The evolution of the angiosperm plastid genome

Elizabeth Robbins

Wolfson College

Department of Biology

University of Oxford

Thesis submitted for the degree of Doctor of Philosophy

October 2024

Abstract

Plastids originated from the endosymbiosis of a cyanobacterium by a eukaryotic cell approximately 1.5 billion years ago. Since then, their genome has undergone substantial reduction, with present day plastids retaining less than 5% of genes found in their cyanobacterial ancestor. Many of these remaining genes are essential for photosynthesis, including components of the photosystems, NAD(P)H dehydrogenase-like complex, cytochrome b_6f complex and rubisco. While previous comparative analyses have shown that the plastid genome has evolved exceptionally slowly, it is unknown why this slow evolution occurs and whether adaptation has occurred given this constraint. This thesis seeks to address both of these questions through an investigation of the constraints on plastid gene evolution and a search for hallmarks of adaptive evolution that may have been previously overlooked. I reveal that there is significant variation in the rate of molecular evolution between the plastid-encoded genes, with those involved in energy production evolving notably slower than those involved in information processing (i.e., transcription and translation). Additionally, I discover that three key factors – gene position, level of gene expression and the encoded protein's composition – collectively account for >50% of the variation in the rate of sequence evolution. Thus, contrary to expectations, factors unrelated to the molecular function of a gene are the major determinants its evolvability. A novel computational approach, herein named RECUR, was subsequently developed to identify hallmarks of adaptive evolution in the plastid genome. In combination with conventional methods for detecting selection, I identified evidence of adaptive evolution at 7% of the residues in genes encoding components of the photosystems. Through the use of various *in silico* approaches, I predicted the impact of these evolutionary changes on the photosystems and revealed that evolution has repeatedly influenced the interaction between photosystem II and its D1 subunit, potentially reducing the energetic barrier for D1 turnover and enhancing photosystem II repair. In conclusion, this thesis has leveraged millions of years of natural selection experiments to identify plausible engineering targets for enhancing photosynthesis. Moreover, the development of RECUR enables the broader scientific community to explore recurrent evolution, facilitating new discoveries and aiding evolutionary guided protein engineering.

Acknowledgments

I would first like to thank my supervisor, Steve Kelly (a.k.a. the bioinformatics wizard). Without your unwavering support and guidance, this thesis would not have been possible. It was during my rotation project with you that sparked my interest in bioinformatics, and I never would have imagined at the start of my DPhil – when I didn't know the first thing about coding – I would end up being a co-developer of a software package. I have thoroughly enjoyed our scientific discussions and am immensely grateful for the academic freedom you have given me over the last four years – so thank you!

I also wish to thank the past and present members of the Kelly lab for making my time in the lab/office so enjoyable. Special thanks go to Pippa, Jacques, Lisa and VJ, who have provided a constant source of intellectual discussions and support through stressful times. I am also very grateful to Yi for all her help with software development.

My thanks extend to Jane and all the members of the Langdale lab, whose company and camaraderie – whether at cake days or lunches – have made the department feel like home. Special thanks to Sophie for your excellent organisational skills, both academic and social, and for being such a great friend throughout these four years. I'm also thankful to Rox, who has been an all-around fantastic office buddy, and Dana, for sharing many nuggets of wisdom. Additional thanks go to Julie, Lizzie, Nina and Rob without whom nothing would run as smoothly.

I would also like to thank the BBSRC for funding my research and the DTP for their training and support throughout.

I am also immensely grateful for the friendships I formed at Wolfson College which have been an integral part of my DPhil experience. I was incredibly fortunate to have Marija and Christine as flatmates in my first year, and I will cherish the memories of all our breakfasts, boxercise, yoga, film nights and Friday night dinners that got me through lockdowns. I would also like to thank the rest of the FriYaY gang – Piotr, Michael, Dimitri, Joost, Rizwaan, Csabi, and Tugrul – for the countless lunches, dinners, walks, bike rides, amazing trips away and bundles of laughs we have had together over the years. Additional thanks go to the members of the Wolfson College Chess Club and Squash Club for providing much-needed breaks from work.

Finally, I would like to extend my heartfelt thanks to my family. Mum, dad, Aidan and Jengo, your love, support and encouragement have carried me throughout the last four years, and for that, I am forever grateful.

Contents

Chapter 1

General Introduction

1.1 Introduction

Photosynthesis is the process by which solar energy is converted into chemical energy which can then be used to fuel life. Primitive forms of photosynthesis first evolved at least 3.5 billion years ago (Ga) (Noffke, et al. 2013) and have helped sustain life on Earth for ~80% of its 4.5 Ga history (Hazen 2010). It is thought that early photosynthetic organisms carried out a form of photosynthesis known as anoxygenic photosynthesis (Hohmann-Marriott and Blankenship 2011). In this process, electrons are donated from a variety of different sources including sulphur containing molecules, molecular hydrogen or reduced iron (Oliver, et al. 2023). As water is not the source of electrons molecular oxygen is not produced, hence the name *anoxygenic* photosynthesis. Although this form of photosynthesis is evolutionary ancient, it is still found in multiple different bacterial phyla today, including Chlorobi (green sulphur bacteria), Chloroflexi (green filamentous autotrophs), Proteobacteria (purple bacteria), Acidobacteria and Firmicutes (Heliobacteria) (Raymond 2008).

At the heart of photosynthesis are the pigment binding multiprotein complexes called photosystems. The photosystem complexes have two main regions, the light-harvesting antenna and the reaction centre (Croce and van Amerongen 2014). The light-harvesting antenna region is packed with pigments whose role is to absorb and funnel light energy to the reaction centre. Once there, a highly organised array of pigments and other cofactors in the reaction centre use this energy to catalyse charge separation (Nelson and Junge 2015). In turn, this instigates a series of electron transfer reactions that generate an electrochemical gradient that can be used to produce carbohydrates, lipids, amino acids and other essential molecules that fuel metabolism and growth (Johnson 2016).

Although all extant photosystems are thought to be derived from a single ancestral photosystem (Nitschke and Rutherford 1991), over the course of their 3.5 Ga history they have diverged to form two discrete types, known as Type I and Type II (Sánchez-Baracaldo and Cardona 2020). Organisms that carry out anoxygenic photosynthesis can possess either of these two photosystem types. For example, Chlorobi, Firmicutes and Acidobacteria possess a Type I photosystem, whereas Proteobacteria and Chloroflexi have a Type II photosystem (Raymond 2008).

A major landmark in the evolution of life on Earth was the establishment of oxygenic photosynthesis through the combination of both photosystem types in a single cell (Sánchez-Baracaldo and Cardona 2020). This pivotal event is thought to have occurred only once in a bacterial phylum known as the Cyanobacteria (Cavalier-Smith 1982; Bhattacharya and Medlin 1995; Keeling 2010). In contrast to anoxygenic photosynthesis, oxygenic photosynthesis requires the sequential action of both photosystem types, named photosystem I and photosystem II, respectively (Hohmann-Marriott and Blankenship 2011). Photosystem II evolved an oxygen-evolution complex responsible for splitting water, and photosystem I was incorporated into the electron flow pathway (Oliver, et al. 2022).

This landmark event changed the course of evolution of life on Earth. It enabled electrons to be stripped from water to produce molecular oxygen, facilitating the rise in molecular oxygen in the atmosphere known as the Great Oxidation Event, estimated to have occurred 2.3 – 2.4 Ga (Bekker, et al. 2004; Lyons, et al. 2014). This turning point allowed the subsequent evolution of aerobic metabolism and advanced life forms (Nursall 1959; Canfield, et al. 2007).

In this introduction, I will first discuss the rise and spread of oxygenic photosynthesis in the eukaryote lineage which began from the imprisonment of a single cyanobacterium. I will then go on to discuss the diverse roles of the endosymbiont in land plants focussing on photosynthesis. Finally, I will review the evolution of the endosymbiont's genome – the topic of this thesis. Specifically, I will highlight its pivotal role in photosynthesis and discuss current views of its limited evolvability.

1.2 The rise and spread of oxygenic photosynthesis in eukaryotes

1.2.1 The rise: Primary endosymbiosis

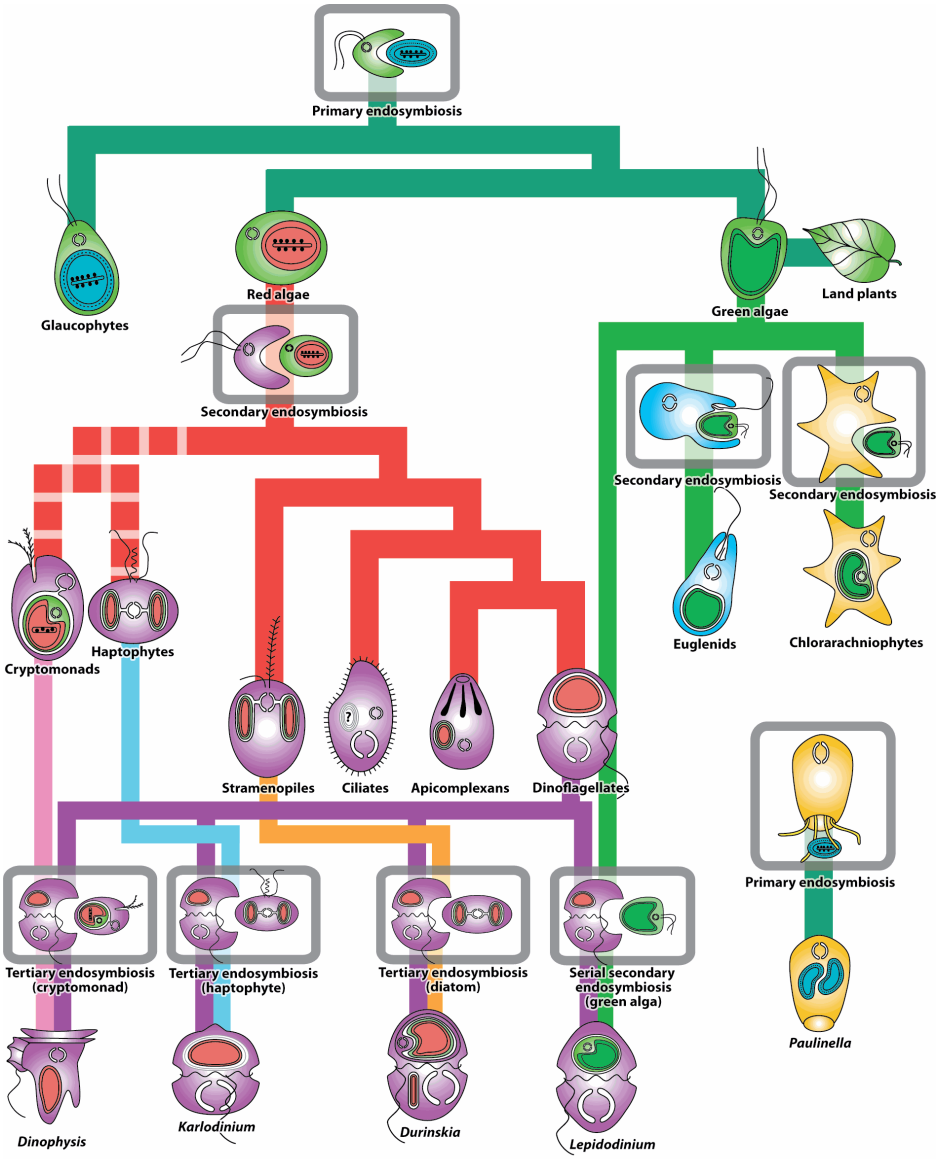
Eukaryotes are one of the three domains of life alongside archaea and bacteria. The name eukaryote derives from the Greek 'eu' meaning *true* and 'karyon' meaning *kernel* and

describes their defining feature – the nucleus, a membrane-bound compartment housing the majority of the cell's DNA. Meanwhile, archaea and bacteria are collectively termed prokaryotes, from the Greek meaning *before kernel* referring to their lack of a nucleus. The increased intracellular complexity of eukaryotes has been proposed to have arisen from a single primary endosymbiosis of an α-proteobacterium by an Asgard archaea approximately ~2 billion years ago (Ga) (Sagan 1967; Spang, et al. 2015; Wang and Wu 2015; Zaremba-Niedzwiedzka, et al. 2017). Over time, the α-proteobacterium endosymbiont evolved into the mitochondria, often termed the 'powerhouse' of the cell. This development enabled more efficient production of adenosine triphosphate (ATP) to power biochemical and cellular processes (Lane and Martin 2010).

A second primary endosymbiosis event occurred later (approximately 1.5 Ga (Hedges, et al. 2004)) whereby a phototrophic bacterium, specifically a cyanobacterium, was engulfed by a eukaryotic cell conferring the ability of photosynthesis to eukaryotes (Sagan 1967; McFadden 2014). Over time this cyanobacterium was domesticated into the organelle known as the plastid which involved several major steps. An early step was the host acquired the ability to tap photosynthate via inserting phosphate translocators, which exchange photosynthate for orthophosphate, into the endosymbiont's inner membrane (Weber, et al. 2006). Next, extensive endosymbiotic gene transfer (discussed in detail later) to the host nucleus rendered the endosymbiont dependent on the host for essential functionals, including primary metabolic pathways, regulation of gene expression and endosymbiont division (McFadden 2014). This necessitated the development of a plastid protein import machinery (translocon at the inner chloroplast membrane and translocon of the outer chloroplast membrane complexes, abbreviated to TIC/TOC, which are embedded in the inner and outer plastid membranes, respectively) to get nuclear encoded proteins into the plastid where they can carry out their function (Richardson and Schnell 2019). Furthermore, obtaining control over plastid replication was a critical step so that the rate can be tuned to avoid being too fast, which would cause plastids to take over the cell, or too slow, which would prevent both daughter cells from inheriting a plastid (McFadden 2014). Collectively, these developments have resulted in the plastid being a semi-autonomous organelle whose metabolism is intricately interwoven with that of the host.

The direct descendants from this first phototrophic eukaryote comprise the monophyletic group known today as the Archaeplastida, containing the glaucophytes, red algae and green algae (Figure 1) (Adl, et al. 2005). The glaucophytes are believed to be the earliest-diverging and consist of a small group of unicellular algae (Keeling 2010). Their plastids contain

chlorophyll *a* and are distinguished by a residual peptidoglycan layer surrounding the plastid (Steiner and Löffelhardt 2002). Meanwhile, red algae (also called Rhodophyta) are a large and diverse lineage with ~7,000 species identified to date, most of which are multicellular organisms (Borg, et al. 2023). Like glaucophytes, red algal plastids also contain chlorophyll *a* but are distinguished by their production of the red protein-pigment complex known as phycoerythrin, from which they get their name (Borg, et al. 2023). Lastly, the green algae (from which plants arose and thus the lineage is also called the Viridiplantae) contain plastids with chlorophyll *a* and *b* (Lewis and McCourt 2004). Unlike both red algae and glaucophytes which contain the ancestral phycobilisome light-harvesting antennae that lie on top of the thylakoid membrane, green algae contain elaborate pigment-filled light harvesting complexes that are lie in the membrane (Liu, et al. 2013; Croce and van Amerongen 2014).



