacteria, and are polyphyletic. The conclusion is that the appearance of chlorophyll b and the loss of phycobilisomes evolved multiple times in different lineages, and hence these characters are subjected to convergent evolution and reversals, probably in connection to environmental pressures.

We note that Prochlorococcus marinus appears to cluster quite clearly apart from the other cyanobacteria on the basis of the analysis of all the tRNA sequences, considering the isoacceptor variation for each codon and the number of copies for each type of tRNA (Fig. 3). This suggests convergent or parallel evolutionary events leading to similar phenotypic traits, because it contradicts phylogenomic analyses in Shih et al. (2013). Such convergent events might be explained by HGT within the environment and therefore environmental pressures (Litvaitis 2002). It is also likely that it represents a frequent trend in cyanobacterial evolution, since other morphological traits, such as multicellularity, have evolved repeatedly (Honda et al. 1999; Robertson et al. 2001; Schirrmeister et al. 2013; Dvorˇa´k et al. 2014a, b). Dvorˇa´k et al. (2014b, Fig. 4) also suggested a model of serial convergence in cyanobacteria, where frequent convergent events might be explained by constant genetic changes via HGT and HR within local habitat gene pools as proposed by Polz et al. (2013).

Diversity of cyanobacteria and their current classification

Taxonomy is usually defined as an operative version of systematics. Both the taxonomy and systematics of cyanobacteria have undergone substantial changes in the last two decades. Previously, the cyanobacteria were placed into three botanical orders, the number of which has changed with respect to the state of investigation of morphological variability and ecology of the species. Geitler (1932) revised the systematics of cyanobacteria established in the nineteenth century and proposed three orders: Chroococcales (coccoidal species reproducing by binary fission), Chamaesiphonales (a heteropolar type of binary fission), and Hormogonales (the filamentous species). Other authors of the twentieth century usually followed Geitler´s botanical system. However,

their systems changed as additional characters were uncovered and additional taxa included, e.g. Desikachary (1959) distinguished five orders (Chroococcales, Chamaesiphonales, Pleurocapsales, Nostocales, and Stigonematales), and Starmach (1966) split the system of cyanobacteria into four classes (Chroococcophyceae, Chamaesiphonophyceae, Pleuastrophyceae, and Hormogoniophyceae).

Later, in the 1970s, a bacteriological approach was used in the classification of cyanobacteria (Stanier et al. 1978). Five subgroups, corresponding to the orders Chroococcales, Pleurocapsales, Oscillatoriales, Nostocales, and Stigonematales, were classified with respect to the type of cell reproduction, cell differentiation, and molecular/ biochemical attributes (Rippka et al. 1979; Boone and Castenholz 2001). This classifica- tion concept facilitated substantial progress in the research on cyanobacteria because of the new methods advocated.

The most comprehensive studies on the classification of cyanobacteria in the modern era were made by Anagnostidis and Koma´rek (1985, 1988, 1990; Koma´rek and Anagnostidis 1986, 1989). The authors combined both botanical and bacteriological approaches, inte- grating traditional cyanobacterial morphology, physiology, and ecology in a total evidence synthesis. They established four orders: Chroococcales (non-filamentous), Oscillatoriales (filamentous, lacking specialized cells), Nostocales (filamentous, facultative specialized cells), and Stigonematales (filamentous, obligatory specialized cells, and division in multiple planes). During the 1990s, analysis of the 16S rDNA gene elucidated and sup- ported phylogenetic relationships among morphologically similar genera, and, surprisingly, among genera from different orders as defined by Anagnostidis and Koma´rek. Hoffmann et al. (2005) proposed a new system of classification where members of the Chroococcales and Oscillatoriales formed two subclasses, the Synechococcophycideae and the Oscillar- iophycidae. Members of the Nostocales and Stigonematales belonged to a separate monophyletic subclass, the Nostochophycideae. Terminal taxonomic units (genera and species) represent a crucial element in the world of cyanobacteria. Numerous new genera are being erected because molecular methods usually show a higher diversity than the traditional botanical (morphological) approach by providing additional character sets (cryptic species). This topic is discussed in greater extent above. This system of higher taxonomic ranks has been recently re-evaluated in a review by Koma´rek et al. (2014). These authors proposed a subdivision of cyanobacteria based on phylogeny and mor- phology in the following orders: Gloeobacterales, Synechococcales, Spirulinales, Pleuro- capsales, Chroococcales, Chroococcidiopsidales, Oscillatoriales, and Nostocales.

The higher level systematic classification of cyanobacteria needs more investigation based on revised genera. A complete revision should include morphological description of natural populations based on light and electron

microscopy, habitat characterization, molecular analysis of 16S rRNA gene and other markers such as ITS region and infor- mation about stored strains or DNA. Moreover, important consideration should be given to biochemical/bioorganic data, e.g. fatty acids composition of cyanobacterial cell wall, which seems to be applicable for species identification (Caudales et al. 2000; Rˇ ezanka et al. 2003; Li and Watanabe 2004).

Whenever a wide agreement on species concepts in cyanobacteria is reached, some practical identification of species and other taxa open another ample array of problems. For instance, cyanobacteria may be described under both the International Code for Algae, Fungi and Plants (ICN, http://www.iapt-taxon.org/nomen/main.php) and the International Code for Nomenclature of Prokaryotes (ICNP), although the vast majority of cyanobac- terial taxa are described under the Botanical Code (Oren 2011). The reason for that are the strict requirements of the ICNP, i.e. axenic culture and DNA–DNA hybridization etc. Detailed values, description and discussion may be found on the website of the Interna- tional Committee on Systematics of Prokaryotes (http://icsp.org/; Starkerbrandt et al. 2002; Oren and Garrity 2014). Some additional problems in the application of the ICNP to cyanobacteria are discussed in Oren (2004, 2011), and Oren and Tindall (2005). An at- tempt to develop a special code valid only for cyanobacteria has been proposed at the Meeting of the International Association for Cyanophyte Research in Luxembourg in 2004 (http://www.cyanodb.cz/files/CyanoGuide.pdf). However, it is an unofficial document that has not yet been accepted. Thus, a schism among cyanobacteriologists still continues, but a number of authors largely favor the Botanical Code, because new taxa might be described without cultures (e.g. Hasˇler et al. 2014a).

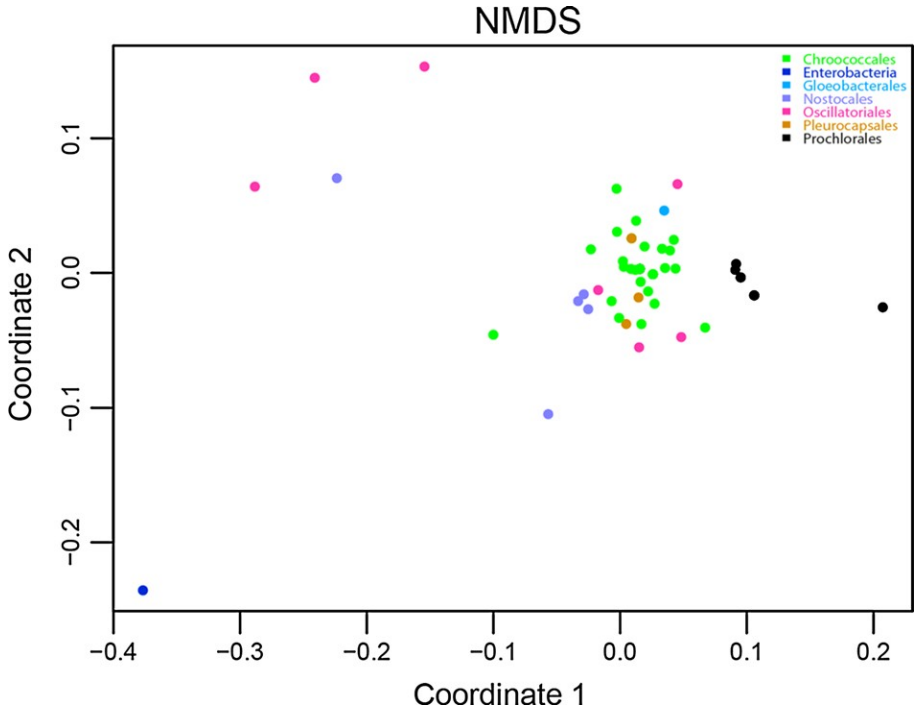


Fig. 3 Plot derived from MDS analysis representing the diversity of cyanobacterial strains and lineages on the basis of variation in tRNA isoacceptors (tRNA targeting considering also different tRNAs but with the same anticodon) for each amino acids types and copy number calculated on the known complete genomes of cyanobacteria. Only 5 of 9 accessions of prochlorophytes are visible, since 4 accessions are completely overlapping with the others. (Color figure online)

Estimate of the total cyanobacterial biodiversity

Culture-independent estimates of prokaryotic biodiversity fall between millions and bil- lions of species (e.g., Dykhuizen 1998; Gans et al. 2005). Estimates of the current cyanobacterial biodiversity range from 2000 (San´tAnna et al. 2006) to 8000 (Guiry 2012). Nabout et al. (2013) applied a discovery curve to cyanobacteria utilizing the CyanoDB database (http://www.cyanodb.cz/) with three asymptotic models, yielding from 3166 to 6280 species, depending on the model of choice. A total of 453 authors have described cyanobacterial taxa, and two of them (J. Koma´rek and K. Anagnostidis) have described

30.9 % of the total described species (Nabout et al. 2013). However, the real number of species can be barely assessed by statistics. It requires extensive

observation of the species diversity and distribution in nature (Foissner 2006) with subsequent quantification. In the future, genetic and molecular data will be increasingly helpful. For example, the last decade of polyphasic studies brought tens of newly erected or revised cyanobacterial genera (e.g. see Koma´rek 2010 for review of older works, afterwards e.g. Strunecky´ et al. 2011; Koma´rek et al. 2013; Koma´rkova´ et al. 2013; Dvorˇa´k et al. 2014a). During the 19th Symposium of the International Society for Cyanophyte Research in 2013, 16 new genera were presented (Koma´rek et al. 2014). Thus, the great atomization of cyanobacterial systematics is now in progress, which is a result of species definition, concept used, and introduction of molecular methods into cyanobacterial systematics. Moreover, with higher resolution abilities, we can expect a further expansion of the number of described taxa. For example, Oculatella erected with single species (Zammit et al. 2012) now contains seven species, which have been described by different researchers. Thus, even the most liberal estimates may be undervalued.