Keeling PJ. 2013.
Annu. Rev. Plant Biol. 64:583–607

Figure 1. The major endosymbiotic events that led to the origin and spread of oxygenic photosynthesis across eukaryotes. At the top is the primary endosymbiosis event, where a cyanobacterium was engulfed by a eukaryotic cell, giving rise to the Archaeplastida lineage. The glaucophytes (middle left) are the earliest diverging lineage, consisting of unicellular algae. The green algae (middle right) gave rise to land plants and have also undergone two independent secondary endosymbiosis events, giving rise to the euglenids and chlorarachniophytes. In the middle is the red algal lineage, whose number of secondary endosymbiosis events is still debated (indicated by red dashed lines). Further endosymbiotic events have occurred in the dinoflagellates, including one serial secondary endosymbiosis event (resulting in *Lepidodinium*) and multiple tertiary endosymbiosis events. In the bottom right is the only other known instance of a second primary endosymbiosis involving a cyanobacterium, which led to *Paulinella chromatophora*. Figure from (Keeling 2013).

The primary endosymbiosis events described above giving rise to the mitochondria and plastid were thought to be unique events. However, recently an independent primary endosymbiosis of a cyanobacterium was discovered in the euglyphid amoeba *Paulinella chromatophora* (Marin, et al. 2005). Molecular clock analysis of the 18S rRNA dates the divergence of the autotrophic *P. chromatophora* from heterotrophic *Paulinella* species approximately 90 – 150 million years ago (Mya) (Delaye, et al. 2016). The cyanobacterial endosymbiont in *P. chromatophora* is referred to as a chromatophore and has been shown to demonstrate important endosymbiont characteristics, including regulated division and transfer of photosynthetic product to the host cell (Kies and Kremer 1979). Further comparative analyses will be required to probe the question as to whether this second acquisition of a cyanobacterium by a eukaryotic cell is on a similar evolutionary trajectory already taken by the plastid.

1.2.2 The spread: Secondary and tertiary endosymbiosis

Although the endosymbiosis of the plastid occurred only once, the plastid spread from the Archaeplastida to other eukaryotic lineages through multiple secondary (and tertiary) endosymbiosis events (Keeling 2013). These processes are thought to proceed via one eukaryotic cell engulfing a second eukaryotic cell that has already undergone primary (or secondary) endosymbiosis. This is followed by the degradation of the eukaryotic features of endosymbiont, presumably after the endosymbiont's nuclear plastid-targeted genes having been transferred to the secondary (or tertiary) host's nucleus. What remains is a four-membrane bound plastid with membranes having derived going out-to-in from the host plasma membrane, the plasma membrane of the endosymbiont (although this can be lost) and the two membranes of the endosymbiont's plastid (Keeling 2010). Occasionally, a drastically reduced remnant of the endosymbiont's nucleus is retained, called a nucleomorph, as observed in cryptomonads and chlorarachniophytes (Archibald 2007). Arguably, these are the most complex cells in existence with four separate genomes found in a single cell - the nucleus, nucleomorph, mitochondria and plastid genomes, respectively.

Secondary and tertiary endosymbiosis has occurred multiple times. In the green lineage, it is widely accepted to have occurred twice, in the euglenids and chlorarachniophytes (Figure 1) (Rogers, et al. 2007). Meanwhile, deciphering the serial endosymbiotic events involving the red algae has been more challenging. It was originally proposed that there was a single secondary endosymbiosis event involving a red algal lineage that gave rise to the Chromalveolata supergroup, including the haptophytes, cryptomonads, stramenopiles,

apicomplexans and dinoflagellates (Cavalier-Smith 2003). However, more recent phylogenomic studies do not support monophyly of these secondary endosymbionts, and it remains an active area of research (Burki, et al. 2016).

In addition to secondary endosymbiosis, additional layers of complexity are required to fully capture the evolution of the plastid in red algae. For example, in multiple lineages photosynthetic capacity has been lost resulting in a highly reduced cryptic plastid such as in ciliates and many apicomplexans (McFadden and Reith 1996; Keeling 2010). Alternatively, although rare, the plastid can be lost all altogether as observed in *Cryptosporidium* (Abrahamsen, et al. 2004; Xu, et al. 2004). Furthermore, tertiary endosymbiosis and serial secondary endosymbiosis have occurred within the dinoflagellates (Figure 1) (Keeling 2013). Tertiary endosymbiosis occurs when a cell engulfs another cell that has already undergone secondary endosymbiosis. Examples include the engulfment of a cryptomonad, haptophyte and diatom by a dinoflagellate in the *Dinophysis*, *Karlodinium* and *Durinskia* genera, respectively (Dodge and Crawford 1969; Schnepf and Elbrächter 1988; Tengs, et al. 2000). Serial secondary endosymbiosis occurs when a cell that has already undergone secondary endosymbiosis engulfs a cell that has undergone primary endosymbiosis. The only known case of serial secondary endosymbiosis is in the *Lepidodinium* genus (Watanabe, et al. 1990; Matsumoto, et al. 2011).

This section has provided an overview of the complicated history of the spread of the plastid in eukaryotes through a myriad of acquisitions, transfers, and losses. From here onwards, this introduction will focus specifically on the evolution of the land plant plastid. In contrast to the tangled web above, the land plant plastid has the relatively straight forward history of a single primary endosymbiosis event followed by diversification of function and form during the radiation of the Viridiplantae.

1.3 The role of plastids in higher plants

1.3.1 Plastid diversity

The plastid is a ubiquitous organelle found in plant cells and is a hub for biosynthetic reactions involved in primary and secondary metabolism. However, plastids have additional functions depending on the type and developmental stage of the tissue (Choi, et al. 2021). To perform these various functions developmental and environmental cues induce plastid differentiation which is detected through changes in colour, morphology and ultrastructure (Wise 2006). From this ability to transform is how the plastid got its name, from the Greek ‘plastikos’

meaning *capable of being moulded*. The types of plastids (proplastid, etioplast, chloroplast, chromoplast, leucoplast and gerontoplast) and their interconversions are shown in Figure 2.

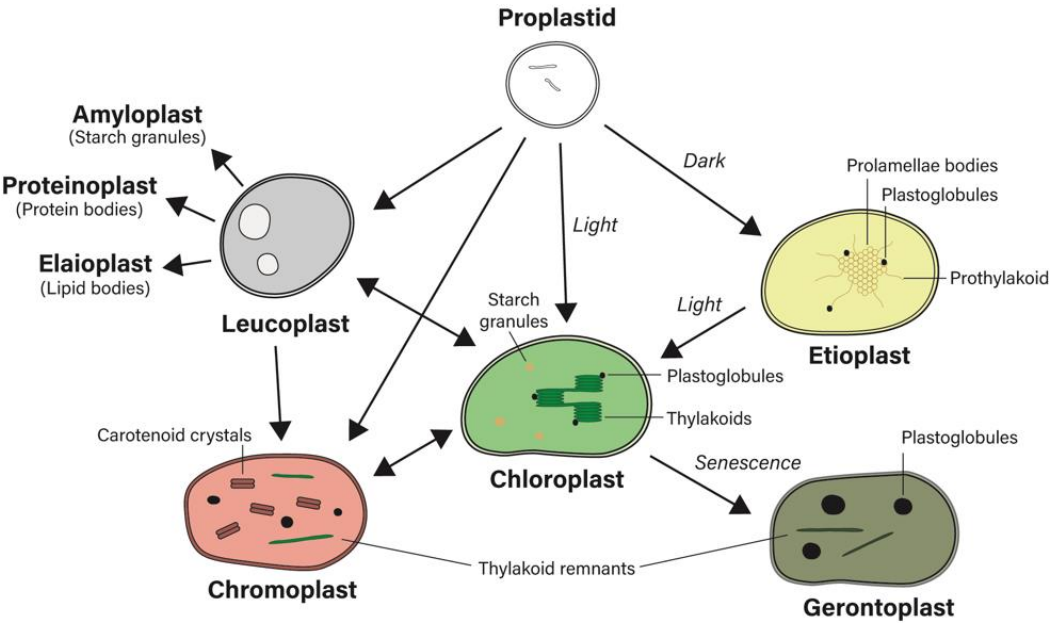


Figure 2. Plastid types and their interconversions. The process by which various plastids such as proplastids, etioplasts, chromoplasts and leucoplasts transition into chloroplasts is referred to as 'greening' and is primarily triggered by light. In senescing leaves, chloroplasts transition into gerontoplasts. Meanwhile, developmental signals initiate the plastid transition to leucoplasts (and the amyloplast, proteinoplasts and elaioplast subtypes) and chromoplasts. Interconversions between plastid types are shown with arrows. Figure adapted from (Choi, et al. 2021).

Proplastids are the undifferentiated form from which all other types of plastids are derived. They are ~1 um in diameter and found in dividing meristem and reproductive tissue (Pyke 2007). In photosynthetic tissue the proplastid primarily differentiates into the chloroplast, the most well-known form of plastid for its role in being the site of photosynthesis. There are ~80 – 120 chloroplasts per mesophyll cell (Crumpton-Taylor, et al. 2011), although this number can vary, and each are ~4-6 um in diameter. Their name comes from the Greek 'chloros' meaning *green* resulting from their high chlorophyll content (Choi, et al. 2021). However, chloroplasts also contain high levels of carotenoid pigments which expand the wavelengths of harvestable light and play an important role in photoprotection (Sun, et al. 2018). Furthermore, chloroplasts contain highly specialised inner membrane structures called thylakoids that encase the thylakoid lumen. The thylakoid membrane has stacked regions called grana which are

connected by flat stromal lamellae (Ostermeier, et al. 2024). Within this membrane structure the photosynthetic apparatus is non-uniformly distributed with photosystem II (Type II) and photosystem I (Type I) being primarily situated in the grana and lamellae regions, respectively (Andersson and Anderson 1980).

If photosynthetic tissue develops in dark conditions, such as a dark grown seedling, chloroplast differentiation is arrested forming etioplasts (Armarego-Marriott, et al. 2020). Etioplasts are equipped for the rapid conversion to the chloroplast with prothylakoid membranes and para-crystalline prolamellar bodies which contain high levels of lipid and NADPH: protochlorophyllide oxidase (POR) enzyme which catalyses the light-dependent last step of chlorophyll biosynthesis (Gabruk and Mysliwa-Kurdziel 2015; Kowalewska, et al. 2016). The lack of chlorophyll is how etioplasts got their name from the French ‘étioler’ meaning to *bleach*. Meanwhile, during leaf senescence chloroplasts terminally differentiate to gerontoplasts (‘geront’ meaning *old man* in Greek) whose primary role is to safely dismantle the photosynthetic machinery and recycle their component atoms (including nitrogen, phosphate and carbon) back to the plant (Zentgraf, et al. 2022).

In tissues with storage functions, such as those in the seed and roots, leucoplasts are present (Choi, et al. 2021). The lack of pigment production in leucoplasts is where they get their name from the Greek ‘leukos’ meaning white. Leucocytes can be subdivided by the macromolecule they store. Specifically, amyloplasts store starch, proteinoplasts store protein and elaioplasts store lipids (Choi, et al. 2021). Lastly, chromoplasts are non-photosynthetic colourful plastids resulting from their high carotenoid content and are found primarily in fruits and flowers, although they are also present in colourful roots such as carrots (Marano, et al. 1993; Rodriguez-Concepcion and Stange 2013). Their main function is to attract insects and animals for pollination and seed dispersal.

In summary, land plants have evolved a high degree of plastid functional diversity to aid them throughout their life cycle. This includes managing resources, optimising growth, adapting to fluctuating environments, and increasing chances of reproductive success.

1.3.2 The role of the chloroplast: Photosynthesis

Photosynthesis is carried out in the chloroplast and has two main stages, the light dependent reactions and the light-independent (or dark) reactions (Johnson 2016). The light-dependent reactions take place in the thylakoid membrane and use light energy to catalyse a series of redox reactions (outlined in detail below) that generate ATP and NADPH and release

molecular oxygen as a byproduct (Figure 3). The ATP and NADPH is then used to fuel the Calvin Benson cycle, a series of enzymatic steps where carbon dioxide is fixed into glyceraldehyde-3-phosphate (G3P) which can then be used to synthesize all other compounds required for growth.

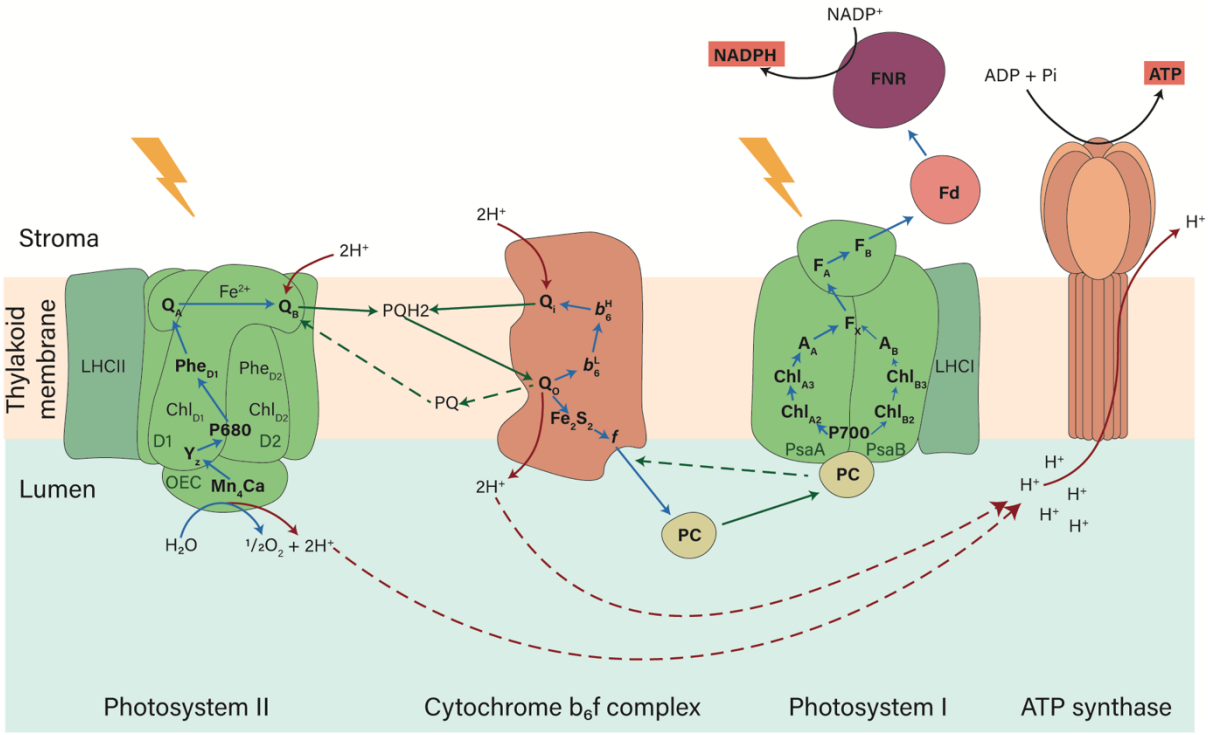


Figure 3. The light dependent reactions in the thylakoid membrane. Electrons derived from the splitting of water by photosystem II are shuttled to $NADP^+$ through the electron transfer chain generating a proton-motive force used to generate ATP. Lightning bolts indicate light requirement in photochemical steps. Blue arrows indicate electron movement, red arrows indicate proton movement and green arrows indicate molecule/protein movement. Abbreviations: *b*, haem; Chl, chlorophyll *a*; *f*, cytochrome *f*; FeS/F iron-sulphur clusters, Fd, ferredoxin; FNR, ferredoxin-NADP reductase; LHC, light harvesting complex; PC, plastocyanin; Pheo, pheophytin *a*; Q, quinone binding site, Y, tyrosine residue.

1.3.2.1 The light dependent reactions

The first step of the light dependent reactions is catalysed by photosystem II (Figure 3). Light energy is absorbed by accessory pigments in the light harvesting complex II (LHCII) and funnelled via resonance energy transfer to the photosystem II reaction centre (Croce and van Amerongen 2014). Here, the excitation energy excites the P680 chlorophyll pair bound at the luminal side of the interface between the pseudo-symmetric D1-D2 heterodimer to form P680* (Müh and Zouni 2020). An electron is then relinquished from P680* and migrates through a molecular wire that terminates on the stromal side of the membrane (Ferreira, et al. 2004; Müh

and Zouni 2020). This molecular wire is composed of a pheophytin *a* (Phe), a tightly bound plastoquinone (Q_A) and a mobile plastoquinone (Q_B). For efficient electron transfer between Q_A and Q_B a non-haem iron is required (Fe^{2+}) and is coordinated by a bicarbonate ion and two histidine residues, one from D1 and one from D2, respectively. Two electrons and two protons, the latter acquired from the stroma, are required to fully reduce plastoquinone to plastoquinol which has a lower affinity for the Q_B binding site and thus dissociates into the milieu of the thylakoid membrane. Meanwhile, on the luminal side of the membrane the strongly oxidising $P680^+$ energetically drives the splitting of water into dioxygen, protons and electrons (Johnson 2016). This reaction is catalysed by the oxygen evolving complex which contains a Mn_4CaO_5 cluster (Li, et al. 2024). A tyrosine residue at position 161 in the D1 protein (*Arabidopsis thaliana* accession: NP_051039) acts as an intermediary sequentially passing the electrons from the oxygen evolving complex to $P680^+$ after each excitation event (Barry and Babcock 1987). Notably, the electron movement described above results in two protons being removed from the stroma and two protons being released in the lumen per plastoquinol molecule produced (Johnson 2016).

Diffusing plastoquinol in the thylakoid membrane can bind the Q_O site located toward the luminal side of the thylakoid membrane in the cytochrome b_6f complex (Malone, et al. 2019). The two-electron oxidation of plastoquinol liberates two protons that are released into the luminal space and two electrons that follow a bifurcated path within the cytochrome b_6f complex (Figure 3) (Mitchell 1976). The first electron flows into the high potential path, constituting the Rieske iron-sulphur cluster and cytochrome *f*, before being donated to plastocyanin bound at the luminal surface which is then released and replaced with an oxidised plastocyanin (Malone, et al. 2019). The second electron flows into a low potential path constituted of two haems prior to being donated to an oxidised plastoquinone at the Q_i site at the stromal side of the thylakoid membrane (Malone, et al. 2019). After two sequential reductions, reduced plastoquinone at the Q_i site acquires two protons from the stroma and dissociates back into the membrane where it can rebind at the Q_O site of cytochrome b_6f complex. This mechanism allows for two protons to be translocated to the lumen from the stroma per electron transferred to plastocyanin.

The next step is the transfer of electrons from plastocyanin in the lumen to ferredoxin in the stroma by photosystem I. Analogous to photosystem II, this reaction requires the harvesting of light energy by antennae regions (Croce and van Amerongen 2014). This excitation energy is then funnelled to the P700 chlorophyll pair at the interface of the PsaA-PsaB heterodimeric core of the photosystem I reaction centre (Mazor, et al. 2017). Excitation of the P700 to P700* is followed by charge separation which can occur along (although not evenly) two quasi-

symmetrical molecular wires converging at an iron-sulphur cluster on the stromal side of the PsaA-PsaB interface (Klukas, et al. 1999). The electron then migrates via two more iron-sulphur clusters, FeS_A and FeS_B, before being passed to the soluble ferredoxin protein in the stroma (Guergova-Kuras, et al. 2001). The $P700^+$ on the luminal side of the membrane is reduced by an electron donated from a reduced plastocyanin generated by cytochrome b_6f complex. The final step of electron flow is then carried out in the stroma where the sequential reduction of $NADP^+$ by ferredoxin-NADP reductase (FNR) using reduced ferredoxin as an electron donor is used to produce NADPH, a reducing agent required for a vast array of cellular processes including the Calvin Benson cycle (Johnson 2016).

The movement of electrons from water to $NADP^+$ described above produces an electrochemical proton concentration gradient across the thylakoid membrane where the lumen is acidified. This concentration gradient is used by the chloroplast-ATP synthase to catalyse the formation of ATP from adenosine diphosphate (ADP) and inorganic phosphate (Boyer 1997). To do this, protons flow down their concentration via a channel in the integral membrane region (cF_0) of the chloroplast-ATP synthase inducing its mechanical rotation. This kinetic energy is transmitted up the stalk and into the catalytic head of the ATP synthase (cF_1) inducing rotatory catalysis (Hahn, et al. 2018). The translocation of fourteen protons through the ATP synthase is required for one revolution of the c-ring in the cF_0 domain forming three molecules of ATP, and thus 4.67 protons is required to produce one ATP molecule (Hahn, et al. 2018).

The linear electron flow pathway shown in Figure 3 does not generate enough ATP to power the Calvin Benson cycle (Allen 2002). To account for this disparity an alternative electron pathway is employed called cyclic electron flow (Allen 2003a). Here, electrons are redirected from ferredoxin back to plastoquinone enabling a greater number of protons to be transferred into the thylakoid lumen per electron donated from water. Cyclic electron flow occurs through two pathways, the major pathway is proton gradient regulation 5 (PGR5)/PRG5-like photosynthetic phenotype 1 (PGRL1) dependent (DalCorso, et al. 2008) and the minor pathway is the NAD(P)H dehydrogenase-like dependent (Joët, et al. 2001). In addition to adequately fuelling the Calvin Benson cycle the ability to tune the ATP/NADPH ratio allows plants to better cope with a wide variety of stress (Huang, et al. 2010; Joliot and Johnson 2011).

1.3.2.2 Dealing with excess light

The photosynthetic apparatus has a maximum capacity for the rate it can convert light energy to chemical energy (Guidi, et al. 2019). This capacity is determined by how quickly

downstream metabolism can utilise the generated NADPH. If the rate of NADPH production is much greater than its subsequent oxidation, i.e., the NADPH/$NADP^+$ ratio gets larger, the insufficient pool size of the final electron acceptor $NADP^+$ causes the over-reduction of the photosynthetic apparatus (Gururani, et al. 2015). Over-reduction can prevent quick transfer of excitation energy between cofactors and lead to reactive oxygen species (ROS) production which can then irreversibly damage proteins, membranes or nucleic acids (Juan, et al. 2021). The major generation sites of ROS are the photosystem reaction centres and their subsequent damage can reduce the photosynthetic capacity in a process known as photoinhibition (Moustakas 2022).

In the photosystem II reaction centre, a saturated reduced plastoquinone pool results in the formation of triplet state chlorophyll ($^3Chl^*$) which can react with ground state molecular oxygen to form singlet oxygen ($^1O_2$) (Pospíšil 2012; Li and Kim 2022). The primary site of $^1O_2$ formation is in the D1 protein and thus results in D1 being a hub of oxidative damage (Mattoo, et al. 1981; Aro, et al. 1993). However, efficient photosystem II repair processes exist such that as long as the rate of damage does not exceed the rate of repair, photoinhibition does not occur. This repair pathway involves photosystem II migration to non-appressed regions of the thylakoid membrane, proteolytic digestion of damaged D1 and its replacement with a newly synthesized functional copy (Järvi, et al. 2015). Meanwhile, when the acceptor side of photosystem I is highly reduced the iron-sulphur clusters donate electrons to alternative electron acceptors (Asada 2006). When this alternative electron acceptor is molecular oxygen the superoxide anion (O_2^-) is produced, which can then react with the iron-sulphur clusters resulting in their inactivation (Sonoike, et al. 1995; Erling Tjus, et al. 1998). Unlike photosystem II, repair of photosystem I is extremely slow and can take days to weeks to fully recover photosynthetic capacity (Bernhard Teicher, et al. 2000; Kudoh and Sonoike 2002).

Plants have evolved intricate mechanisms to cope with excess light and prevent damage to the photosynthetic apparatus. A primary mechanism is non-photochemical quenching (NPQ) where excess light can be dissipated as heat (Murchie and Ruban 2020). NPQ is triggered by over-acidification of the thylakoid lumen which signals the stress response to excess light. This signal can be further amplified by activating the cyclic electron pathway to move electrons away from the acceptor side of photosystem I and prevent its photoinhibition. The acidification of the lumen is sensed by the integral membrane protein PsbS through protonation of two lumen-exposed glutamate residues (Li, et al. 2000; Li, et al. 2002). The protonated form of PsbS undergoes conformational changes allowing it to act as a seeding site for LHCII aggregation causing dissociation from the photosystem II reaction centre (Correa-Galvis, et al. 2016). Thus, the LHCII transitions from a light-harvesting to a dissipative state.

Furthermore, the acidification of the thylakoid lumen activates the violaxanthin de-epoxidase enzyme to convert violaxanthin to zeaxanthin via an antheraxanthin intermediate (Yamamoto, et al. 1962; Demmig-Adams 1990). These de-epoxidated xanthophyll pigments aid thermal dissipation of excess light energy absorbed by light-harvesting proteins (Gilmore and Yamamoto 1993). It is important that induction and removal of NPQ is a highly dynamic process and tuned such that the trade-off between photosynthetic efficiency and prevention of photoinhibition is optimal.

1.4 The plastid genome

1.4.1 Architecture of the plastid genome

The land plant plastid contains its own genome which is generally thought to be a circular molecule (Figure 4), although linear and branched forms have been observed (Deng, et al. 1989; Oldenburg and Bendich 2004; Oldenburg and Bendich 2016). The size of the genome is ~120 – 180 kb in size, although plant species with significant deviations have been identified (Kolodner and Tewari 1975a; Chumley, et al. 2006; Daniell, et al. 2016). Although small in size, the plastid houses multiple copies of its genome and can make up significant proportions of cellular DNA. For example, in light-grown spinach leaves plastid DNA can reach 23% of total cellular DNA resulting from ~900 genome copies in each of the ~65 plastids (Lawrence and Possingham 1986; Bendich 1987). However, genome copy number is highly variable depending on tissue type and developmental stage (Oldenburg and Bendich 2015).

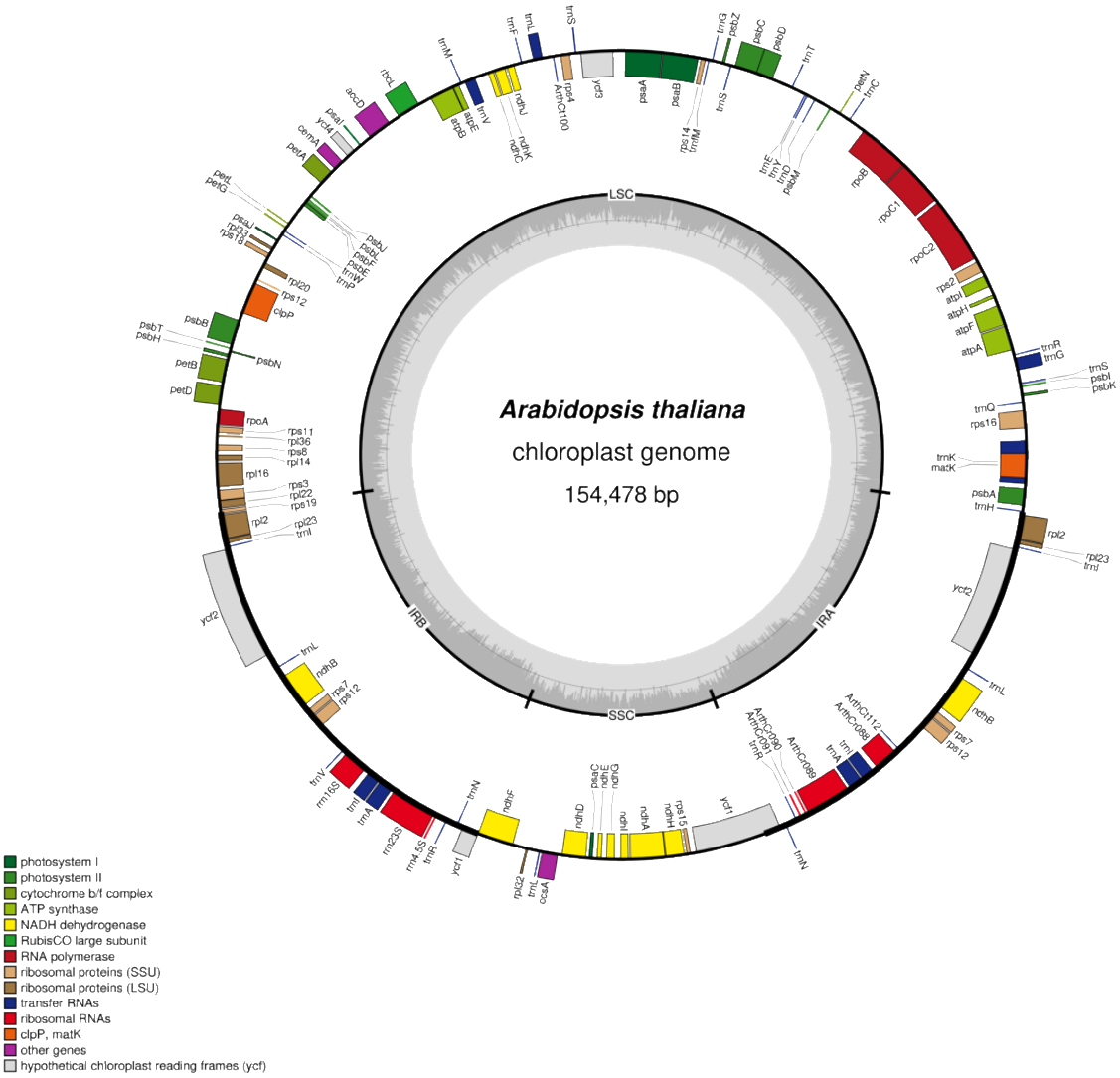


Figure 4. Map of the *Arabidopsis thaliana* plastid genome (accession: NC_000932). The inner concentric circle highlights the quadripartite structure consisting of the large single copy region (LSC) and small single copy region (SSC) separated by two inverted repeat regions (IRA and IRB). The GC content is shown in dark grey shading and the complementary AT content in light grey. The outer circle shows gene positions which are coloured by their function. Figure generated using OGDRAW (Greiner, et al. 2019).