Conclusions and future directions

Great challenges lie ahead in regards to the taxonomy and systematics of cyanobacteria. Fortunately, molecular techniques have facilitated a renaissance in describing and elucidating cyanobacterial biodiversity. In this review, we showed that although cyanobacteria lack sexual reproduction, we are able to apply, in terms of evolutionary- lineage coherence, a species concept similar to that one used for eukaryotic macroor- ganisms, even though it might be considered ‘‘fuzzy’’ due to the molecular markers ap- plied, homologous recombination or horizontal gene transfer. However, many questions remain regarding cyanobacterial species definitions and concepts. Caution must be maintained, though, as morphology is sometimes in conflict with molecular markers, or has limited resolution. Therefore, cryptic species and extremely polyphyletic genera caused by serial convergence represent problematic phenomena resulting with uncertainty of proper morphological identification. We suggest that more attention should be paid to the use of molecular markers in taxonomy and practical identification of taxa. On the other hand, ecological and morphological criteria are also important, which should be taken into consideration. Thus, deposited sequences in GenBank and other databases should also be completed with such data or they should be made easily accessible by providing the original papers. These data may be afterwards a source for a barcoding database, which would provide correct and fast identification workflow, and would resolve cryptic taxa and polyphyletic genera problems.

The rapidly growing number of described taxa signifies large gaps in our current knowledge of cyanobacterial biodiversity and distribution. Although the total biodiversity of any microbial lineage is probably unknown, all

estimations suggest a significant increase of described taxa. Moreover, the selected species definition will impact on how many species are identified and will be recognized in the future. It also largely influences pos- sible patterns of distribution. Thus, evidently, we are now in a period of important changes in taxonomy, and knowledge of cyanobacterial biodiversity is amplified by novel tech- niques, and increasing sampling effort.

Methods of analysis

Multivariate statistics have been carried out by using the R software 3.0 (R Development Core Team 2013) and some functions included in Vegan and MASS packages (Venables and Ripley 2002; Oksanen et al. 2013). Transfer RNA data of all the analyzed organisms have been imported in R as dataframe. The distance matrix has been computed using the ‘‘vegdist’’ function and selecting ‘‘jaccard’’ as method. The quantitative form of the Jac- card distance in Vegan actually is the Ruzicka index and has been preferred over the Euclidean distance for its better performances in presence of species containing missing tRNA (counts equal to zero). In order to visualize the distances between organisms, data have been statistically explored through unconstrained ordination by computing a non- metric multidimensional scaling (NMDS) using the ‘‘metaMDS’’ function included in the Vegan package. Multidimensional Scaling helps to visualize the distance between samples through a low-dimensional spatial map. The non-metric scaling methods are able to map non-Euclidean distances.

References

Achtman M, Wagner M (2008) Microbial diversity and the genetic nature of microbial species. Nat Rev Microbiol 6:431–440

Anagnostidis K, Koma´rek J (1985) Modern approach to the classification system of cyanophytes 1—introduction. Algol Stud 38(39):291–302

Anagnostidis K, Koma´rek J (1988) Modern approach to the classification system of cyanophytes

3—Oscillatoriales. Algol Stud 50–53:327–472

Anagnostidis K, Koma´rek J (1990) Modern approach to the classification system of cyanophytes 5—Stigonematales. Algol Stud 59:1–73

Baas Becking LGM (1934) Geobiologie of inleiding tot de miliekunde. W. P. van Stockum, The Hague

Bekker A, Holland HD, Wang P-L et al (2004) Dating the rise of atmospheric oxygen. Nature 427:117–120 Boone, Castenholz RW (2001) Bergey‘s manual of systematic bacteriology volume 1: the Archaea and the
Deeply branching and phototrophic Bacteria. Springer, New York
Boyer SL, Fletchner V, Johansen JR (2001) Is the 16S-23S rRNA internal transcribed spacer (ITS) region a good tool for use in molecular systematics and population genetics? A case study in cyanobacteria. Mol Biol Evol 18:1057–1069
Boyer SL, Johansen JR, Howard GL (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. J Phycol 38:1222–1225
Burger-Wiersma T, Veenhuis M, Korthals HJ, Van de Wiel CCM, Mur LR (1986) A new prokaryote containing chlorophylls a and b. Nature 320:262–264
Casamatta DA, Vis ML, Sheath RG (2003) Cryptic species in cyanobacterial systematics: a case study of
Phormidium retzii (Oscillatoriales) using 16S rDNA and RAPD analyses. Aquat Bot 77:295–309 Castenholz RW (1992) Species usage, concept, and evolution in the cyanobacteria (blue-green algae).
J Phycol 28:737–745
Castenholz RW, Norris TB (2005) Revisionary concepts of species in the Cyanobacteria and their appli- cations. Algol Stud 117:53–56
Caudales R, Wells JM, Butterfield JE (2000) Cellular fatty acid composition of cyanobacteria assigned to
subsection II, order Pleurocapsales. Int J Syst Evol Microbiol 50:1029–1034
Chisholm SW, Frankel SL, Goericke R, Olson RJ, Palenik B, Waterbury JB, West-Johnsrud L, Zettler ER (1992) Prochlorococcus marinus gen. nov.: an oxyphototrophic marine prokaryote containing divinyl
chlorophyll a and b. Arch Microbiol 157:297–300
Cohan FM (2001) Bacterial species and speciation. Syst Biol 50:513–524 Cohan FM (2002) What are bacterial species? Annu Rev Microbiol 56:457–487
Cohan FM (2006) Towards a conceptual and operational union of bacterial systematics, ecology, and evolution. Philos Trans R Soc Lond B Biol Sci 361:1985–1996

Cohan FM (2011) Are species cohesive? A view from bacteriology. In: Walk ST, Feng PCH (eds)
Population genetics of bacteria: a tribute to Thomas S. Whittam, ASM Press, Washington DC Cohan FM, Perry EB (2007) A systematics for discovering the fundamental units of bacterial diversity. Curr
Biol 17:R373–R386
Daubin V, Gouy M, Perriere G (2002) A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. Genome Res 12:1080–1090
David LA, Alm EJ (2011) Rapid evolutionary innovation during an Archean genetic expansion. Nature
469:93–96
de Queiroz K (2005) Ernst Mayr and the modern concept of species. Proc Natl Acad Sci USA 102:6600–6607
de Queiroz K (2007) Species concepts and species delimitation. Syst Biol 56:879–886
Desikachary TV (1959) Cyanophyta. In: Randhawa MS (ed) I.C.A.R. monographs on algae. Indian Council of Agriculture Research, New Dehli
Doolittle WF, Zhaxybayeva O (2009) On the origin of prokaryotic species. Genome Res 19:744–756 Dvoˇra´k P, Hasˇler P, Poul´ıcˇkova´ 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

Dvoˇra´k P, Casamatta DA, Hasˇler P, Ondˇrej V, Poul´ıcˇkova´ A, Sanges R (2014a) Synechococcus: 3 billion years of global dominance. Mol Ecol 23:5538–5551
Dvoˇra´k P, Hinda´k F, Hasˇler P, Hinda´kova´ A, Poul´ıcˇkova´ A (2014b) Morphological and molecular studies of
Neosynechococcus sphagnicola, gen. et sp. nov. (Cyanobacteria, Chroococcales). Phytotaxa 170:24–34 Dykhuizen DE (1998) Santa Rosalia revisited: why are there so many species of bacteria? Antonie Van
Leeuwenhoek 73:25–33

Eckert EM, Fontaneto D, Coci M, Callieri C (2015) Does a barcoding gap exist in prokaryotes? Evidences from species delimitation in cyanobacteria. Life 5:50–64

Engene N, Choi H, Esquenazi E, Rottacker EC, Ellisman MH, Dorrestein PC, Gerwick WH (2011) Un- derestimated biodiversity as a major explanation for the perceived prolific secondary metabolite ca-
pacity of the cyanobacterial genus Lyngbya. Environ Microbiol 13:1601–1610

Engene N, Gunasekera SP, Gerwick WH, Paul VJ (2013) Phylogenetic inferences reveal large extent of novel biodiversityin chemically rich tropical marine cyanobacteria. Appl Environ Microbiol 79:1882–1888

Foissner W (2006) Biogeography and dispersal of micro-organisms: a review emphasizing protists. Acta Protozool 45:111–136

Fontaneto D, Barraclough TG, Chen K, Ricci C, Herniou EA (2008) Molecular evidence for broad-scale
distribution in bdelloid rotifers: everything is not everywhere but most things are very widespread. Mol Ecol 17:3136–3146

Fox GE, Wisotzkey JD, Jurtshuk P Jr (1992) How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. Int J Syst Bacteriol 42:166–170

Fraser C, Hanage WP, Spratt BG (2007) Recombination and the nature of bacterial speciation. Science
315:476–480

Friedman J, Alm EJ, Shapiro BJ (2013) Sympatric speciation: when is it possible in bacteria? PLoS One 8:e53539

Gans J, Wolinsky M, Dunbar J (2005) Computational improvements reveal great bactgerial diversity and
high metal toxicity in soil. Science 309:1387–1390

Geitler L (1932) Cyanophyceae. In: Rabenhorst L (ed) Kryptogamenflora. Akademische Verlagsge- sellschaft, Leipzig

Gevers D, Cohan FM, Lawrence JG et al (2005) Opinion: re-evaluating prokaryotic species. Nat Rev Microbiol 3:733–739

Gevers D, Dawyndt P, Vandamme P, Willems A, Vancanneyt M, Swings J, De Vos P (2006) Stepping stones towards a new prokaryotic taxonomy. Philos Trans R Soc Lond B Biol Sci 361:1911–1916

Giddings TH, Withers NW, Staehelin LA (1980) Supramolecular structure of stacked and unstacked regions of the photosynthetic membranes of

Prochloron sp., a prokaryote. Proc Natl Acad Sci USA
77:352–356
Goris J, Konstantinidis JT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM (2007) DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol
Microbiol 57:81–91
Guiry MD (2012) How many species of algae are there? J Phycol 48:1057–1063
Hacker J, Carniel E (2001) Ecological fitness, genomic islands and bacterial pathogenicity—a Darwinian view of the evolution of microbes. EMBO Rep 2:376–381
Hanage WP (2013) Fuzzy species revisited. BMC Biol 11:41
Hanage WP, Fraser C, Spratt BG (2005) Fuzzy species among recombinogenic bacteria. BMC Biol 3:6 Hanage WP, Fraser C, Spratt BG (2006) Sequences, sequence clusters and bacterial species. Philos Trans R
Soc Lond B Biol Sci 361:1917–1927
Hasˇler P, Dvoˇra´k P, Johansen JR, Kitner M, Ondˇrej V, Poul´ıcˇkova´ A (2012) Morphological and molecular study of epipelic filamentous genera Phormidium, Microcoleus and Geitlerinema (Oscillatoriales,
Cyanophyta/Cyanobacteria). Fottea 12:341–356
Hasˇler P, Dvoˇra´k P, Poul´ıcˇkova´ P (2014a) A new genus of filamentous epipelic cyanobacteria, Johansenia.
Preslia 86:81–94
Hasˇler P, Dvoˇra´k P, Poul´ıcˇkova´ P, Casamatta DA (2014b) A novel genus Ammassolinea gen. nov. (Cyanobacteria) isolated from sub–tropical epipelic habitats. Fottea 14:241–248
Hess WR (2011) Cyanobacterial genomics for ecology and biotechnology. Curr Opin Microbiol 14:608–614 Hess WR, Partensky F, van der Staay GWM, Garcia-Fernandez JM, Borner T, Vaulot D (1996) Coexistence
of phycoerythrin and a chlorophyll a/b antenna in a marine prokaryote. Proc Natl Acad Sci USA 93:11126–11130
Hoffmann L, Koma´rek J, Kasˇtovsky´ J (2005) System of Cyanoprokaryotes (cyanobacteria) state in 2004.
Algol Stud 117:95–115
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. J Mol Evol 48:723–739

Johansen JR, Casamatta DA (2005) Recognizing cyanobacterial diversity through adoption of a new species paradigm. Algol Stud 117:71–93

Johnson ZI, Zinse ER, Coe A, McNulty NP, Woodward EMS, Chisholm SW (2006) Niche partitioning

among Prochlorococcus ecotypes along ocean-scale environmental gradients. Science 311:1737–1740 Jungblut AD, Lovejoy C, Vincent WF (2010) Global distribution of cyanobacterial ecotypes in the cold

biosphere. ISME J 4:191–202

Kauff F, Bu¨del B (2011) Phylogeny of cyanobacteria: an overview. Prog Bot 72:209–224

Kim M, Oh HS, Park SC, Chun J (2014) Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarty for species demarcation of prokaryotes. Int J Syst Evol

Microbiol 64:346–351

Koeppel AF, Wertheim JO, Barone L, Gentile N, Krizanc D, Cohan FM (2013) Speedy speciation in bacterial microcosm: new species can arise as frequently as adaptations within a species. ISME J 7:1080–1091

Koma´rek J (2003) Problem of the taxonomic category ‘‘species’’ in cyanobacteria. Algol Stud 109:281–297 Koma´rek J (2010) Recent changes (2008) in cyanobacteria taxonomy based on a combination of molecular

background with phenotype and ecological consequences (genus and species concept). Hydrobiologia

1:245–259

Koma´rek J (2011) Introduction to the 18th IAC Symposium in C˘ eske´ Budeˇjovice 2010, Czech Republic— some current problems of modern cyanobacteria taxonomy. Fottea 11:1–7

Koma´rek J, Anagnostidis K (1986) Modern approach to the classification system of cyanophytes 2— Chroococcales. Algol Stud 43:157–226

Koma´rek J, Anagnostidis K (1989) Modern approach to the classification system of cyanophytes 4— Nostocales. Algol Stud 56:247–345

Koma´rek J, Anagnostidis K (1998) Cyanoprokaryota 1. Teil: Chroococcales. In: Ettl H, Ga¨rtner G, Heynig H, Mollenhauer D (eds) Su¨sswasserflora von Mitteleuropa 19/1. Gustav Fischer, Jena-

Stuttgart-
Lu¨beck-Ulm
Koma´rek J, Sant'Anna C, Bohunicka´ M, Mareš J, Hentschke GS, Rigonato J, Fiore M (2013) Phenotype diversity and phylogeny of selected Scytonema-species (Cyanoprokaryota from SE Brazil). Fottea
13:173–200
Koma´rek J, Kaštovsky´ J, Mareš J, Johansen JR (2014) Taxonomic classification of cyanoprokaryotes (cyanobacterial genera) 2014 using a polyphasic approach. Preslia 86:295–335
Koma´rkova´ J, Zapomeˇlova´ E, Koma´rek J (2013) Chakia (cyanobacteria), a new heterocystous genus form
Belizean marshes identified on the basis of the 16S rRNA gene. Fottea 13:227–233
Konstantinidis KT, Tiedje JM (2005) Genomic insights that advance the species definition for prokaryotes.
Proc Natl Acad Sci USA 102:2567–2572
Konstantinidis KT, Ramette A, Tiedje JM (2006) The bacterial species definition in the genomic era. Philos Trans R Soc Lond B Biol Sci 361:1929–1940
Kopac S, Wang Z, Wiedenbeck J, Sherry J, Wu M, Cohan FM (2014) Genomic heterogeneity and ecological speciation within one subspecies of Bacillus subtilis. Appl Environ Microbiol 80:4842–4853
Li R, Watanabe MM (2004) Fatty acid composition of planktonic species of Anabaena (cyanobacteria) with coiled trichomes exhibited a significant taxonomic value. Curr Microbiol 49:376–380
Litvaitis MK (2002) A molecular test of cyanobacterial phylogeny: inferences from constraint analyses.
Hydrobiology 468:135–145
Lodders N, Stackebrandt E, Nubel U (2005) Frequent genetic recombination in natural populations of the marine cyanobacterium Microcoleus chthonoplastes. Environ Microbiol 7:434–442
Maiden MC, Bygraves JA, Feil EJ et al (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci USA
95:3140–3145
Marston MF, Pierciey FJ Jr, Shepard A, Gearin G, Qi J, Yandava C, Schuster

SC, Henn MR, Martiny JB (2012) Rapid diversification of coevolving marine Synechococcus and virus. Proc Natl Acad Sci USA 109:4544–4549

Martiny JBH, Bohanna BJM, Brown JH et al (2006) Microbial biogeography: putting microorganisms on the map. Nat Rev Microbiol 4:102–112

Mayr E (1942) Systematics and the origin of species from the viewpoint of a zoologist. Columbia University

Press, New York

Mayr E (1946) What is a species, and what is not? Philos Sci 63:262–277

Mazard S, Ostrowski M, Partensky F, Scanlan DJ (2012) Multi-locus sequence analysis, taxonomic resolution and biogeography of marine Synechococcus. Environ Microbiol 14:372–386

Melendrez MC, Lange RK, Cohan FM, Ward DM (2011) Influence of molecular resolution on sequence- based discovery of ecological diversity among Synechococcus populations in an alkaline siliceous hot

spring microbial mat. Appl Environ Microbiol 77:1359–1367

Mes TH (2008) Microbial diversity—insights from population genetics. Environ Microbiol 10:251–264 Miller KR, Jacob JS, Burger-Wiersma T, Matthijs HC (1988) Supramolecular structure of the thylakoid

membrane of Prochlorothrix hollandica: a chlorophyll b-containing prokaryote. J Cell Sci 91:577–586

Miller SR, Castenholz RW, Pedersen D (2007) Phylogeography of the thermophilic cyanobacterium

Mastigocladus laminosus. Appl Environ Microbiol 73:4751–4759

Morden CW, Golden SS (1989) psbA genes indicate common ancestry of prochlorophytes and chloroplasts.