Comparison of plastid genomes between different land plant species has uncovered a highly conserved quadripartite structure made up of a large single copy region and a small single copy region separated by two inverted repeat regions (Figure 4) (Kolodner and Tewari 1979; Palmer 1985; Daniell, et al. 2016). In most plant species these inverted repeat regions are identical (or near identical) and are ~20 – 25 kb in size (Palmer 1985; Mohanta, et al. 2020). Expansions and contractions of the inverted repeats are a reasonably common occurrence

(Wang, et al. 2008; Zhu, et al. 2016). More rarely, the inverted repeat can be lost altogether such as in the Fabaceae clade of legumes (Palmer, et al. 1987). After loss, subsequent large structural rearrangements to the plastid genome ensue which led to belief that inverted repeats protect plastid genome stability (Palmer and Thompson 1982). Furthermore, the rate of nucleotide substitution of genes in the inverted repeats is three times lower than in the single copy regions (Wolfe, et al. 1987; Perry and Wolfe 2002; Zhu, et al. 2016). However, loss of the inverted repeat results in elevation of the rate of nucleotide substitution to that found in the single copy regions (Perry and Wolfe 2002). This is thought to be the result of increased error correction in the inverted repeats, specifically through homologous recombination mediated repair and gene conversion mechanisms (Birky and Walsh 1992; Khakhlova and Bock 2006). Although these mechanisms can occur in single copy regions between plastid genome molecules, the inverted repeat has twice as much template and thus is likely to occur more frequently resulting in lower rates of nucleotide substitution over time.

1.4.2 Gene content

The gene content of the plastid genome in land plants is also highly conserved and consists of ~120 genes (~80 of which are protein coding) and are predominantly involved in photosynthesis, transcription or translation (Table 1) (Wicke, et al. 2011). Genes are organised into operons that are transcribed as polycistronic mRNA (Stern, et al. 2010). The photosynthetic genes include subunits of photosystem I, photosystem II, cytochrome b_6f complex, NAD(P)H dehydrogenase-like complex, ATP synthase and the large subunit of rubisco. Transcription related genes include the subunits of the plastid-encoded RNA polymerase (PEP). Translation related genes include the plastid-encoded ribosomal proteins and various structured RNAs (ribosomal RNA and transfer RNA). A handful of genes do not fit into these categories and have functions in photosystem assembly, RNA maturation, protein degradation and fatty acid synthesis (Table 1).

Table 1. Highly conserved genes in land plant plastid genomes

Function	Functional category	Genes
Photosynthesis	Photosystem I	*psaA, psaB, psaC, psaI, psaJ*
	Photosystem II	*psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ*
	Cytochrome b6f complex	*petA, petB, petD, petG, petL, petN*
	ATP synthase	*atpA, atpB, atpE, atpF, atpH, atpI*
	NAD(P)H dehydrogenase-like complex	*ndhA, ndhB, ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK*
	Rubisco	*rbcL*
Transcription	RNA polymerase	*rpoA, rpoB, rpoC1, rpoC2*
Translation	Ribosomal small subunit	*rps2, rps3, rps4, rps7, rps8, rps11, rps12, rps14, rps15, rps16, rps18, rps19*
	Ribosomal large subunit	*rpl2, rpl14, rpl16, rpl20, rpl22, rpl23, rpl32, rpl33, rpl36*
	Ribosomal RNAS	*rrn16, rrn23, rrn4.5, rrn5*
	Transfer RNAs	Typically around 30 genes
Miscellaneous	Miscellaneous	*accd* (fatty acid synthesis), *ccsA* (cytochrome c assembly), *clpP* (protein degradation), *cemA*, *matK* (intron maturation), *ycf3* and *ycf4* (both photosystem I assembly factors), *ycf1* and *ycf2* (function unknown)

1.4.3 Plastid genome replication

Multiple mechanisms for the replication of the plastid genome have been proposed. Initially, the discovery of bacterial-like origins of replication in the inverted repeat regions led to the belief that replication proceeded by double displacement loop and rolling circle mechanisms (Kolodner and Tewari 1975b; Kolodner and Tewari 1975c; Kunnimalaiyaan and Nielsen 1997). However, deletion of the origins of replication in *Nicotiana tabacum* showed that they were not essential for plastid genome replication suggesting additional mechanisms of replication exist (Scharff and Koop 2006). Since then, the discovery of branched plastid genome structures and the requirement of plastid targeted RecA to maintain plastid genome copy number have led to the proposal of recombination mediated replicative mechanisms (Bendich 2004; Oldenburg and Bendich 2004; Rowan, et al. 2010). Together, these findings suggest multiple

mechanisms of plastid genome replication exist, but further research is required to understand how these mechanisms are coordinated.

1.4.4 Evolution of the plastid genome

1.4.4.1 Genome reduction

One of the most striking features of plastid evolution is the dramatic reduction of its genome content. The closest identified relative to the ancestral endosymbiont is *Gloeomargarita lithophora* that has a genome size of 3 Mb and contains ~3000 genes (Ponce-Toledo, et al. 2017). However, extensive gene loss and transfer to the nuclear genome via endosymbiotic gene transfer has resulted in land plant plastids having lost ~95% of genes from the ancestral cyanobacterial genome (Figure 5) (Timmis, et al. 2004). More extreme genome reduction has occurred in parasitic plants due to loss or pseudogenisation of the non-essential photosynthetic machinery (Krause 2008).

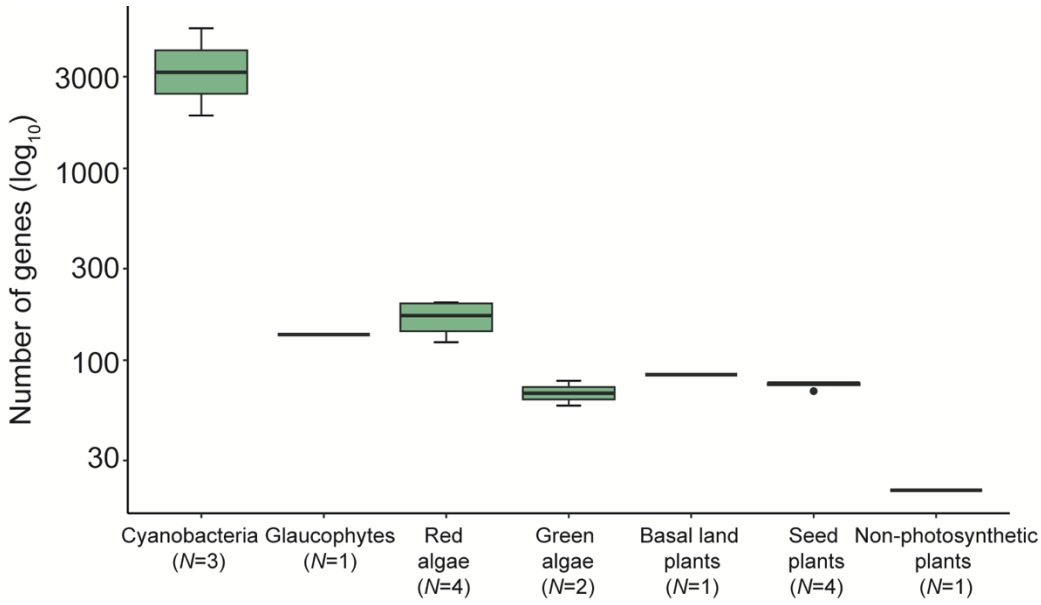


Figure 5. Number of genes encoded in cyanobacterial genomes and in the plastid genome of several groups of the Archaeplastida lineage. *N* indicates the sample size of each group. Data from (Timmis, et al. 2004).

The favoured model of endosymbiotic gene transfer is the uptake of free plastid DNA into the host nucleus after organelle rupture (Timmis, et al. 2004). Here, plastid derived DNA can be integrated into the host genome at chromosomal breaks via the non-homologous end-joining repair pathway (Wang and Timmis 2013). Endosymbiotic gene transfer occurs frequently (and continues to do so) and has had large impacts on host nuclear genome composition (Stern and Lonsdale 1982; Ayliffe, et al. 1998; Huang, et al. 2003; Stegemann, et al. 2003). For

example, 18% of the *A. thaliana* nuclear protein coding genes have endosymbiont origin (Martin, et al. 2002). Furthermore, plastid DNA insertions can be large as found in rice with a 130 kb plastid-derived DNA chunk inserted into the nuclear genome (The Rice Chromosome 10 Sequencing 2003). Endosymbiotic gene transfer has been observed to occur more frequently in land plants, which have multiple plastids per cell, compared to algae which contain a single plastid (Lister, et al. 2003). This led to the development of the 'limited transfer hypothesis' which states that cells with more endosymbionts are more likely to successfully transfer genes to the nucleus (Lister, et al. 2003; Smith, et al. 2011). In part, this is simply because there are more potential DNA donors. However, host cell survival is more likely after organelle rupture if the cell contains multiple endosymbionts compared to a catastrophic rupture in a cell that contains a single endosymbiont.

Once transferred to the nucleus, genes can follow several evolutionary paths (Martin and Herrmann 1998). The simplest and probably most frequent fate is that the transferred gene cannot be expressed, or a useful function for the expressed gene is not identified prior to the rapid accumulation of mutations, leading to its decay. Alternatively, if the successfully expressed donated gene is better than existing nuclear homologues it can replace their function, as seen extensively in the glycolytic pathway (Martin and Herrmann 1998). Furthermore, donated genes may evolve new functions through genetic innovation as seen in the TIC/TOC chloroplast protein import machinery (Bölter, et al. 1998; Reumann and Keegstra 1999). Finally, the donated gene might acquire an amino terminal transit peptide that targets the gene product to the plastid compartment where it can carry out its original function. This alleviates the selection pressure acting on the plastid gene which can lead to gene loss or pseudogenization (Millen, et al. 2001; Park, et al. 2020; Shrestha, et al. 2020).

Several explanations as for why organelle gene transfer to the nucleus is beneficial have been proposed. One hypothesis is that endosymbiotic gene transfer increases host control over the endosymbiont and can aid its integration into cellular metabolism (Martin and Herrmann 1998). Another hypothesis is that endosymbiotic gene transfer removes organelle genes from a hostile environment where ROS induced damage could lead to a high mutational burden (Allen 2003b). This is exacerbated by the uniparental inheritance of the plastid genome preventing the removal of deleterious mutations through recombination that would occur in sexual reproduction. This phenomenon is known as Muller's ratchet (Muller 1964). However, moving organelle genes to the nucleus, where sexual recombination does occur, would allow escape from Muller's ratchet. Paradoxically, nucleotide substitution rates of plant organelles are significantly lower than the nuclear genome (Wolfe, et al. 1987; Smith 2015). Thus, a compensatory mechanism preventing the accumulation of mutations must exist. This

mechanism is proposed to be gene conversion enabled by the high copy number of the plastid genome and presence of homologous recombination machinery (Birky and Walsh 1992; Khakhlova and Bock 2006). Lastly, it was recently proposed that gene transfer and re-import of proteins into organelles is more energetically efficient for the cell than maintaining the high copy numbers of organelle genomes (Kelly 2021).

Given the benefits of organelle gene transfer to the nucleus, the question arises as to why some genes have remained in the plastid. A leading theory is the Co-location for Redox Regulation (CoRR) hypothesis (Allen 1993, 2015; Allen 2017). This states that there is a selective advantage for rapid and direct redox-mediated regulation of gene expression of the redox-sensitive components of photosynthesis. Thus, even if transfer of these redox-sensitive genes to the nucleus occurred by chance, the plastid located gene copy remains beneficial and would be selected over the nuclear copy. A second theory for organelle gene retainment is the hydrophobicity hypothesis (Johnston and Williams 2016). This poses that the integral membrane proteins encoded by the plastid genome may encounter difficulties being retargeted back to the plastid compartment. For example, they may aggregate in the cytosol or be mistargeted to the endoplasmic reticulum (Björkholm, et al. 2017). Protein translation in close proximity to their target membrane may increase the chance of proper folding and thus complex assembly. Furthermore, the high copy number of the plastid genome may aid the production of plastid genes in the quantity they are required (Bendich 1987). For example, rubisco is the most abundant enzyme on the planet and makes up to 50% of the soluble protein found in the leaf (Carmo-Silva, et al. 2015). Thus, while there are the benefits of ploidy, recombination and purifying selection for genes being housed in the nucleus, factors including rapid regulation of gene expression, successful protein folding and adequate expression level underscore the functional importance of maintaining a small set of genes in the plastid genome.

1.4.4.2 The sudden appearance of RNA editing

RNA editing is the post-transcriptional conversion of a nucleotide base that occurs in an RNA transcript. Intriguingly, RNA editing is not present in algae but found in all land plants (with the exception of Marchantiales) suggesting it evolved during the transition from water to land (Groth-Malonek, et al. 2007; Small, et al. 2020). In plants, RNA editing is limited to the organelles and does not occur in the nucleus and is dominated by cytosine to uracil conversions. Furthermore, edit sites are concentrated in coding regions and correct missense mutations to restore highly conserved amino acids (Takenaka, et al. 2013). Thus, RNA editing may provide a mechanism to overcome Muller's ratchet. For example, the RNA edit that gives rise to a serine to phenylalanine substitution at position 26 in PsbF prevents defects in

photosystem II function (Bock, et al. 1994). Interestingly, despite their rapid emergence RNA edit sites are seemingly disappearing from the plastid genome. In basal land plants hundreds, and occasionally thousands, of RNA edits have been identified in the plastid genome (Oldenkott, et al. 2014). However, in the more recently diverging flowering plants (angiosperms) RNA edit sites have been much reduced such that plastid genomes contain ~30 – 40 sites (Tsudzuki, et al. 2001; Gao, et al. 2023).