Nature 337:382–385

Mu¨hlsteinova´ R, Johansen JR, Pietrasiak N, Martin MP, Osoirio-Santos K, Warren SD (2014) Polyphasic characterization of Trichocoleus desertorum sp. nov. (Pseudanabaenales, Cyanobacteria) from desert

soils and phylogenetic placement of the genus Trichocoleus. Phytotaxa 163:241–261

Nabout JC, da Silva Rocha B, Carneiro FM, Sant'Anna CL (2013) How many species of Cyanobacteria are there? Using a discovery curve to predict the species number. Biodivers Conserv 22:2907–2918

Nakhleh L (2013) Computational approaches to species phylogeny inference

and gene tree reconciliation.
Trends Ecol Evol 28:719–727
NarechaniaA Baker RH, Sit R, Kolokotronis SO, DeSalle R, Planet PJ (2012) Random addition concate- nation analysis: a novel approach to the exploration of phylogenomic signal reveals strong agreement between core and shell genomic partitions in the cyanobacteria. Genome Biol Evol 4:30–43
Nu¨rnberg DJ, Mariscal V, Parker J, Mastroianni G, Flores E, Mullineaux CW (2014) Branching and intercellular communication in the Section V cyanobacterium Mastigocladus laminosus, a complex multicellular prokaryote. Mol Microbiol 91:935–949
Oksanen J, Blanchet FG, Kindt R, Legendre P et al. (2013) Vegan: community ecology package. R package version 2.0-10. http://CRAN.R-project.org/package=vegan
Oren A (2004) A proposal for further integration of the cyanobacteria under the Bacteriological Code. Int J
Syst Evol Microbiol 54:1895–1902
Oren A (2011) Cyanobacterial systematics and nomenclature as featured in the International Bulletin of Bacteriological Nomenclature and Taxonomy/International Journal of Systematic Bacteriology/International Journal of Systematic and Evolutionary Microbiology. Int J Syst Evol Microbiol 61:10–15
Oren A, Garrity GM (2014) Then and now: a systematic review of the systematics of prokaryotes in the last 80 years. Antonie Van Leeuwenhoek 106:43–56
Oren A, Tindall BJ (2005) Nomenclature of the cyanophyta/cyanobacteria/cyanoprokaryotes under the
International Code of Nomenclature of Prokaryotes. Algol Stud 117:39–52
Osorio-Santos K, Pietrasiak N, Bohunicka´ M, Miscoe LH, Kova´cik L, Martin MP, Johansen JR (2014) Seven new species of Oculatella (Pseudanabaenales, Cyanobacteria). Eur J Phycol 49:450–470
Papke RT, Ramsin NB, Bateson MM, Ward DM (2003) Geographical isolation in hot spring cyanobacteria.
Environ Microbiol 5:650–659
Perkerson RB, Johansen JR, Kova´cˇik L, Brand J, Kasˇtovsky´ J, Casamatta DA (2011) A unique pseudan- abaenalean (cyanobacteria) genus

Nodosilinea gen. nov. based on morphological and molecular data.
J Phycol 47:1397–1412
Pinevich AV, Averina SG, Velichko NV (1997) Another view on the role of photosynthetic pigments in taxonomy of oxygenic-phototrophic bacteria: proposed rejection of the order Prochlorales Florenzano,
Balloni, and Materassi 1986 (emend. Burger-Wiersma, Stal, and Mur 1989), the Family Prochloraceae
Florenzano, Balloni, and Materassi 1986, and family Prochlorotrichaceae Bruger-Wiersma, Stal, and Mud 1989. Int J Syst Bacteriol 47:1264–1267
Polz MF, Alm EJ, Hanage WP (2013) Horizontal gene transfer and the evolution of bacterial and archeal population structure. Trends Genet 29:170–175
Popa O, Hazkani-Covo E, Landan G, Martin W, Dagan T (2011) Directed networks reveal genomic barriers
and DNA repair bypasses to lateral gene transfer among prokaryotes. Genome Res 21:599–609 Preheim SP, Timberlake S, Polz MF (2011) Merging taxonomy with ecological population prediction in a
case study of Vibrionaceae. Appl Environ Microbiol 77:7195–7206
R Development Core Team (2013) R: a language and environment for statistical computing. The R Foundation for Statistical Computing, Vienna, Austria. Available at http://www.R-project.org/
Ramette A, Tiedje JM (2007) Biogeography: and emerging cornerstone for understanding prokaryotic
diversity, ecology and evolution. Microb Ecol 53:197–207
Reno ML, Held NL, Fields CJ, Burke PV, Whitaker RJ (2009) Biogeography of the Sulfolobus islandicus
pan-genome. Proc Natl Acad Sci USA 106:8605–8610
R˘ ezanka T, Dor I, Prell A, Dembitsky VM (2003) Fatty acid composition of six freshwater wild cyanobacterial species. Folia Microbiol 48:71–75
Richter M, Rosello´-Mo´ra R (2009) Shifting the genomic gold standard for the prokaryotic species definition.
Proc Natl Acad Sci USA. doi:10.1073/pnas.0906412106
Rippka R, Deruelles J, Waterbury JB, Herdman M, Stanier RY (1979) Generic assignments, strain histories and properties of pure cultures of cyanobacteria. J Gen Microbiol 111:1–61

Robertson BR, Tezuka N, Watanabe M (2001) Phylogenetic analyses of Synechococcus strains (cyanobacteria) using sequences of 16S rDNA and part of the phycocyanin operon reveal multiple
evolutionary lines and reflect phycobilin content. Int J Syst Evol Microbiol 51:861–871

Rodriguez-Valera F, Martin-Cuadrado AB, Rodriguez-Brito B, Pasic L, Thingstad TF, Rohwer F, Mira A (2009) Explaining microbial population genomics through phage predation. Nat Rev Microbiol 7:828–836

Sabehi G, Shaulov L, Silver DH, Yanai I, Harel A, Lindell D (2012) A novel lineage of myoviruses infecting cyanobacteria is widespread in the oceans. Proc Natl Acad Sci USA 109:2037–2042

San´tAnna CL, Azevedo MTP, Agujaro LF, Carvalho MC, Carvalho LR, Souza RCR (2006) Manual ilus-
trado para identificac¸a˜o e contagem de cianobacte´rias plancto^nicas de a´guas continentais brasileiras, lst edn. Editora Intercie^ncia, Rio de Janeiro

Schirrmeister BE, de Vos JM, Antonelli A, Bagheri HC (2013) Evolution of multicellu-larity coincided with increased diversification of cyanobacteria and the great oxidation event. Proc Natl Acad Sci USA
110:1791–1796

Schopf JW (2000) The fossil record: tracing the roots of the cyanobacterial lineage. In: Whitton BA, Potts M (eds) The ecology of cyanobacteria: their diversity in time and space. Springer, Berlin

Seckbach J (2007) Algae and Cyanobacteria in extreme environments. In: Seckbach J (ed) Cellular origin, life in extreme habitats and astrobiology, vol 11. Springer Science & Business Media, New York,
pp 1–811

Shapiro BJ, Friedman J, Cordero OX et al (2012) Population genomics of early events in the ecological differentiation of bacteria. Science 336:48–51

Shi T, Falkowski PG (2008) Genome evolution in cyanobacteria: The stable core and the variable shell. Proc
Natl Acad Sci USA 105:2510–2515

Shih PM, Wu D, Latifi A et al (2013) Improving the coverage of the cyanobacterial phylum using diversity- driven genome sequencing. Proc Natl Acad Sci USA 110:1053–1058

Shishido TK, Kaasalainen U, Fewer DP et al (2013) Convergent evolution of [D-Leucine1] microcystin-LR in taxonomically disparate cyanobacteria.

BMC Evol Biol 13:86
Siegesmund MA, Johansen JR, Karsten U, Friedl T (2008) Coleofasciculus gen. nov. (cyanobacteria): morphological and molecular criteria for revision of the genus Microcoleus Gomont. J Phycol
44:1572–1585
Simpson GG (1953) The major features of evolution. Columbia University Press, New York
Smillie CS, Smith MB, Friedman J, Cordero OX, David LA, Alm EJ (2011) Ecology drives a global network of gene exchange connecting the human microbiome. Nature 480:241–244
Stacerbrandt E, Goebel BM (1994) Taxonomic note: a place for DNA–DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. Int J Syst Bacteriol 44:846–849
Stackebrandt E, Ebers J (2006) Taxonomic parameters revisited: tarnished gold standards. Microbiol Today
33:152–155
Stackebrandt E, Frederiksen W, Garrity GM et al (2002) Report of the ad hoc committee for the re- evaluation of the species definition in bacteriology. Int J Syst Evol Microbiol 52:1043–1047
Staley JT (2006) The bacterial species dilemma and the genomic–phylogenetic species concept. Philos
Trans R Soc Lond B Biol Sci 361:1899–1909
Stanier RY, Sistrom WR, Hansen TA et al (1978) Proposal to place the nomenclature of the cyanobacteria (blue-green algae) under the rules of the international code of nomenclature of bacteria. Int J Syst Bacteriol 28:335–336
Starmach K (1966) Cyanophyta–sinice. Panstwowe Wydawnictwo Naukowe, Warszawa
Stewart EJ (2012) Growing unculturable bacteria. J Bacteriol 196:4151–4160
Strunecky´ O, Elster J, Koma´rek J (2011) Taxonomic revision of the freshwater cyanobacterium ‘‘Pho- midium’’ murrayi = Wilmottia murrayi. Fottea 11:57–71
Strunecky´ O, Koma´rek J, Sˇ marda J (2014) Kamptonema (Microcoleaceae, Cyanobacteria), a new genus
derived from the polyphyletic Phormidium on the basis of combined molecular and cytomorphological markers. Preslia 86:193–207

Sullivan MB, Huang KH, Ignacio-Esponoza JC et al (2010) Genomic analysis of oceanic cyanobacterial myoviruses compared with T4-like myoviruses from diverse hosts and environments. Environ Microbiol 12:3035–3056

Van Gremberghe I, Leliaert F, Mergeay J, Vanormelingen P, Van der Gucht K, Debeer AE, Lacerot G, De Meester L, Vyverman W (2011) Lack of phylogeographic structure in the freshwater cyanobactrium Microcystis aeruginosa suggests global dispersal. PLoS One. doi:10.1371/journal.pone.0019561

Venables WN, Ripley BD (2002) Modern applied statistics with S, 4th edn. Springer, New York

Whitton BA, Potts M (2000) The ecology of cyanobacteria. Their diversity in time and space. Springer, Berlin

Wiedenbeck J, Cohan FM (2011) Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. FEMS Microbiol Rev 35:957–976

Zammit G, Billi D, Albertano P (2012) The subaerophytic cyanobacterium Oculatella subterranea (Os- cillatoriales, Cyanophyceae) gen. et sp. nov.: a cytomorphological and molecular description. Eur J Phycol 47:341–354

Zhaxybayeva O, Gogarten JP (2004) Cladogenesis, coalescence and the evolution of the three domains of life. Trends Genet 20:182–187

Zheng Q, Jiao N, Zhang R, Wei J, Zhang F (2014) The evolutionary divergence of psbA gene in Syne- chococcus and their myoviruses in the East china Sea. PLoS One 9:e86644. doi:10.1371/journal.pone.0086644

REVIEW ARTICLE In submission

II EVOLUTION OF TRNA GENES IN PLASTID GENOMES: A MINIMAL SET SELECTED TO EFFECTIVELY CARRY OUT THE MRNA TRANSLATION.