1.4.4.3 Evolvability of plastid genome

Evolvability describes the capability of a system to evolve, and multiple factors contribute to the prevailing view that the evolvability of genes encoded in the plastid genome has been severely limited. First, given that the majority of plastid-encoded genes are single copy (except those encoded in the inverted repeat) there is no genetic redundancy to protect against deleterious mutations. Therefore, if a beneficial allele requires multiple evolutionary steps with worse performing intermediates, the intermediate alleles are likely to be quickly removed by purifying selection before the accumulation of subsequent mutations that collectively result in adaptation. Consequently, the sequence space required for complex adaptations is likely to remain unexplored. Secondly, the plastid has highly efficient DNA repair and mutation removal mechanisms (the latter being gene conversion) that supress genetic variation (Birky and Walsh 1992; Khakhlova and Bock 2006). This is evident from the exceptionally low rate of nucleotide evolution observed in plastid genes (Bendich 1987; Smith 2015). These repair mechanisms, while important for maintaining genome integrity, limit the pool of genetic variation that could fuel adaptive evolution. Thirdly, the transmission of the plastid genome via uniparental inheritance prevents recombination, effectively resulting in each gene having 100% linkage disequilibrium. Consequently, beneficial alleles arising in one plastid genome cannot be shared with others in the community, but rather each lineage must ‘reinvent’ its own beneficial alleles, further reducing the likelihood of adaptive evolution occurring in plastid-encoded genes. Moreover, the absence of recombination prevents the separation of beneficial substitutions from deleterious ones, reducing the likelihood that advantageous substitutions become fixed in the population, and thus limiting the adaptive potential of the plastid genome.

As a consequence of the belief that the evolvability of the plastid genome is highly constrained, there has been little effort in investigating adaptation during the evolution of the plastid-encoded genes. The exception to this is the large subunit of rubisco where extensive efforts have been made to understand its evolution in the hope of making it a better enzyme to improve crop yields (Kapralov and Filatov 2007; Prywes, et al. 2023; Bouvier, et al. 2024).

1.5 Aims of the thesis

The overall aim of this thesis was to explore the evolutionary potential of the plastid-encoded genes by investigating both the constraints on their evolution and search for signatures of adaptation that may have been previously overlooked. Identifying such adaptive changes in genes vital for photosynthesis, and hence growth, could offer valuable insight for precision edits. Although these adaptive changes may not naturally combine due to the uniparental inheritance of the plastid genome, they could be artificially integrated into a single plastid genome, potentially enhancing crop productivity and resilience. The specific aims of the three results chapters in this thesis are outlined below:

Chapter 2: The aim of this chapter was to investigate constraints on the molecular evolution of the plastid-encoded genes during the radiation of the angiosperms. This involved curating a dataset of ~800 angiosperm genomes and measuring the rates of molecular evolution of their highly conserved genes. Multiple factors such as gene position, amino acid composition, nucleotide composition, levels of gene expression, transcript biosynthesis cost and translational efficiency were then examined to see whether they could explain any of the observed variation in the rate of molecular evolution. The main results of this chapter have been published in the journal Genome Biology and Evolution (Robbins and Kelly 2023).

Chapter 3: The aim of this chapter was to search for adaptive evolution in the plastid genome's slowest evolving genes, the photosystem genes. To do this, a new method for identifying recurrent evolution, which is a hallmark of adaptation, was developed. This method was used in combination with positive selection tests and uncovered widespread signatures of adaptive evolution in the photosystem genes. The impact of the evolutionary changes at these sites was investigated through *in silico* modelling on protein structures, providing insights to the adaptive benefit they confer. The work presented in this chapter has been accepted for publication in The Plant Cell.

Chapter 4: The aim of this final results chapter was to develop a software package, called RECUR, to detect recurrent evolution, and was based on the methodology established in Chapter 3. An exemplary analysis using RECUR on all non-redundant SARS-CoV-2 surface glycoprotein sequences is provided. The SARS-CoV-2 surface glycoprotein protein was chosen given the general interest of its evolution as it may impact effectiveness of COVID vaccines or transmissibility of the SARS-CoV-2 virus. The results presented in this chapter show RECUR is fast, easy to use and scalable to thousands of sequences.

The results chapters described above are three independent bodies of work and are presented as three manuscripts, each with an abstract, introduction, results, discussion and methods section. Additionally, author contributions are provided at the end of each chapter and supplementary material is provided in the Appendices. The final chapter, Chapter 5, discusses the discoveries and work in this thesis within the wider literature and highlights avenues for future work.

Chapter 2

The evolutionary constraints on angiosperm chloroplast adaptation

Elizabeth HJ Robbins and Steven Kelly

2.1 Abstract

The chloroplast (plastid) arose via the endosymbiosis of a photosynthetic cyanobacterium by a non-photosynthetic eukaryotic cell approximately 1.5 billion years ago. Although the plastid underwent rapid evolution by genome reduction, its rate of molecular evolution is low and its genome organisation is highly conserved. Here, we investigate the factors that have constrained the rate of molecular evolution of protein coding genes in the plastid genome. Through phylogenomic analysis of 773 angiosperm plastid genomes we show that there is substantial variation in the rate of molecular evolution between genes. We demonstrate that the distance of a plastid gene from the likely origin of replication influences the rate at which it has evolved, consistent with time and distance dependent nucleotide mutation gradients. In addition, we show that the amino acid composition of a gene product constrains its substitution tolerance, limiting its mutation landscape and its corresponding rate of molecular evolution. Finally, we demonstrate that the mRNA abundance of a gene is a key factor in determining its rate of molecular evolution, suggesting an interaction between transcription and DNA repair in the plastid. Collectively, we show that the location, the composition, and the expression of a plastid gene can account for >50% of the variation in its rate of molecular evolution. Thus, these three factors have exerted a substantial limitation on the capacity for adaptive evolution in plastid-encoded genes, and ultimately constrained the evolvability of the chloroplast.

2.2 Introduction

The chloroplast (plastid) is a membrane bound organelle that arose from the endosymbiosis of a photosynthetic cyanobacterium by a non-photosynthetic eukaryotic cell approximately 1.5 billion years ago (Schwartz and Dayhoff 1978; Martin and Kowallik 1999; Hedges, et al. 2004). Early in the transition from cyanobacterium to organelle, the plastid underwent a substantial genome reduction whereby thousands of genes were either lost or transferred to the nucleus (Blanchard and Schmidt 1995; Martin, et al. 2002; Kelly 2021). As a consequence, modern-day plastid genomes in plants contain fewer than 5% of the genes found in their free-living bacterial relatives, with the majority of plant plastids containing only ~80 protein-coding genes (Palmer 1985; Timmis, et al. 2004). As a result, despite comprising over 80% of the protein (Makino and Osmond 1991; Li, et al. 2017) and 80% of the mRNA (Forsythe, et al. 2022) in a plant cell, the combined forces of gene loss and endosymbiotic gene transfer mean that plastids typically contain fewer than 0.5% of the genes found in a typical plant cell.

Although the plastid underwent rapid evolution by genome reduction, the rate of molecular evolution of the plastid-encoded genes has been markedly slow and the organisation in the plastid chromosome has been highly conserved. Comparisons between plant species has revealed that the plastid genome has a mutation rate that is an order of magnitude lower than that of the nuclear genome (Wolfe, et al. 1987; Smith 2015). The finding that the mutation rate of the plastid was low in comparison to the nuclear genome was initially unexpected given that the plastid genome is predominantly uniparentally inherited (Birky 1995; Mogensen 1996), where an absence of sexual recombination is thought to lead to an accumulation of deleterious mutations in a phenomenon known as Muller's ratchet (Muller 1964). Accordingly, it was proposed that gene conversion in the presence of a high genome copy number per plastid (~900 per plastid (Bendich 1987)) acts to counteract the effect of Muller's ratchet to prevent accumulation of mutations (Birky and Walsh 1992; Khakhlova and Bock 2006). However, it is unclear whether other factors intrinsic to the plastid genome itself, also act to limit the rate of molecular evolution of plastid genes.

One factor that could differentially impact the rate of molecular evolution of plastid-encoded genes is their position in the chromosome relative to the point of replication initiation. Most plant species contain a plastid genome that has a quadripartite structure (Palmer 1985; Daniell, et al. 2016) composed of two single-copy regions (the large single-copy region and the small single-copy region) separated by two inverted repeat regions. Although multiple replicative mechanisms have been hypothesised (Kolodner and Tewari 1975; Takeda, et al. 1992; Kunnimalaiyaan and Nielsen 1997; Oldenburg and Bendich 2004; Oldenburg and

Bendich 2016), the origin(s) of replication have been repeatedly proposed to be located in the inverted repeat regions, and both the replication mechanism and position relative to these origins are thought to impact the evolution of the nucleotide sequence along the plastid genome (Wolfe, et al. 1987; Smith 2015). For example, adenine to guanine deamination gradients disseminate away from the inverted repeat regions, and mutation rates in the single-copy regions are much higher than in the inverted repeats (Wolfe, et al. 1987; Krishnan and Rao 2009). The formation of the deamination gradients is thought to be attributable to the differential length of time in which the regions of the plastid genome are single-stranded during DNA replication, and the difference in mutation rate between the single-copy and inverted repeat regions is thought to be due to enhanced frequency of homologous recombination mediated repair in the inverted repeat (Khakhlova and Bock 2006). Other context and location dependent effects on mutation in the plastid genome have been described which also alter the rate and type of mutations that occur along the plastid chromosome (Wolfe, et al. 1992; Kelchner 2000; Jansen, et al. 2007; Zhu, et al. 2016; Niu, et al. 2017). However, it is unknown whether these position-dependent effects have influenced the rate of molecular evolution of plastid gene sequences, or how these phenomena interact with other factors to modulate the rate of molecular evolution of plastid-encoded genes.

Here, we present an analysis of the molecular evolution of the single-copy plastid-encoded protein-coding genes from 773 species of angiosperms. We show that there is substantial variation in the rate of molecular evolution between genes, with genes whose products function in energy production having evolved more slowly than those whose products function in information processing (transcription and translation). We then perform an analysis of multiple factors that could contribute to the variation in the rate of molecular evolution between genes, including gene position, amino acid composition, nucleotide composition, levels of gene expression, transcript biosynthesis cost, and transcript translational efficiency. In doing so, we reveal that the level of mRNA abundance of a gene, its position relative to the inverted repeat region, and its amino acid composition each significantly constrain the rate of evolution of plastid-encoded genes. Thus, the combined action of compositional factors and production requirements impose a severe constraint on the molecular adaptation of the chloroplast.

2.3 Results

2.3.1 There is a large variation in rate of molecular evolution between plastid-located genes

To understand the factors that have influenced the rate of molecular evolution of plastid-encoded genes, a dataset of fully sequenced plastid genomes was compiled. To do this, the

full set of sequenced plastid genomes was downloaded from NCBI (Wolfsberg, et al. 2001) and subjected to filtration to remove genomes that did not contain a full set of the genes typically found in angiosperms. This resulted in a dataset comprising complete plastid genomes from 773 species of angiosperms, each containing 69 single-copy protein-coding genes that are ubiquitously present and in all species. Gene trees were inferred from nucleotide multiple sequence alignments for all ubiquitously conserved genes (see Methods) and the phylogenetic trees were used to evaluate and compare the gene wide rates of synonymous (d_S) and non-synonymous substitution (d_N). The synonymous substitution rate is computed from mutations that do not alter the encoded protein and thus informs the background substitution rate to which the gene is exposed. In contrast, the non-synonymous substitution rate is computed from mutations that alter the encoded protein sequence and shows the rate at which the protein is evolving. Moreover, the ratio of the rate of non-synonymous substitution to synonymous substitution (d_N/d_S) indicates the strength of selection to which a gene is subject, whereby genes with smaller values encode proteins subject to greater purifying selection due to increased functional constraint.

The total length of each gene tree, i.e., the sum of all the branch lengths in the gene tree, was used as the estimate for the rate of synonymous substitution and the rate of non-synonymous substitution for each gene, respectively. This measure accounts for differences in alignment length between genes because the branch lengths in the tree are estimates of the number of substitutions per sequence site (Figure A.1). This measure is also not affected by species sampling, variation in model of sequence evolution, or variation in topology of the underlying tree, as each of these factors was held constant in all analyses that were conducted. Specifically, all 773 plant species are in each gene tree, and each gene is inferred from a strictly single-copy gene in each species. Furthermore, the topology of each gene tree was constrained to the consensus topology inferred from a concatenation of all genes (see Methods), and the model of codon evolution was held constant across all individual tree inferences using parameters that allow for comparison of branch lengths between trees. Thus, this analysis removes any potential phylogenetic bias that could occur from variation in these factors.

Comparison of the rates of synonymous substitution between genes revealed that there was a 3-fold difference between the genes with the fastest and slowest background substitution rate, corresponding to *ndhF* and *psbF*, respectively (Figure 1A, Table A.1). In contrast, the rate of non-synonymous substitution varied much more substantially with a 47-fold difference between genes with the fastest and slowest evolving encoded proteins, *rpl22* and *atpH*, respectively (Figure 1B, Table A.1). Moreover, although all gene wide values of d_N/d_S were

small (<0.5), there was still a 23-fold difference between genes subject to the weakest and strongest purifying selection, *matK* and *psaC*, respectively (Figure 1C, Table A.1). Thus, even though the plastid genome has evolved slowly in comparison to the nuclear genome (Wolfe, et al. 1987; Smith 2015), there is still substantial variation in rate of background substitution, rate of protein evolution and functional constraint between plastid-encoded genes.

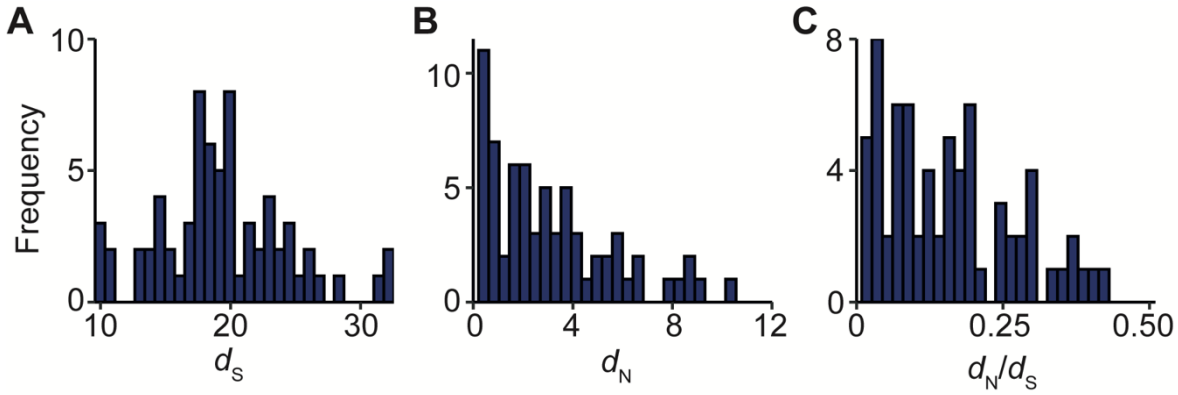


Figure 1. The rates of molecular evolution for 69 single-copy protein-coding plastid genes. **A-C**) Frequency distributions of the rates of synonymous substitution (d_S), non-synonymous substitution (d_N), and strength of selection (d_N/d_S), respectively. d_S has the units of number of synonymous substitutions per synonymous sequence site and d_N has the units of the number of non-synonymous substitutions per non-synonymous sequence site.