Mattia Belli, Alessio Papini.

ABSTRACT

Plastid genomes encode their own expression apparatus. We explored about 600 plastid genomes and 60 cyanobacteria genomes in order to study their tRNA gene set in term of both composition and copy number.

Developing suitable computer programs we were able to analyze thousand GenBank annotations and study the differences among main groups of plastid-harboring organisms. In addition, we studied the relationships between tRNA gene copy number and codon usage. Excluding the import of nuclear encoded tRNAs in plastids, wobbling and extended wobbling are fundamental to properly translate all the proteins encoded in plastid genomes. We showed that the correlation between the tRNA gene copy number and the codon usage in mRNAs is statistically significant only in relation to the wobble rules. Moreover we observe that in plastid genomes codon usage correlates with the tRNA pool more than in cyanobacteria genomes.

These observations suggest that the tRNA gene set of plastid genomes have been selected for its capability to effectively carry out the mRNA translation without losing efficiency.

INTRODUCTION

Plastids are a large family of cellular organelles that can be found in many organisms (e.g. plants and algae). The most common are chloroplasts, so named for the presence of chlorophyll pigments and responsible for the conversion of CO_2 to carbohydrates. Other members of the plastid family are chromoplasts and leucoplasts which have synthetic and storage functions. Plastid related organelles include also the cyanelles of glaucocystophyte algae (Lee 2008) and the apicoplasts of the parasitic protists belonging to the apicomplexa phylum (Lau et al. 2009).

As stated by Lynn Margulis (Sagan 1967) and many other biologists, symbiogenesis led to the origin of heterotrophic and autotrophic eukaryotes. Mitochondria originated from the engulfment of an alpha-proteobacterium by an eukaryotic progenitor (Bullerwell 2011) whereas plastids derived from the symbiosis between an eukaryotic cell and a cyanobacterium-like prokaryote (Lee 2008). Three lineages of photosynthetic eukaryotes arose from primary endosymbiosis: glaucophytes, green algae and red algae. Plants evolved from green algae at a later stage (Lee 2008). Events of secondary endosymbiosis led to the formation of plastids surrounded by more than two membranes like euglenids, chlorarachniophytes, dinoflagellates, cryptophytes and other heterokont algae (Lee 2008; McFadden 2001). Phylogenetic analysis suggested that glaucophytes were the first lineage to diverge, since their plastid-like photosynthetic organelles (cyanelles) could resemble their cyanobacterial ancestor more than those of green and red algae lineages (Lee 2008). To date *C. paradoxa* is the only glaucophyta having its plastid genome sequenced and annotated. A recent example of primary endosymbiosis was found in the amoeba *Paulinella chromatophora*, which harbors in its cytoplasm a cyanobacterium-like symbiont related to the genera *Prochlorococcus* and *Synechococcus*.

Like bacteria, plastids have their own genome and are able to self-replicate by division. The plastid genome or “plastome” is present in multiple copies and consists of generally circular DNA molecules. The first completely sequenced plastomes were those of *Merchantia polymorpha* (Ohyama et al. 1986) and *Nicotiana tabacum* (Shinozaki et al. 1986). At the time we were writing this article, the number of sequenced plastomes was close to 700 and was increasing constantly due to the recent improvements in sequencing techniques.

The average size of a plastome is 150 kbp and the gene number ranges roughly from 80 to 200. The chlorophycean *Floydiella terrestris* owns the largest chloroplast genome known so far, having a size of 521 kbp (Brouard et al. 2010). Plastids belonging to organisms that do not rely anymore on

photosynthesis like the parasitic plant *Epifagus virginiana* lost many genes and got very condensed genomes (70 kbp or less) (dePamphilis and Palmer 1990).

Many plastid genes are involved in gene transcription and gene translation, and despite the consistent reduction of their genome during evolution, plastids conserved a consistent pool of transfer RNA genes. Studying the composition of the tRNA pool of plastids is important to understand which tRNAs are essential for the mRNA translation and why. According to the universal standard genetic code, 61 out of 64 possible triplets specify for the 20 standard amino acids, thus the some amino acid can be transferred by more than one tRNA species (named also isoaccepting tRNA or isoacceptor). Considering the wobble rules described by Crick (Crick 1966) at least 32 tRNA species must be available in order to accomplish the translation process. Actually, recent experiments showed that 25 tRNA isoacceptors (Alkatib et al. 2012a, 2012b) or even less (Koning and Keeling 2006) could be sufficient to read all triplets of the genetic code due to extended wobble rules (superwobbling). According to the superwobbling hypotheses (Lagerkvist 1978; Rogalski et al. 2008), 23 tRNA species (24 including tRNA-fMet) could be sufficient to decode all the codons by fully exploiting the superwobbling mechanisms. In addition, theoretical and experimental assumptions (Rogalski et al. 2008) suggest that plastids hold in their plastomes enough tRNA species to carry out the translation process without importing any tRNAs from the cytosol.

Theoretically, the composition of the tRNA pool in plastids affects not only the translation ability but also the translation efficiency. Indeed, evidence showed that the number of tRNA genes is correlated to the abundance of tRNAs availability (Tuller et al. 2010a, 2010b) thus influencing the translation efficiency (Gingold and Pilpel 2011). Coding genes of plastomes generally encode for proteins involved in important biochemical processes such as photosynthesis and energy metabolism (e.g. *rbcL* for the photosynthesis) and so they need to be translated with high efficiency. Given the above-mentioned statements the gene number of a particular tRNA isoacceptor should be positively related to the frequency of the targeted codon in protein-coding sequences. A correlation between tRNA dosage and codon usage has been found in unicellular species, either prokaryotes or eukaryotes (Coghlan and Wolfe 2000; Dittmar et al. 2004; Ran and Higgs 2010), whereas in many multicellular eukaryotes the results were less clear (Reis et al. 2004). Furthermore, the studies carried out on plants were mainly focused on nuclear genomes or only on a limited number of plastid genomes (Xu et al. 2011).

Assuming that also transfer RNA genes were subjected to a selection, it becomes clear that evolutionary relationships could be inferred by analyzing how many and which tRNA genes were conserved in plastids of different organisms. In the past, phylogenetic analysis using tRNA genes were limited by their short length and their tendency to be transferred horizontally. On the contrary, phylogenetic studies based on the tRNA gene content overcame these

issues and by analyzing gains or losses in more than 500 nuclear genomes it was possible to infer differences between the main life kingdoms (Novoa and Pavon-Eternod 2012).

In this work we analyze hundreds of plastid genome sequences in order to provide an extended view of the tRNA gene content in plastomes. Accomplishing a similar task by hand would have been impossible so we wrote a set of bioinformatics tools to handle the huge amount of tRNA gene annotations available on GenBank. We inspected each plastome sequences to assess if the tRNA gene set was sufficient to translate all the amino acid codons without importing any tRNAs from the cytosol. In a second stage we used the data to infer evolutionary relationships between plastids as previous works had done for nuclear genomes (Novoa and Pavon-Eternod 2012). Rather than using the relative gene frequency (RGF) of tRNA isoacceptor we worked directly with their copy number. Avoiding any data transformation allowed us to use all the information (quantitative and qualitative) about tRNA genes. Then, by using multivariate techniques we explored the differences in term of tRNA gene copy number between cyanobacteria and eukaryotic plastids as well as among plastids belonging to different supergroups of eukaryotic organisms (Archaeplastida, Excavata, Rhizaria, and Chromalveolata). Finally, we retrieved the codon composition of the protein coding genes and we verified if a positive correlation with the tRNA gene copy number actually exists.

RESULTS

tRNA annotations and data enrichment

At the time we wrote this paper about 38% of the plastome sequences submitted to NCBI GenBank showed tRNA annotations lacking information about the anticodon sequence (Fig. 1).

By re-analyzing the tRNA genes of plastid sequences we succeeded in recovering most of the incomplete annotations. In addition, we recovered the tRNA gene annotations of two organisms important in plastid evolution, *Cyanophora paradoxa* (Glaucocystophyceae) and *P. chromatophora* (Rhizaria). For only the remaining 11% of plastomes we were not able to find an anticodon sequence in agreement with the information provided by the existing annotations.

tRNA gene content

Among the plastome sequences collected from the NCBI Genome database, Streptophytes were overrepresented (Fig. 1) reflecting the actual status of the plastid genome sequencing project.

Table S1 shows the plastome sequences of 67 species that lack annotations for at least one standard amino acid. A consistent lack of tRNA genes can be related to a loss of plastid functionality in the organisms living as parasites (e.g. *Rhizanthella gardneri* or *Epifagus virginiana*). If we exclude the import of tRNAs from the cytosol, the missing information can be ascribed to an incomplete annotation procedure.

Plastid genomes harbor a median value of 37 tRNA genes but this does not mean that 37 different tRNA species are present, indeed some tRNA genes are present in more than one copy (Fig. 2). Francis Crick stated that the minimum set of tRNA genes able to decode all the standard codons is composed by 32 tRNA isoacceptors, taking into account the wobble pairing rules. Table S2 in supplementary information lists the plastomes having a number of tRNA isoacceptors greater than or equal to 32. Only few plastids, including the symbiont of *P. chromatophora*, have at least 32 tRNA isoacceptors in their genomes.

Figure 2 shows that the number of tRNA isoacceptors annotated in plastomes is around 28 (median value) but in some organisms it results to be less than 23, which is the minimum amount of tRNA species needed to decode all codons exploiting the superwobbling rules. Table S3 in supplementary information lists the plastomes with a number of tRNA isoacceptors less than or equal to 23. Examples of organisms with such a reduced set of tRNAs are the apicomplexan parasites *Babesia bovis* and *Theileria parva*. Streptophytes show a lower number of tRNA isoacceptors, e.g. *Conopholis americana*, *Cuscuta gronovii*, *Cuscuta obtusiflora*, *Epifagus virginiana*, *Rhizanthella gardneri* and *Selaginella moellendorffii*.

tRNA gene copy number

We calculated the median value of the gene copy number of each tRNA isoacceptor for the plastomes of six eukaryotic supergroups plus the nuclear genome of some Cyanobacteria and the model bacteria *E. coli K12*. In order to have an overall view of the tRNA gene content, we clustered the data in a heatmap (Fig. 3).

In table 1 we reported the tRNA isoacceptors found for every amino acid codon of the standard genetic code. We show that 23 different types of tRNAs are generally conserved in all plastomes and these tRNAs are all capable of wobble (G or U in wobble position) or superwobble (e.g. tRNA-Arg) pairings. Exceptions exist for amino acids decoded by only one codon like Methionine and Tryptophan. In supplementary information we provide the median copy number of every tRNA genes for each group of organisms (Fig 1S).

Multivariate analysis

We used the non-metric multidimensional scaling (NMDS) analysis to obtain a spatial map (Fig. 4) representing the distances between supergroups of organisms, on the basis of the gene copy number of tRNA isoacceptors. The NMDS had a stress value of 0.14 which is very low considering the high number of points (almost 500) and the bi-dimensionality (Sturrock and Rocha 2000). Significant differences ($p<0.0001$) between the groups were confirmed by the PERMANOVA test, along with the homogeneity of variances assessed by the PERMDISP2 test ($p>0.05$).

Table 2 reports the 10 eigenvectors (i.e. tRNA isoacceptor genes) which have the highest correlation ($r^2>0.5$) with the NMDS coordinates and thus highly influence the relative position of every plastid genome on the NMDS plot.

tRNA gene copy number and codon usage

As described before, the tRNA set in plastid genomes is very reduced but retains the ability to decode all the amino acid codons by exploiting the wobble pairings. We calculated the Spearman's correlation coefficient between the tRNA gene copy number and the codon frequency in order to verify whether these characteristics are related or not (Fig. 5). The correlation was estimated by considering either the standard pairing rules alone or together with wobble pairings described by Francis Crick (see the section "Methods" for a detailed description).

Taking into account only the standard pairings the correlation between the tRNA gene copy number and the codon frequency did not result significant (p-value > 0.05) for nearly every genome of both cyanobacteria and plastids. On the contrary, by considering the wobble pairing rules the correlation had high significance (p-value < 0.05) especially for plastid genomes.

Moreover plastid genomes showed a correlation between tRNA pool and codon usage higher than cyanobacteria genomes (Wilcoxon test, p-value < 0.05) either considering or not the wobble pairings.

DISCUSSION

Wobbling and superwobbling rules together with experimental studies (Alkatib et al. 2012b; Rogalski et al. 2008) have shown that only few tRNAs are essential for the mRNA translation process: a single tRNA for 2-codons amino acids; two tRNAs for 4-codons amino acids; three tRNAs for 6-codons amino acids. In addition, using superwobble mechanisms, the number of essential tRNAs becomes lower since a single tRNA having U at the wobble position is sufficient in case of 4-codons amino acids. Exceptions exist for the codon

families of Arginine, Methionine, Isoleucine and Tryptophan (Alkatib et al. 2012b).

Studying the tRNA gene pool of a genome and excluding post-transcriptional modifications, we can imagine at least four possibilities: 1) all the tRNA isoacceptors are present so that each codon is decoded by its corresponding tRNA through the standard pairings; 2) at least 32 tRNA isoacceptors are present and the wobble pairings are used; 3) less than 32 but at least 23 tRNA isoacceptors are present the extended wobble pairings (e.g. "two out of three" rule) are used; 4) less than 23 tRNA isoacceptors are present and the mRNA translation cannot be fully accomplished.

The first scenario generally does not occur, indeed also the large genomes do not use all the 61 tRNA isoacceptors. In fact, the second scenario is the most common for bacterial genomes or nuclear genomes in eukaryotes, and our results showed that the genomes of Bacteria (including all Cyanobacteria) and the chromatophore symbiont of *Paulinella chromatophora*, retain genes for at least 32 tRNA isoacceptors.

Analyzing all the plastid genomes annotated so far, we showed that nearly all of them contain genes for less than 32 tRNA isoacceptors (third scenario) and with few exceptions (e.g. *Bryopsis hypnoides*, *Chlorokybus atmophyticus* and *Pyropia haitanensis*). The total number of tRNA genes in plastid genomes is 37 (median value) but the tRNA isoacceptors are only 28 (median value). Such a reduced set of tRNA isoacceptors can decode all the amino acid codons only by exploiting the wobble and the extended wobble pairing rules (Koning and Keeling 2006).

We found that some plastid genomes have less than 23 tRNA, as described by the fourth scenario. Theoretically, plastomes with a very low number of tRNA isoacceptors cannot fully translate mRNA molecules, in fact they belong to organisms that have lost their original photosynthetic capability, like many parasitic plants (*Conopholis americana*, *Cuscuta gronovii, Cuscuta obtusiflora, Epifagus virginiana* and *Rhizanthella gardneri*). Surprisingly also two species of photosynthetic lycophytes, *Selaginella uncinata* and *S. moellendorffii*, showed a very reduced tRNA pool. *S. uncinata* owns a unusual genome with many gene losses (e.g. rps15, rps16, rpl32 and ycf10) and this may partially explain the very low number of tRNA isoacceptors (Tsuji et al. 2007). Though, the fact that *S. uncinata* is a normally photosynthetic plant with many plastid-encoded photosynthetic genes, suggest that in some cases the import of nuclear tRNAs can occur. Another explanation lies in the post-translational editing mechanisms that could modify the decoding specificity of tRNAs by altering their anticodon sequences (Wolf et al. 2003).

Our data suggest that plastids belonging to different groups of organisms have a specific set of tRNA isoacceptors. In the NMDS map, Glaucocystophyta have an intermediate position between Cyanobacteria and the rest of Archaeplastida. This picture matches the one obtained by previous

phylogenetic analyses conducted on DNA sequences of both plastidial and nuclear markers. Glaucocystophytes represent the currently living organisms more resembling the photosynthetic ancestor derived from the first endosymbiosis event. As a matter of fact, plastids of Glaucocystophytes show intermediate features between Cyanobacteria and chloroplasts and for this reason are called cyanelles. Unfortunately, only one species of Glaucocystophyta, *Cyanophora paradoxa*, has been completely sequenced so far and only two monospecific genera are known.

Table 2 lists the 10 tRNA genes that influence the most our NMDS analysis (Fig. 4) and thus the differences between genomes in term of tRNA gene copy number. In many cases such tRNAs are non-essential due to wobbling or superwobbling mechanisms, and probably during plastid evolution they were lost. In particular, tRNAs having C as the first base of their anticodon (wobble position) like tRNA-Leu (CAA), tRNA-Thr (CGT), tRNA-Arg (CCG), tRNA-Leu (CAG), are non-essential according to the standard wobble rules. Supporting this hypothesis, the tRNA-Arg targeting CGG codons was found to be lost by several clades of lycophytes, ferns, and seed plants (Gao et al. 2010). Furthermore, considering the "two out of three" superwobble rule, also tRNA-Ala (GGC) and tRNA-Val (GAC) might be non-essential. Genes coding tRNA-Asn (GTT) and tRNA-Arg (ACG) were found in all genomes and interestingly only the plastomes of Streptophytes and Cyanobacteria showed more than one copy.

By taking into account only the standard pairings, we initially found that gene copy number of tRNAs and global codon usage of the coding sequences did not correlate (Fig. 5) in both cyanobacteria and plastids. On the contrary, by considering the wobble pairings, a significant correlation was found in nearly all plastids and cyanobacteria. In particular plastid genomes showed a higher correlation respect to cyanobacteria either with or without wobble rules. Since the copy number of tRNA genes influences the abundance of tRNA molecules (Tuller et al. 2010a, 2010b) such correlation could have been maintained during the evolution in order to improve the translation efficiency (Gingold and Pilpel 2011). We suppose it could be a way to compensate for the reduction of the translation efficiency caused by wobble pairings (Rogalski et al. 2008). Our conclusion is that, relying on wobble rules and the correlation between tRNA pool and codon usage, a minimal set of tRNA genes can effectively carry out the mRNA translation process.