2.3.2 Plastid genes with functions in energy production have evolved slower and are more constrained than genes that function in information processing

Inspection of the rank order of the rate of synonymous substitution, rate of non-synonymous substitution and strength of purifying selection (Figure 2A-C) suggested that genes involved in energy production (i.e., genes whose products function in photosystem I, II, cytochrome b_6f complex, ATP-synthase and the NAD(P)H dehydrogenase-like complex or CO_2 fixation) had lower rates of molecular evolution and were subject to greater purifying selection than those involved in information processing (i.e., genes that function in transcription, mRNA maturation, translation or protein degradation). Comparison of these two groups revealed that genes with functions in energy production had evolved slower and were subject to greater purifying selection than those with functions in information processing, with a mean fold difference of 1.2 in the rate of synonymous substitution (Welch's t-test, $P < 0.01$) and median fold differences of 3.1 and 2.8 in the rate of non-synonymous substitution and the ratio of non-synonymous to synonymous substitution (Wilcoxon tests, both $P < 0.001$) (Figure 2D-F). Within each category there was little to no difference in the evolutionary rate or functional constraint of genes between multiprotein complexes (Figure A.2). Thus, by all measures

considered here genes with functions in energy production have evolved slower and are more constrained than genes that function in information processing.

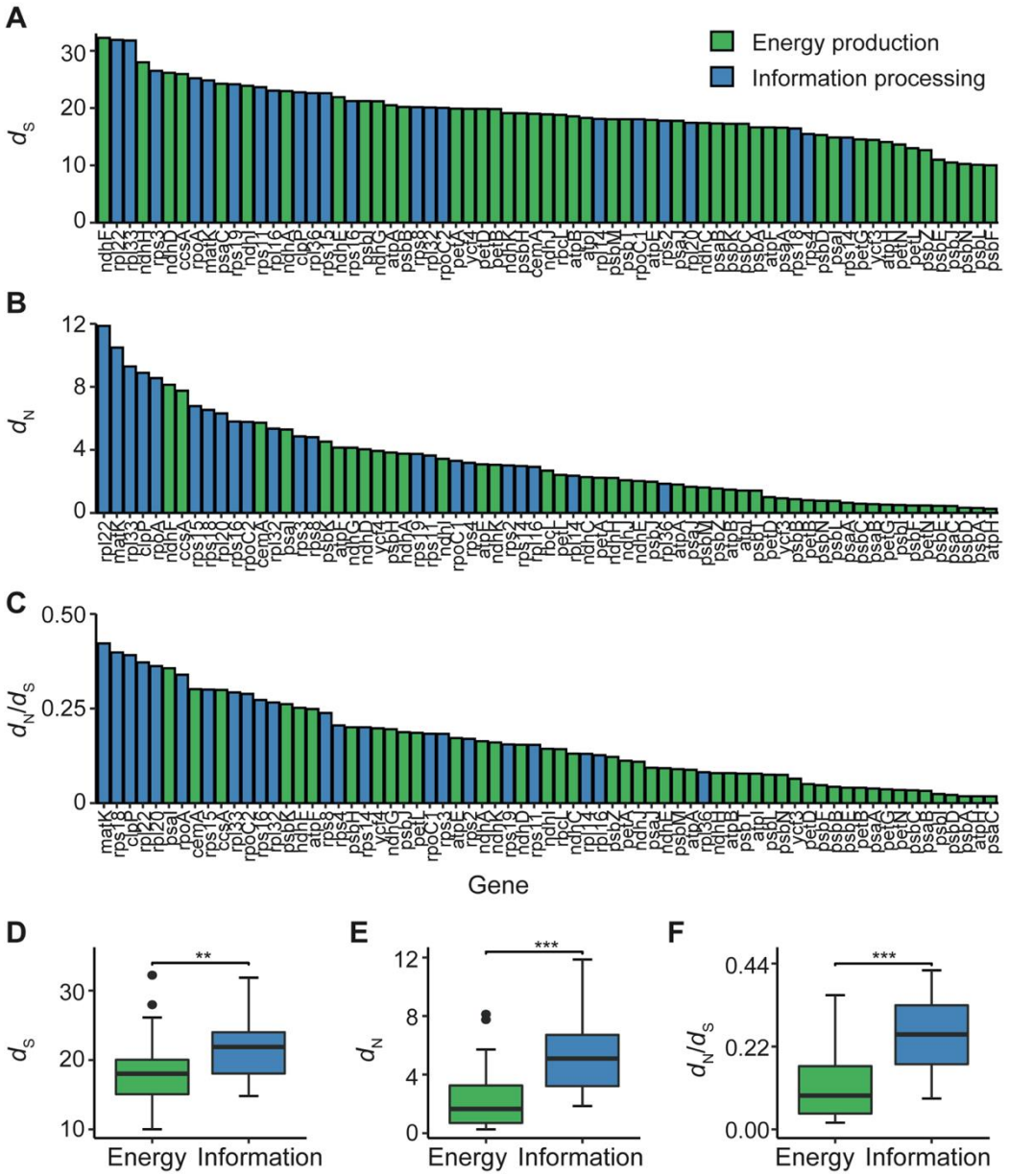


Figure 2. Relationship between gene function with rate of molecular evolution and functional constraint. **A-C**) Bar graphs showing d_S (**A**), d_N (**B**) and d_N/d_S (**C**) for individual genes in order of descending value from left to right. Genes are coloured by their high-level function: green, energy production; blue, information processing. The energy production category includes genes that function in, or are associated with photosystem I, photosystem II, cytochrome b_6f complex, ATP-synthase, NAD(P)H dehydrogenase-like complex or rubisco. The information processing category includes genes involved in mRNA maturation in addition to genes that encode subunits of the 70S ribosome, the plastid-encoded RNA polymerase and the plastid protease system. **D-F**) The difference in d_S (**D**), d_N (**E**) and d_N/d_S (**F**) between genes involved in energy production (green, n=47) versus information processing (red, n=22), respectively. Statistical significance was assessed using a Welch's t-test ($P < 0.01$, indicated by **) for d_S and a Wilcoxon test for d_N and d_N/d_S (both $P < 0.001$, indicated by ***).

2.3.3 The location of a gene in the plastid genome has influenced its rate of molecular evolution

As noted in the introduction, the rate and type of mutation along the plastid genome is non-uniform. Therefore, it is likely that the position of a gene in the plastid genome could influence its rate of molecular evolution. Given that all the genes analysed here are in the single-copy regions of the plastid genome, gene position was assessed by calculating the average distance from each gene's midpoint to the closest border of an inverted repeat region in all 773 species. This distance was then compared with the rate of synonymous substitution, non-synonymous substitution and functional constraint of each gene. This revealed that the further away a gene is located from the inverted repeat, the lower its rate of synonymous substitution (Figure 3A). In total, distance to the inverted repeat border could explain ~30% of the variance the rate of synonymous substitution between genes (R^2 = 0.30, P < 0.001). A similar effect was observed for the rate of non-synonymous substitution (R^2 = 0.13, P < 0.01) (Figure 3B). In contrast, a significant association was not observed between average distance to the inverted repeat border and the strength of selection acting on the encoded protein (Figure 3C). Similar relationships were observed if the rates of molecular evolution and strength of selection of individual genes were aggregated at the level of operons (Figure A.3). Thus, all locations in the plastid genome are not equal, and the rate at which a plastid gene has evolved during the radiation of the angiosperms is dependent on how distant that gene is positioned from the inverted repeat in the plastid genome.

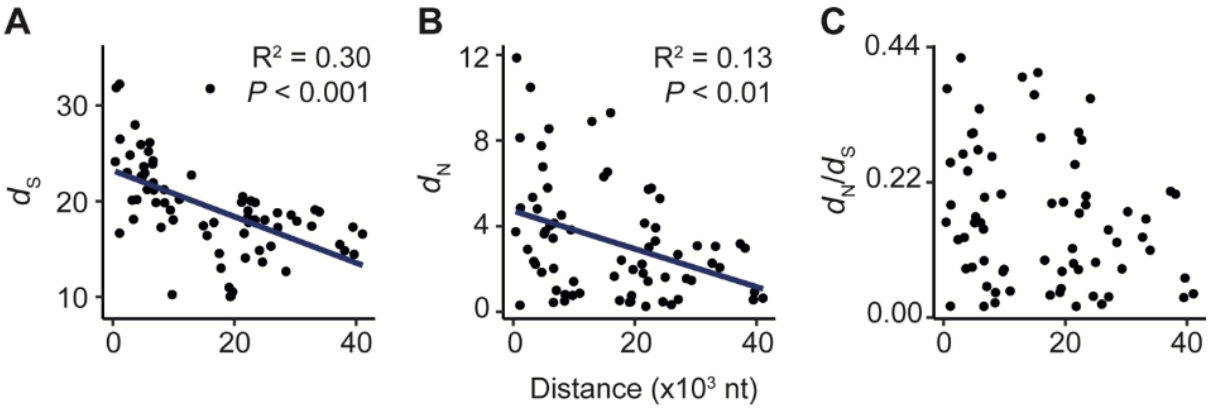


Figure 3. The relationship between the average distance of a gene to the closest inverted repeat border and the rate of molecular evolution of 69 single-copy plastid-encoded genes. Scatter plots showing the average distance of each gene's midpoint to the closest inverted repeat border measured in nucleotides (nt) versus the rate of synonymous substitution (**A**), non-synonymous substitution (**B**) and functional constraint (**C**). Where appropriate, linear models are shown in blue with their associated R^2 and P-values.

2.3.4 Physicochemical constraints on amino acid substitution have modulated the strength of purifying selection acting on plastid-encoded genes

Given that the functional constraint of a protein was independent of any genome position effects, we sought to determine whether any component of this constraint was attributable to compositional differences between the encoded protein sequences. Such a compositional effect may arise from the biochemical differences and similarities between amino acid monomers found in a given protein sequence. This is because amino acids are not all equally dissimilar to each other, and many share similarities in their size, charge, and reactivity of their side chains. The extent of similarity between amino acids varies such that each different amino acid monomer has a different tolerance to substitution in protein sequences (Sneath 1966; Epstein 1967; Grantham 1974; Henikoff and Henikoff 1992). Accordingly, variation in constraint on the rate of non-synonymous substitution could arise simply from compositional differences between genes. A substitution tolerance can be effectively described using substitution frequency data summarised in the BLOSUM62 amino acid substitution matrix (Henikoff and Henikoff 1992). We hypothesised that proteins containing a higher proportion of amino acids that were less tolerant to substitution would have slower protein sequence evolution and would be subject to stronger purifying selection because mutations would be more likely to be deleterious to the encoded protein sequence. To investigate this, a substitution tolerance score was calculated for each gene (see Methods) and compared to the rate of synonymous substitution, rate of non-synonymous substitution and strength of purifying selection of that gene. While there was no association between the substitution tolerance of an encoded protein with the rate of synonymous substitution (Figure 4A), there were significant associations between substitution tolerance and the rate of non-synonymous substitution ($R^2 = 0.20$, $P < 0.001$) (Figure 4B) and the strength of purifying selection acting on the gene ($R^2 = 0.22$, $P < 0.01$) (Figure 4C). Here, genes with lower substitution tolerance scores had lower rates of non-synonymous substitution and were subject to stronger purifying selection. Thus, all sequences are not equal, and the rate at which a plastid gene has evolved during the radiation of the angiosperms is dependent on the amino acid composition of that gene.

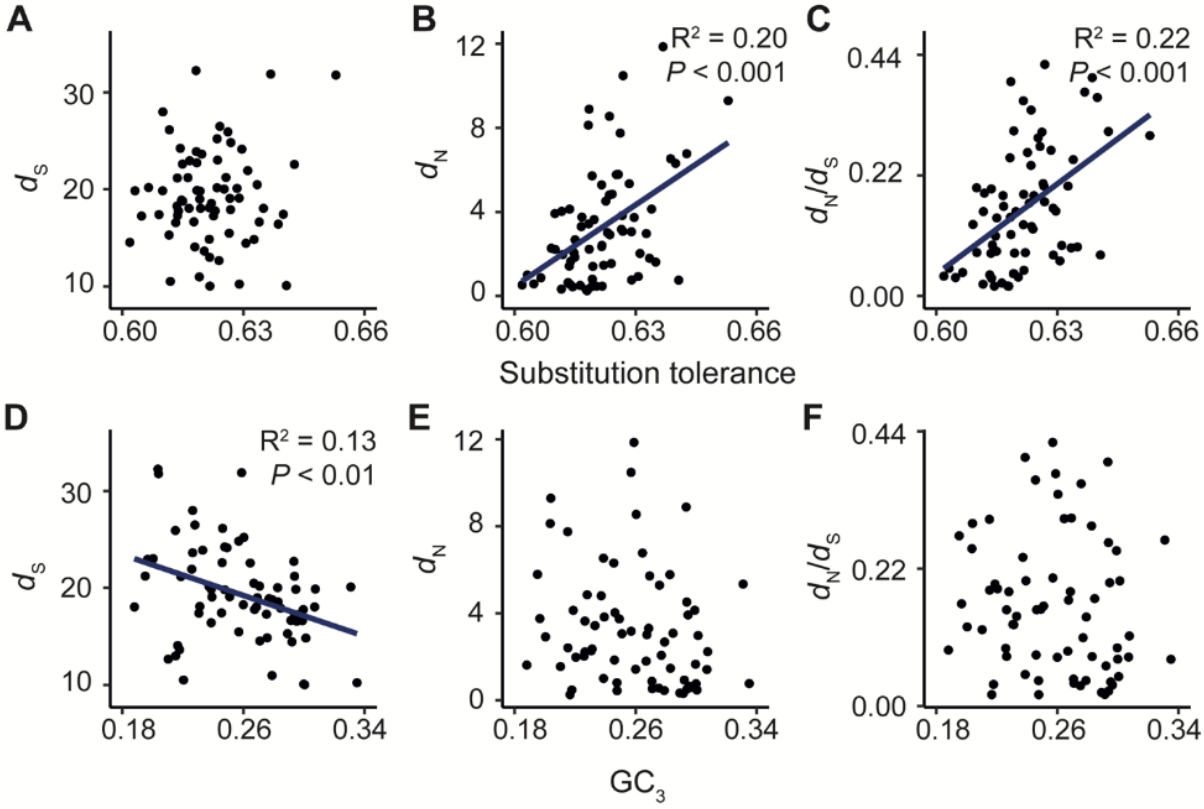


Figure 4. The effect of gene composition on the rate of molecular evolution for plastid-encoded genes. The nucleotide composition of a gene was determined by analysing the GC content at the 3rd codon position (GC_3). Meanwhile, amino acid composition of the encoded protein was summarised in a substitution tolerance score based on the BLOSUM62 matrix, where a lower value corresponds to proteins with a composition less tolerant to substitution. **A-C**) Scatter plots showing protein substitution tolerance scores versus the rate of synonymous substitution, rate of non-synonymous substitution and the ratio of non-synonymous substitution to synonymous substitution, respectively. **D-F**) Scatter plots showing GC content at the 3rd codon position (GC_3) versus the rate of synonymous substitution, rate of non-synonymous substitution and the ratio of non-synonymous substitution to synonymous substitution, respectively. Where appropriate, linear models are shown in blue with their associated R^2 and *P*-values.