Our results showed that plastids could not accomplish the translation process without relying on wobble and superwobble pairing rules, If we exclude the import of nuclear encoded tRNAs. Moreover we showed that there is no significant correlation between gene copy number of tRNA isoacceptors and codon usage in protein-coding genes, while this correlation becomes very high considering wobbling and superwobbling rules. We conclude that, in very reduced genomes like those of plastids, the evolutionary pressure selected

specific species and gene copy number of tRNA isoacceptors in order to maximize the translation efficiency of protein-coding genes.

MATERIAL AND METHODS

Plastid genome sequences

All the plastid genome sequences were retrieved from the NCBI Genome website:

(http://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=2759&opt=plastid).

Accession codes were used to download the corresponding GenBank full records from the NCBI Nucleotide database (RefSeq) and build a dataset of plastid genomes. Cyanobacteria genome sequences were obtained by searching the NCBI Nucleotide database for the following string: txid1117[orgn] AND "complete genome"[title] AND "refseq"[Filter]. The genome sequence of *Escherichia coli* str. K-12 was downloaded from NCBI Nucleotide and included in the dataset.

The organisms analyzed in this paper were grouped as explained in table 5 (adapted from table 4.1 of (Bullerwell 2011) according to their taxonomic data and the current knowledge of the evolutionary relationships among the eukaryotes.

Bioinformatic tools

We wrote computer scripts in Perl programming language to extract all the tRNA and CDS (coding sequence) annotations from the Genbank records. All our computer scripts are freely available at https://github.com/bioinfoplant/tRNA/releases/latest.

The script “tRNA-finder” searches genomes for tRNA genes and retrieves their amino acid specificity and the anticodon sequence (triplet). When the anticodon annotations are not available, tRNA-finder fetches each tRNA gene sequences (merging the exons when introns are present) and uses tRNAscan-SE 1.3.1 (Lowe and Eddy 1997) to identify the anticodon sequence. At the end of this procedure we discarded all the plastomes still lacking such kind of information. A second script, named “CODON-finder”, counts the number of codons in each CDS, using CodonW 1.4.4 (http://codonw.sourceforge.net/). Stop codons were not considered in the analysis because they are recognized by release factors and not by tRNA molecules. Finally, the third script named “CORR-checker” performs the correlation statistics.

Wobble pairing rules

The wobble rules described by Crick (Crick 1966) assume non-standard base pairings between 5' base of tRNA anticodons and 3' base of mRNA codons. Experimental data derived from transplastomic knockout experiments on *Nicotiana tabacum* suggested the existence of the extended wobbling in plastids (Alkatib et al. 2012a, 2012b). Table 3 summarizes the pairing rules considered in our work.

In addition to the standard wobble rules, two types of RNA editing were taken into account: the conversion of cytidine to lysidine (C to L) and the conversion of adenosine to inosine (A to I). In most bacteria and also in plastids the cytidine at the wobble position of tRNA-Ile (CAT) is post-transcriptionally modified to lysidine allowing the decoding of ATA codons (Alkatib et al. 2012a). A second type of RNA editing that can occur in bacteria and plastids is the conversion of adenine to inosine in tRNA-Arg (ACG) (Karcher and Bock 2009; Zhou et al. 2013). Table 4 lists the known special cases and exceptions to the wobble pairing rules that were considered in our analysis.

Correlation coefficient

The data mining process aimed to find the tRNA gene copy number and the codon count of every plastid genomes. In order to account for the wobble rules we used an additive strategy: if a tRNA isoacceptor is able to decode more than one codon, the number of targeted codons were summed up and assigned to it.

Since tRNA gene copy numbers and codon counts do not follow a normal distribution we estimated the correlation between codon preferences and tRNA gene copy number by using the non-parametric Spearman's statistics. Spearman's rank coefficients were calculated through the Statistics::RankCorrelation Perl module. To test the hypothesis that correlation coefficients are significantly greater than zero (one-tailed test) two methods have been implemented in our scripts: Student's t approximation (fast) and random permutations (slow). The first method converts the Spearman's correlation coefficient *rho* to a value distributed approximately as a Student's t-distribution according to the following formula:

$$t = rho\sqrt{\frac{n-2}{1-rho^2}} \qquad (1)$$

The *p*-value can be calculated using the Student's t distribution.

The second method is based on random permutations and our Perl script can use two different strategies. The first strategy does not need any external program and relies on the "shuffle" function provided by List::Util module that generates random permutations of the data. The Spearman's correlation coefficient is calculated for both real and random data. The probability value (P_{perm}) can be obtained by counting the number of times the resulting Spearman

coefficient is equal or greater than the observed Spearman coefficient (one-tailed test):

$$P_{perm} = \frac{\sum_{n=1}^{N_{all}} rho_n \geq rho^*}{N_{all}+1} \qquad (2)$$

where N_{all} is the total number of permutations used, rho_n is the Spearman coefficient calculated for each random set and rho^* is the observed Spearman coefficient.

Finally, the second strategy of random permutation requires the R software. We used the Statistics::R module in order to get access to R commands. The function "spearman_test" included in the R package "coin" allows to calculate the Spearman statistics faster than the Perl module and uses Monte-Carlo resampling.

Multivariate analysis

Multivariate statistics was carried out by using the R software version 3 (R Core Team 2014) and the functions included in Vegan and MASS packages (Oksanen et al. 2013; Ripley et al. 2014).

The tRNA gene copy number of all the analyzed organisms was imported in R as a dataframe. A distance matrix was computed using the "vegdist" function and selecting "jaccard" as method. The quantitative form of the Jaccard distance implemented in Vegan is actually the Ruzicka index and was preferred over the Euclidean distance for its better performances in presence of species containing missing tRNA type (i.e. excludes joint absences). Moreover the Jaccard index compared to the Bray-Curtis distance is fully metric (Anderson et al. 2011; Anderson 2006).

In order to obtain a graphical representation of the distances between every genome sequences an unconstrained ordination was performed by computing a Non-metric Multidimensional Scaling (NMDS) by using the "metaMDS" function included in the Vegan package. Since we had non-ecological data we changed the default options of metaMDS by setting "autotransform=FALSE" and "noshare=FALSE". Multidimensional Scaling helps to visualize distances between samples by producing a low-dimensional spatial map and, in addition the non-metric scaling methods are able to handle non-Euclidean distances. We estimated the "quality" of the non-metric multidimensional scaling using the Shepard diagram and the "stress" value, a normalized version of the sum of squared errors considered acceptable if lower than 0.2 (Sturrock and Rocha 2000).

The statistical significance of the differences between groups was estimated using the "adonis" function included in the Vegan package. The "adonis" function performs a Permutational Multivariate Analysis of Variance (PERMANOVA) (Anderson 2001) avoiding the use of Bonferroni correction.

The function “betadisper” in combination with the “permutest” function allowed the estimation of the homogeneity of variance by using the PERMDISP2 method (Anderson 2006).

FUNDING

This work was supported by the Biology Department of the University of Florence, “Fondi di Ateneo ex 60%”. Funding for open access charge: University of Florence.

REFERENCES

Alkatib S, Fleischmann TT, Scharff LB, Bock R. 2012a. Evolutionary constraints on the plastid tRNA set decoding methionine and isoleucine. *Nucleic Acids Res* **40**: 6713–6724.

Alkatib S, Scharff L, Rogalski M. 2012b. The Contributions of Wobbling and Superwobbling to the Reading of the Genetic Code. *PLoS Genet* **8**: e1003076.

Anderson MJ. 2001. A new method for non-parametric multivariate analysis of variance. *Austral Ecol* **26**: 32–46.

Anderson MJ. 2006. Distance-Based Tests for Homogeneity of Multivariate Dispersions. *Biometrics* **62**: 245–253.

Anderson MJ, Crist TO, Chase JM, Vellend M, Inouye BD, Freestone AL, Sanders NJ, Cornell H V, Comita LS, Davies KF, et al. 2011. Navigating the multiple meanings of β diversity: a roadmap for the practicing ecologist. *Ecol Lett* **14**: 19–28.

Brouard J-S, Otis C, Lemieux C, Turmel M. 2010. The exceptionally large chloroplast genome of the green alga Floydiella terrestris illuminates the evolutionary history of the Chlorophyceae. *Genome Biol Evol* **2**: 240–56.

Bullerwell CE. 2011. *Organelle Genetics: Evolution of Organelle Genomes and Gene Expression*. Springer-Verlag.

Coghlan A, Wolfe KH. 2000. Relationship of codon bias to mRNA concentration and protein length in Saccharomyces cerevisiae. *Yeast* **16**: 1131–1145.

Crick FH. 1966. Codon--anticodon pairing: the wobble hypothesis. *J Mol Biol* **19**: 548–555.

dePamphilis CW, Palmer JD. 1990. Loss of photosynthetic and chlororespiratory genes from the plastid genome of a parasitic flowering plant. *Nature* **348**: 337–9.

Dittmar KA, Mobley EM, Radek AJ, Pan T. 2004. Exploring the Regulation of tRNA Distribution on the Genomic Scale. *J Mol Biol* **337**: 31–47.

Gao L, SU Y, Wang T. 2010. Plastid genome sequencing, comparative genomics, and phylogenomics: current status and prospects. *J Syst Evol* **48**: 77–93.

Gingold H, Pilpel Y. 2011. Determinants of translation efficiency and accuracy. *Mol Syst Biol* **7**.

Karcher D, Bock R. 2009. Identification of the chloroplast adenosine-to-inosine tRNA editing enzyme. *RNA* **15**: 1251–1257.