2.3.5 The intrinsic nucleotide composition has only weakly influenced the rate of background substitution

Given that the composition of the encoded amino acid sequence has contributed significantly to the rate at which the underlying molecular sequence of a gene has evolved, it was next determined whether similar composition-based constraints were manifest at the level of the nucleotide sequence. There is some precedent for hypothesising that this might occur, as

genes encoded in the inverted repeat regions have higher guanine-cytosine (GC) contents and have evolved more slowly than genes located in the single-copy regions (Li, et al. 2016), and accordingly similar effects may be apparent at the level of an individual gene. To prevent re-detecting differences in amino acid composition at the nucleotide level, only the GC content of the third codon position (GC_3) was investigated, excluding amino acids that have no degeneracy at this position. Comparison of the GC_3 content of a gene to its rate of synonymous substitution, non-synonymous substitution and strength of purifying selection revealed that the GC_3 could explain ~13% of the variance in the rate of synonymous substitution ($R^2 = 0.013$, $P < 0.01$), where genes with higher GC_3 contents have evolved more slowly (Figure 3D). Meanwhile, no significant relationships were identified between GC_3 and the rate of non-synonymous substitution and strength of purifying selection (Figure 3E and F). This latter result was not surprising given that there was no relationship between GC_3 and the GC content of the first two codon positions, i.e., the positions that are major determinants of residue identity (Figure A.4). Thus, while genes that have higher GC_3 contents had slower evolving nucleotide sequences, this interaction is weak and GC_3 has not had an impact on the rate of protein evolution or strength of protein functional constraint.

2.3.6 mRNA abundance is a determinant of plastid gene evolutionary rate

High gene expression has been previously linked to low rates of molecular evolution in multiple systems (Pál, et al. 2001; Subramanian and Kumar 2004; Wright, et al. 2004; Seward and Kelly 2018). To investigate whether this phenomenon was also true of genes encoded in the plastid genome, the rate of molecular evolution and strength of purifying selection of a plastid gene was compared to both the mRNA and protein abundance of that gene. In agreement with previous analysis in other systems, genes with high mRNA abundance had low levels of molecular evolution and *vice versa* (Figure 5). Specifically, variance in mRNA abundance could explain 29% of the variation in the rate of synonymous substitution ($R^2 = 0.29$, $P < 0.001$) (Figure 5A), 32% of variation in the rate of non-synonymous substitution ($R^2 = 0.32$, $P < 0.001$) (Figure 5B) and 25% of variation in the strength of purifying selection ($R^2 = 0.25$, $P < 0.001$) (Figure 5C), respectively. In contrast, there was no significant relationship between protein abundance and rate of molecular evolution or strength of purifying selection (Figure 5D-F). This suggests that the constraint on the evolutionary rate is mediated via an interaction with transcription rather than through the requirement for protein production. Furthermore, previously detected interactions between mRNA abundance and gene evolutionary rate have been manifest in part through selection acting on transcript biosynthesis cost (the number and type of atoms contained within the transcript and the number of high-energy phosphate bonds required biosynthesis) and transcript translation efficiency (a function of the number of tRNAs that can translate that codon encoded in the plastid genome) (Seward and Kelly 2016, 2018).

However, there was no evidence for an interaction between these two factors and variation in the rate of molecular evolution or the strength of purifying selection of genes encoded in the plastid genome (Figure A.5 and A.6). Thus, the extent to which a transcript accumulates in the chloroplast has constrained the evolutionary rate of its corresponding gene, and this is independent of selection acting to preserve the biosynthetic cost, or translational efficiency of the coding sequence, or the abundance of the encoded protein.

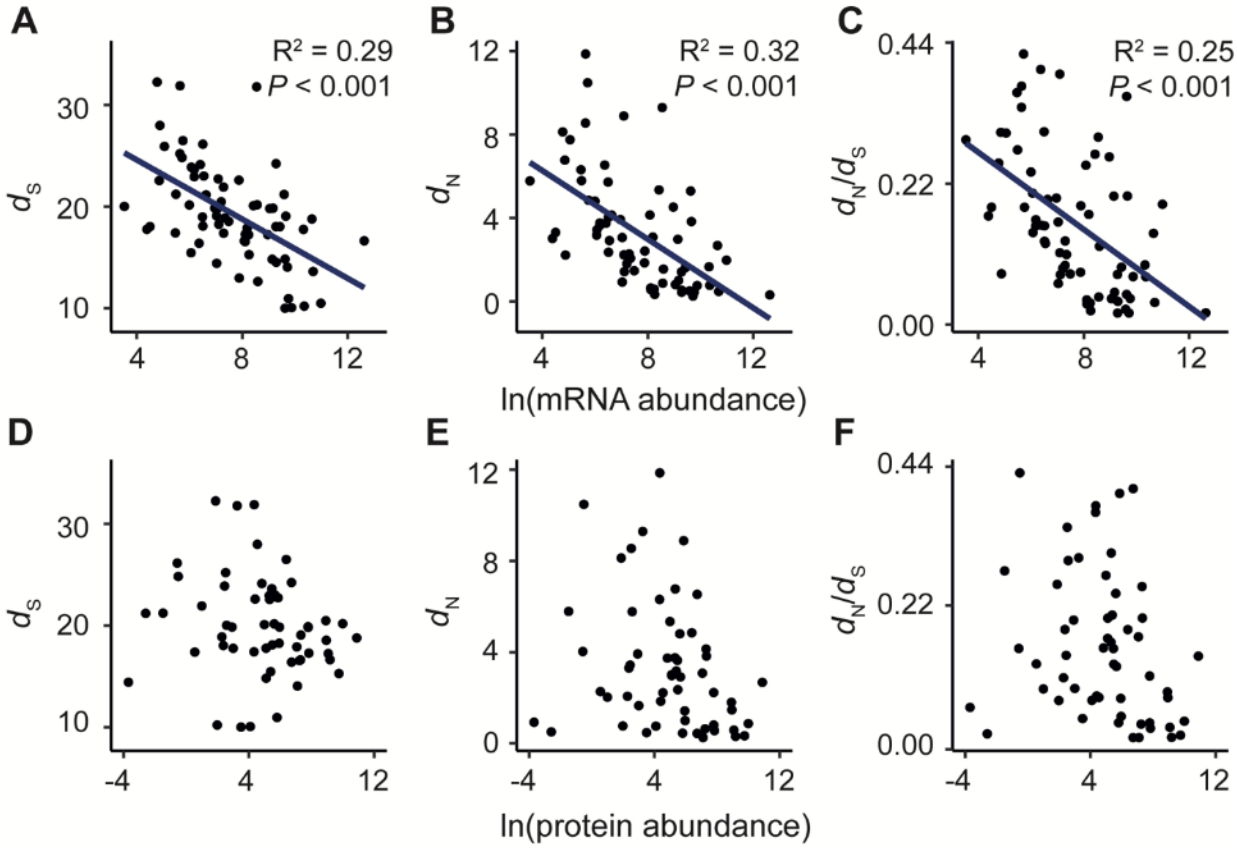


Figure 5. Relationship between the rate of molecular evolution and gene expression for plastid-encoded protein-coding genes. **A-C**) Scatterplots of mRNA abundance (data from (Forsythe, et al. 2022)) versus the rate of synonymous substitution, rate of non-synonymous substitution and the ratio of non-synonymous substitution to synonymous substitution, respectively. Linear regressions are shown as blue lines with their associated R^2 and *P*-values. **D-F**) Scatterplots of protein abundance (data from (Wang, et al. 2012)) versus rate of synonymous substitution, rate of non-synonymous substitution and the ratio of non-synonymous substitution to synonymous substitution, respectively.

2.3.7 Covariate models can explain >50% of the variance in the rate of molecular evolution of plastid-encoded genes

Given that several factors were identified that could each explain a significant component of variation in the rate of molecular evolution of plastid genes in isolation, we next sought to identify the cumulative variance attributable to all factors while accounting for interdependencies (Figure 6A) which may deflate the explained variance observed when single variables are considered. To do this, regression tests were conducted for all possible subsets of the set of independent variables described above. This revealed that gene position and mRNA abundance could explain 52% of the variation in the rate of synonymous substitution (Figure 6B, Figure A.7). Relative importance metrics showed gene position and mRNA abundance were similarly important covariates, contributing to 51% and 49% of the total variance explained, respectively (Figure 6B). A similar proportion of variance in the rate of non-synonymous substitution could be explained by the same factors above with the addition of the protein's substitution tolerance (Figure 6C, 53% variance explained, Figure A.8). Specifically, mRNA was the most important covariate (48% variance explained), followed by protein substitution tolerance (32% variance explained) and then gene position with respect to the inverted repeat border (19% variance explained) (Figure 6C). Although protein sequence constraint is independent of gene position, mRNA abundance and protein substitution tolerance together explained 43% of the variation in purifying selection between genes, contributing 53% and 47% of the explained variance, respectively (Figure 6D, Figure A.9). Thus, in total >50% of the variance in the rate of molecular evolution and >40% of the variance in strength of purifying selection is attributable to compositional factors and production requirements. In all cases, the weak associations between GC_3 content and the rates of molecular evolution and strength of purifying selection (Figure 6) were rendered insignificant when other factors above were accounted for.

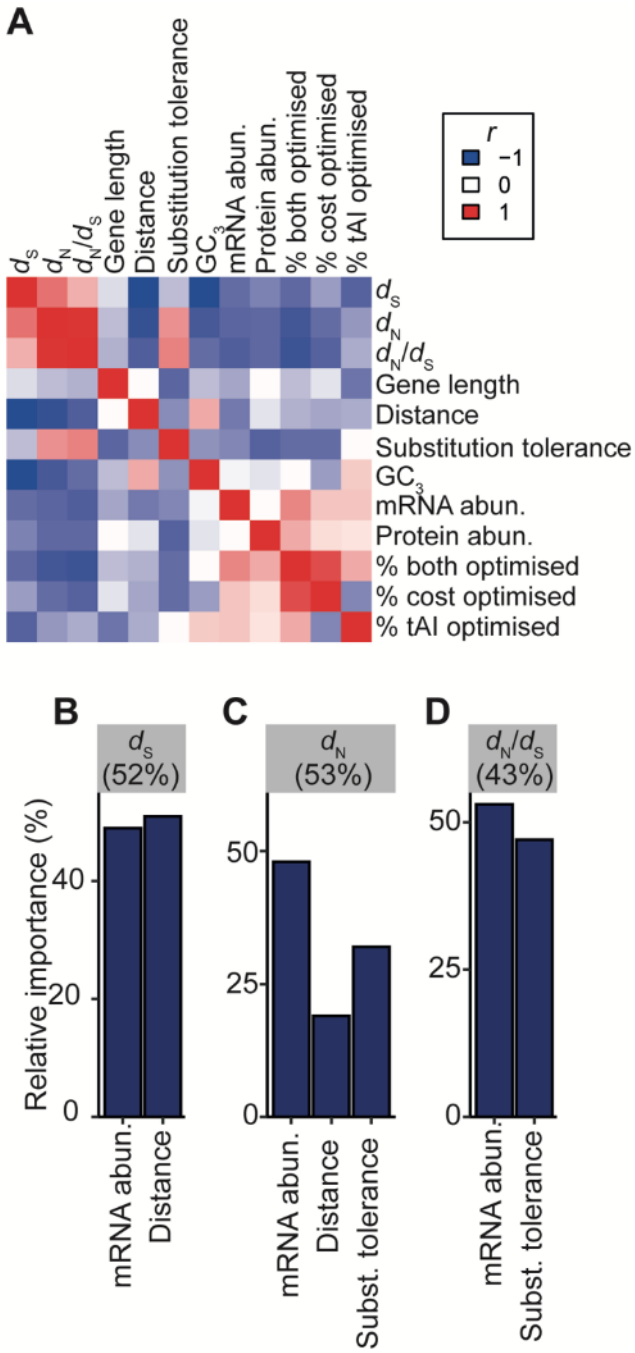


Figure 6. Relative importance of covariates in the best linear models identified to explain the rates of molecular evolution. **A)** A correlation heatmap of all continuous variables used in this study. Red, Pearson correlation coefficient (*r*) = 1; white, *r* = 0; blue, *r* = -1. **B-D)** The relative importance (given as a percentage) of mRNA abundance, gene distance to the inverted repeat and protein substitution tolerance in explaining 52%, 53% and 43% of the variance in the rate of synonymous substitution, rate of non-synonymous substitution and the purifying selection, respectively.

To ensure model space was effectively explored, random forests were also run on the complete set of factors to determine if there were alternative models that could explain the variance in the rate of molecular evolution of the strength of purifying selection (Supplemental File A.1). None of the random forest models outperformed the linear regression analyses shown above, despite including more parameters (Supplemental File A.1). Thus, mRNA abundance, gene position and protein tolerance to substitution together account for >50% of the variance in the molecular evolution and >40% of the variance in protein sequence constraint in the plastid genome, and have substantially constrained the capacity for adaptive evolution in the chloroplast genome.

2.4 Discussion

The endosymbiosis of the bacterial progenitor of the chloroplast is a landmark event in the evolution of eukaryotes. It first enabled oxygenic photosynthesis in eukaryotes and helped direct the Earth on a continued trend of decreasing atmospheric CO_2 and increasing O_2 over the subsequent ~1.5 billion years (Holland 2002; Canfield 2005). Following this endosymbiosis there was a dramatic reduction in the gene content of the plastid genome, such that it now harbours fewer than 5% of the genes found in its free-living bacterial relatives (Palmer 1985; Timmis, et al. 2004). These remaining genes have been trapped in the same genome and inherited in the absence of sexual recombination throughout their evolution (Birky 1995; Mogensen 1996). Here, we investigate the factors that have contributed to the variation in evolutionary rate of plastid-encoded genes over the last 125 million years during the radiation of the angiosperms. We show that the rate of molecular evolution has been governed by factors extrinsic to gene function, specifically, the position of the gene in the plastid genome, the substitution tolerance of the encoded protein and the level of gene expression. Combined, these production requirements and compositional factors explain >50% of the variation in the rates of synonymous and non-synonymous substitution, moreover they also explain >40% of the variation in the rate of purifying selection. Thus, the capacity for molecular evolution of a plastid-encoded gene has been determined to a large extent by the location, composition, and expression level of that gene.