Koning AP de, Keeling PJ. 2006. The complete plastid genome sequence of the parasitic green alga Helicosporidium sp. is highly reduced and structured. *BMC Biol* **4**: 12.

Lagerkvist U. 1978. “Two out of three”: an alternative method for codon reading. *Proc Natl Acad Sci U S A* **75**: 1759–62.

Lau AOT, McElwain TF, Brayton KA, Knowles DP, Roalson EH. 2009. Babesia bovis: a comprehensive phylogenetic analysis of plastid-encoded genes supports green algal origin of apicoplasts. *Exp Parasitol* **123**: 236–43.

Lee RE. 2008. *Phycology*. Cambridge University Press.

Lowe TM, Eddy SR. 1997. tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. *Nucleic Acids Res* **25**: 0955–964.

McFadden G. 2001. Primary and secondary endosymbiosis and the origin of plastids. *J Phycol* **37**: 951–959.

Novoa E, Pavon-Eternod M. 2012. A role for tRNA modifications in genome structure and codon usage. *Cell* **149**: 202–213.

Ohyama K, Fukuzawa H, Kohchi T, Shirai H, Sano T, Sano S, Umesono K, Shiki Y, Takeuchi M, Chang Z, et al. 1986. Chloroplast gene organization deduced from complete sequence of liverwort Marchantia polymorpha chloroplast DNA. *Nature* **322**: 572–574.

Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O’Hara RB, Simpson GL, Solymos P, Stevens MHH, Wagner H. 2013. vegan: Community Ecology Package.

R Core Team. 2014. R: A Language and Environment for Statistical Computing.

Ran W, Higgs PG. 2010. The Influence of Anticodon–Codon Interactions and Modified Bases on Codon Usage Bias in Bacteria. *Mol Biol Evol* **27**: 2129–2140.

Reis M dos, Savva R, Wernisch L. 2004. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res* **32**: 5036–5044.

Ripley B, Venables B, Bates DM, 1998) KH (partial port ca, 1998) AG (partial port ca, Firth D. 2014. MASS: Support Functions and Datasets for

Venables and Ripley's MASS. *R Packag version*.

Rogalski M, Karcher D, Bock R. 2008. Superwobbling facilitates translation with reduced tRNA sets. *Nat Struct Mol Biol* **15**: 192–198.

Sagan L. 1967. On the origin of mitosing cells. *J Theor Biol* **14**: 255–274.

Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayashida N, Matsubayashi T, Zaita N, Chunwongse J, Obokata J, Yamaguchi-Shinozaki K, et al. 1986. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. *EMBO J* **5**: 2043–2049.

Sturrock K, Rocha J. 2000. A Multidimensional Scaling Stress Evaluation Table. *Field methods* **12**: 49–60.

Tsuji S, Ueda K, Nishiyama T, Hasebe M, Yoshikawa S, Konagaya A, Nishiuchi T, Yamaguchi K. 2007. The chloroplast genome from a lycophyte (microphyllophyte), Selaginella uncinata, has a unique inversion, transpositions and many gene losses. *J Plant Res* **120**: 281–90.

Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, Pan T, Dahan O, Furman I, Pilpel Y. 2010a. An Evolutionarily Conserved Mechanism for Controlling the Efficiency of Protein Translation. *Cell* **141**: 344–354.

Tuller T, Waldman YY, Kupiec M, Ruppin E. 2010b. Translation efficiency is determined by both codon bias and folding energy. *Proc Natl Acad Sci* **107**: 3645–3650.

Wolf PG, Rowe CA, Sinclair RB, Hasebe M. 2003. Complete nucleotide sequence of the chloroplast genome from a leptosporangiate fern, Adiantum capillus-veneris L. *DNA Res* **10**: 59–65.

Xu C, Cai X, Chen Q, Zhou H, Cai Y, Ben A. 2011. Factors Affecting Synonymous Codon Usage Bias in Chloroplast Genome of Oncidium Gower Ramsey. *Evol Bioinform Online* **7**: 271–278.

Zhou W, Karcher D, Bock R. 2013. Importance of adenosine-to-inosine editing adjacent to the anticodon in an Arabidopsis alanine tRNA under environmental stress. *Nucleic Acids Res* **41**: 3362–72.

TABLES

Table 1. Codon coverage by the tRNA isoacceptors found in plastid genomes.Different levels of codon coverage according to the presence of the corresponding tRNA species in plastid genomes (see table 5).

Second Base

First Base	T	C	A	G	Third Base
T	TTT F Phe	TCT S Ser	TAT Y Tyr	TGT C Cys	T
	TTC F Phe	TCC S Ser	**TAC Y Tyr**	**TGC C Cys**	C
	TTA L Leu	**TCA S Ser**	TAA * Ter	TGA * Ter	A
	TTG L Leu	TCG S Ser	TAG * Ter	**TGG W Trp**	G
C	CTT L Leu	CCT P Pro	CAT H His	**CGT R Arg**	T
	CTC L Leu	CCC P Pro	**CAC H His**	CGC R Arg	C
	CTA L Leu	**CCA P Pro**	**CAA Q Gln**	CGA R Arg	A
	CTG L Leu	CCG P Pro	CAG Q Gln	CGG R Arg	G
A	ATT I Ile	ACT T Thr	AAT N Asn	AGT S Ser	T
	ATC I Ile	ACC T Thr	**AAC N Asn**	**AGC S Ser**	C
	ATA I Ile	**ACA T Thr**	**AAA K Lys**	**AGA R Arg**	A
	ATG M Met	ACG T Thr	AAG K Lys	AGG R Arg	G
G	GTT V Val	GCT A Ala	GAT D Asp	GGT G Gly	T
	GTC V Val	GCC A Ala	**GAC D Asp**	GGC G Gly	C
	GTA V Val	**GCA A Ala**	**GAA E Glu**	**GGA G Gly**	A
	GTG V Val	GCG A Ala	GAG E Glu	GGG G Gly	G

tRNA isoacceptor genes :
Conserved (23)
Differentially Conserved (17)
Lost (21)

Table 2. Explanatory power of individual variables.

tRNA (anticodon)	mRNA Codon	r^2	p-value
tRNA-Asn (GTT)	AAC	0.70	0.0001
tRNA-Val (GAC)	GTC	0.69	0.0001
tRNA-Leu (CAA)	TTG	0.67	0.0001
tRNA-Arg (ACG)	CGT	0.61	0.0001
tRNA-Ile (TAT)	ATA	0.59	0.0001
tRNA-Ala (GGC)	GCC	0.58	0.0001
tRNA-Thr (CGT)	ACG	0.56	0.0001
tRNA-Leu (CAG)	CTG	0.51	0.0001
tRNA-Arg (CCG)	CGG	0.51	0.0001

Table 3. Standard wobble pairings. “A, G, T, C” represent IUPAC symbols for the standard bases.

Anticodon first base	**Codon third base**
A	T
G	C, T
T	A, G
C	G

Table 4. Special cases of wobble pairings (tRNA anticodons and corresponding mRNA codons).

tRNA Anticodon	mRNA Codon	Notes
A(I)CG tRNA-Arg	CGT, CGC, CGA	Conversion of Adenine to Inosine
TCA tRNA-STOP	TGA	Does not target TGG (Trp)
TAT tRNA-Ile	ATA	Does not target ATG (Met) to avoid mistranslation
C(L)AT tRNA-Ile	ATA	Conversion of Cytidine to Lysidine

"A, G, T, C" represent IUPAC symbols for the standard bases, "I" stands for the non-standard base Inosine, and "L" stands for Lysidine.

Table 5. Supergroups of eukaryotic organisms.

Supergroup	Groups included	Plastid Origin
ARCHAEPLASTIDA	Streptophyta Chlorophyta Rhodophyta Glaucophyta or Glaucocystophyta *(Cyanophora paradoxa)*	Primary Endosymbiosis
EXCAVATA	Euglenozoa	Secondary Endosymbiosis from GREEN LINE
RHIZARIA	Cercozoa	Secondary Endosymbiosis

	(Paulinella chromatophora)	from GREEN LINE
CHROMALVEOLATA	Alveolata Dinoflagellata Apicomplexa Chromista Cryptophyta Heterokonta (Stramenopiles) Haptophyta	Secondary Endosymbiosis from RED LINE

FIGURES LEGENDS

Fig. 1. Data enrichment. A) Pie chart summarizing the annotation status. We enriched 27% of plastome annotations by finding the missing anticodon sequences. B) Number of plastome sequences in each group before and after the enrichment procedure.

Fig. 2. Number of plastid genomes (y-axis) per number of tRNAs (x-axis). Plastid genomes have a median number of 37 tRNA genes (white bars) but the median number of tRNA isoacceptors is 28 (black bars).

Fig. 3 The heatmap shows for each codon the gene copy number (median) of its corresponding tRNA isoacceptor. Asterisks indicate the tRNA isoacceptor genes that influence the most the position of each plastome on the NMDS map.

Fig. 4. NMDS map derived from the tRNA gene copy number of cyanobacteria and plastid genomes (divided according to supergroups).

Fig. 5. Boxplots of the Spearman's *rho* coefficients obtained correlating the tRNA copy number with the codon frequency in protein coding genes. Y-axis reports the *rho* values while x-axis represents the organism groups. The dotted horizontal line represents the threshold level of significance below which the correlation coefficient is considered significant (p-value < 0,05). (A) Almost all

cyanobacteria (left) and plastid genomes (right) do not show significant correlation taking into account only the standard pairings. (B) Considering also the wobbling most genomes show significant correlations. (C) The extended wobbling (Adenine to Inosine) of tRNA-Arg (ACG) increases even more the number of genomes with significant correlations.