The finding that genes which are positioned further from the inverted repeat border, and hence the likely origin of replication, are evolving more slowly than genes that are closer has two of important implications. First, this finding is compatible with models of chloroplast DNA replication where replication initiates at origins located in the inverted repeat, and where DNA closer to the inverted repeat border spends more time in a single-stranded state than DNA that is more distant to the origin. This is because single-stranded DNA is more susceptible to

base deamination which can result in increased likelihood of mutation and therefore elevated rates of evolution (Lindahl 1993; Krishnan, et al. 2004). Thus, if DNA closer to the origin of replication spends more time single-stranded during replication, this could lead to a positional effect where genes closer to the inverted repeat have elevated rates of molecular evolution. One model of plastid genome replication that satisfies both these criteria is the bi-directional dual displacement loop mechanism (Kolodner and Tewari 1975). However, there is increasing evidence that plastid genome replication proceeds via a recombination dependent mechanism (Oldenburg and Bendich 2004; Oldenburg and Bendich 2016), and while these mechanisms have also been proposed to initiate in the inverted repeat, it is difficult to conceptualise how DNA would spend differential time single-stranded in these models. Therefore, our findings are consistent with the bi-directional dual displacement loop mechanism and lend further support to analyses of intergenic regions and the third codon position that indicate the wide-spread operation of this form of plastid DNA replication in angiosperms (Krishnan and Rao 2009). Second, the fact that origin-proximal genes are evolving faster than origin-distal genes means that all locations in the plastid genome are not equal in terms of their capacity to generate genetic diversity. Accordingly, the genes located further from the inverted repeat have had the capability to explore more sequence space, and have generated more diversity on which natural selection can act, during angiosperm evolution. Thus, the position of a gene and the structure of the plastid genome has constrained the evolvability of adaptive changes in the chloroplast.

The 20 amino acids that are used to build proteins vary in the physical and chemical properties of their side chains: some are hydrophobic, some are charged, some are large, and some are small. Consequently, a mutation that results in the substitution of one amino acid for another with similar physicochemical properties is more likely to have a smaller impact on the structure and function of a protein than substitution with an amino acid with substantially different properties (Sneath 1966; Epstein 1967; Henikoff and Henikoff 1992; Bohórquez, et al. 2017). As there are different numbers of similar amino acids in each physicochemical category, and each amino acid can occupy different physicochemical categories, each amino acid therefore has a different tolerance to substitution. We estimated the substitution tolerance of each protein sequence encoded in the plastid genome and found that the aggregate substitution tolerance of a protein sequence is a major component of variation in its evolutionary rate. This means that the evolvability of novel adaptive phenotypes arising from variation in plastid-encoded genes is limited by the identity of the encoded amino acids in that gene.

In agreement with previous analyses in other systems, we observed that plastid genes with high expression levels had lower rates of molecular evolution (Pál, et al. 2001; Subramanian

and Kumar 2004; Wright, et al. 2004; Seward and Kelly 2018). However, in contrast to previous studies, this phenomenon was not attributable to selection acting to maintain a low transcript biosynthesis cost or a high translational efficiency, nor was it attributable to the requirement for high protein abundance. The lack of association with protein abundance could be due to increased noise in the data as it is an integrated dataset, or alternatively, protein abundance is not a good proxy for protein production as protein turnover is not accounted for. Nevertheless, the results presented here raise the question as to what phenomenon could give rise to a strong interaction between mRNA abundance with the rate of molecular evolution that is not attributable to transcript biosynthetic cost, translational efficiency or protein abundance. One mechanism that could potentially explain this dependency is the presence of transcription mediated DNA repair, whereby increased transcriptional activity directly enhances DNA repair. This would give rise to a phenomenon whereby the more highly transcribed a gene, the more likely DNA damage is repaired in that gene and the lower its mutation rate. Interestingly, this potential explanation is supported by the presence of a nuclear-encoded, chloroplast-targeted ortholog of mutation frequency decline protein (Mfd), alternatively named transcription-repair coupling factor (TRCF) (Hays 2002). In bacteria, Mfd facilitates DNA repair at stalled RNA polymerase complexes by recruiting the nucleotide excision repair machinery (Selby and Sancar 1993). Additionally, it is proposed to play a role in promoting homologous recombination mediated repair (Ayora, et al. 1996). While plants lack orthologs of other nucleotide excision repair genes (Hays 2002), they contain orthologs of the recombination mediated repair factors (Maréchal and Brisson 2010) and recombination-based repair is a primary DNA repair mechanism in the plastid (Cerutti, et al. 1995; Khakhlova and Bock 2006; Maréchal and Brisson 2010). Thus, it is possible that Mfd in plants functions to promote recombination-based repair in response to DNA damage-stalled RNA polymerases. This effect would likely arise from Mfd's ability to form R-loops by distorting the DNA around the stalled polymerase (Portman, et al. 2021). This, in turn, is known to lead to recruitment of plastid homologous recombination machinery to facilitate DNA repair (Wang, et al. 2021). Thus, although it is impossible to prove the existence of transcription-coupled DNA repair from the data presented here, it is consistent with the well-described phenomenon. Furthermore, it will be interesting to determine whether loss of function of Mfd in plants leads to increased mutation in the chloroplast genome in a manner that is consistent with a reduction in transcription mediated DNA repair, and thus can provide a mechanistic explanation for the transcription dependent constraint on the rate of molecular evolution of plastid genes.

Our finding that the expression, location, and amino acid composition of a gene impose significant constraints on rate of molecular evolution of plastid-encoded genes is important. It means that factors other than the function of the encoded protein have been important in

constraining the evolvability of plastid genes. In this context, it is noteworthy that the constraints identified here help provide a mechanistic explanation for the strong phylogenetic constraints that have limited the adaptation of rubisco enzyme kinetics (Bouvier, et al. 2021) and why rubisco is one of the slowest evolving enzymes on earth (Bouvier, et al. 2022). Specifically, evolutionary constraints intrinsic to the location, mRNA abundance and amino acid composition of rubisco (in addition to functional constrains on the protein) may help to explain why it has been so slow to adapt to changing atmospheric conditions, and consequently why it is poorly suited to the O_2-rich, CO_2-poor atmosphere of the present day. Furthermore, it is tempting to speculate that endosymbiotic gene transfer to the nuclear genome allowed other plastid-encoded genes to escape some of these evolutionary constraints, facilitated more rapid adaptation of chloroplast proteins, and provided a further advantage for migration of organellar genes to the nuclear genome.

In summary, the molecular adaptation of the land plant chloroplast has been constrained by the location, composition and expression of its genes. The presence of these evolutionary constraints limits the evolvability of adaptive phenotypes arising from chloroplast encoded genes, and thus motivates strategies to improve photosynthesis in synthetic biology contexts by circumventing these evolutionary barriers to optimisation.

2.5 Methods

2.5.1 Dataset curation

The complete set of fully sequenced plastid genomes was downloaded from the National Center for Biotechnology and Information (NCBI) (https://www.ncbi.nlm.nih.gov/) in July 2021. Those genomes that did not contain a full set of canonical plastid-encoded photosystem I, photosystem II, ATPase, NAD(P)H dehydrogenase-like, cytochrome b_6f complex, ribosomal and RNA polymerase genes were discarded. All other plastid-encoded genes were not included as they were missing from a large number of angiosperm plastid genomes (Table A.2). The full set of genes included in the analysis and their accession numbers are provided in Supplemental File A.2. The protein sequences encoded by these genes were aligned using MAFFT L-INS-I (Katoh and Standley 2013), and visually checked in AliView (Larsson 2014).

2.5.2 Phylogenetic tree inference

The protein multiple sequence alignments were used to guide codon alignment using Pal2Nal (Suyama, et al. 2006) so that there were no inconsistencies between the protein and nucleotide alignments. Nucleotide and protein multiple sequence alignments were trimmed to remove columns containing >90 % gaps using trimAl (Capella-Gutiérrez, et al. 2009). All 69

trimmed nucleotide multiple sequence alignments were concatenated together and used to identify the best-fitting model of sequence evolution using the inbuilt ModelFinder function in IQ-TREE (Kalyaanamoorthy, et al. 2017). In total, 243 models of sequence evolution were tested and the best-fitting model according to the Bayesian information criterion (BIC) was GTR+F+R7. A maximum-likelihood phylogenetic tree was then inferred from the concatenated nucleotide alignment and the best-fit model GTR+F+R7 using IQ-TREE's ultrafast bootstrapping method with 1000 replicates (Hoang, et al. 2018). The resulting tree was then rooted manually in Dendroscope using the Nymphaeales species as the outgroup as they were the most basal-branching clade in the tree (Huson and Scornavacca 2012).

2.5.3 Substitution rates and selection pressure

To estimate the non-synonymous and synonymous substitution rates for each individual gene, synonymous and non-synonymous gene trees were inferred from the trimmed gene nucleotide multiple sequence alignments and the MG94+F3X4 codon substitution model using HYPHY's FitMG94 method (Muse and Gaut 1994; Pond, et al. 2004). Here, the topology of the trees was constrained to that identified from the concatenated alignment above. This was done as this is the topology that is most likely to be the true topology for all genes given the uniparental inheritance of the plastid in the absence of sexual recombination – i.e. all gene trees should have the same topology. For each non-synonymous and synonymous gene tree, the total tree length, which is the sum of all branch lengths, was extracted from the HYPHY output and used as a measure of gene wide non-synonymous (d_N) and synonymous substitution (d_S) rate. These values were then used to calculate a gene wide ratio of the rate of non-synonymous to synonymous substitution (d_N/d_S) which is a measure of selection pressure. Due to the high species sampling and low rate of molecular evolution of the plastid genome, all per-branch estimated values of d_S and d_N were less than 0.88 and 0.32, respectively (Table A.1). Thus, saturation did not occur on any branch in any tree and therefore could not have influenced the results presented in this study. There were too few indels to facilitate an analysis of indel evolution (Supplemental File A.3).

2.5.4 Protein tolerance to substitution

Protein tolerance to substitution scores were calculated using the BLOSUM62 matrix (Henikoff and Henikoff 1992). BLOSUM62 was chosen as it is the most widely used substitution matrix the biological sciences, being used in BLAST (and related sequence search methods) and nearly all protein multiple sequence alignment methods. Here, the odds ratios for all substitutions were back calculated from the log-odds scores provided. Each residue was then given a substitution tolerance by calculating the geometric mean of the odds ratios of its

substitution to the remaining 19 common residues. Then, for each of the 69 proteins analysed, the average residue substitution tolerance was calculated by taking the arithmetic mean of the concatenated protein sequence (composed from all 773 species sequences for that protein).

2.5.6 Gene position and GC_3 content

Although the plastid genome is a highly conserved structure in the angiosperm lineage, its size, and thus gene position, varies between species. Since all of the genes included in this analysis are in the single-copy region, the average genetic distance of a gene's midpoint to the closest inverted repeat border was used as a measure of gene position relative to the putative origin of replication (the inverted repeat). To do this, the positions of the inverted repeat regions were identified for each species by searching for large-duplicated regions using BLASTN (Altschul, et al. 1990; Camacho, et al. 2009). Then, the distance between the gene's midpoint (obtained from the GenBank file for each species' plastid genome) and the closest inverted repeat border was calculated. An average distance was then determined and this average for each gene was used in the analysis of gene position. Positional effects were also analysed at the level of operons. To do this, operon maps were obtained from (Shahar, et al. 2019) for six diverse species in our dataset. The operon's distance to the closest inverted repeat border was determined by averaging the distances of its encoded genes.

The average GC content of the 3rd codon position (GC_3) for all 773 species was calculated for each gene. The GC_3 excluded methionine and tryptophan codons as these amino acids are encoded by a single codon, and thus have no flexibility in base identity at the 3rd position.

2.5.7 mRNA abundance and protein abundance

mRNA abundance data were obtained from (Forsythe, et al. 2022) and values were available for all 69 genes analysed in this study. The whole organism integrated (3702-WHOLE_ORGANISM-integrated.txt) protein abundance estimates for *Arabidopsis thaliana* were obtained from the Protein Abundance Database (PaxDb) (Wang, et al. 2012). Abundance data was available for 56/69 genes included in this analysis.

2.5.8 Transcript biosynthetic cost and translational efficiency

The strength of selection acting to reduce transcript cost, increase translational efficiency and the trade-off between these two forces was inferred for each species using CodonMuSe (Seward and Kelly 2018). The tRNA copy numbers required by CodonMuSe were inferred for each species by inputting the species plastid genome to tRNAscan-SE 2.0 (Chan, et al. 2021).

2.5.9 Statistical analysis and modelling

All statistical analysis and linear regressions were calculated in R. Identification of the best subset of variables for multiple linear regressions was determined using the olsrr package. Relative importance metrics were generated using the reliampo package (Groemping 2006). Random forests were using the randomForest package (Svetnik, et al. 2003).

2.6 Acknowledgements

This work was funded by the Royal Society and the European Union's Horizon 2020 research and innovation program under grant agreement number 637765. ER is funded by the Biotechnology and Biological Sciences Research Council (BBSRC) [grant numbers BB/M011224/1 and BB/P003117/1]. This research was funded in whole, or in part, by the BBSRC. For the purpose of open access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

2.7 Data Availability

All data used in this study is provided in the supplemental material, and all accession numbers are provided.

2.8 Author Contributions

ER and SK conceived the study. ER conducted the analysis. ER and SK wrote the manuscript.

2.9 References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403-410.

Ayora S, Rojo F, Ogasawara N, Nakai S, Alonso JC. 1996. The Mfd protein of Bacillus subtilis168 is involved in both transcription-coupled DNA repair and DNA recombination. Journal of Molecular Biology 256:301-318.

Bendich AJ. 1987. Why do chloroplasts and mitochondria contain so many copies of their genome? Bioessays 6:279-282.

Birky CW, Jr. 1995. Uniparental inheritance of mitochondrial and chloroplast genes: mechanisms and evolution. Proc Natl Acad Sci U S A 92:11331-11338.

Birky CW, Jr., Walsh JB. 1992. Biased gene conversion, copy number, and apparent mutation rate differences within chloroplast and bacterial genomes. Genetics 130:677-683.

Blanchard JL, Schmidt GW. 1995. Pervasive migration of organellar DNA to the nucleus in plants. J Mol Evol 41:397-406.

Bohórquez HJ, Suárez CF, Patarroyo ME. 2017. Mass & secondary structure propensity of amino acids explain their mutability and evolutionary replacements. Scientific Reports 7:7717.

Bouvier JW, Emms DM, Kelly S. 2022. Slow molecular evolution of rubisco limits adaptive improvement of CO2 assimilation. bioRxiv:2022.2007.2006.498985.

Bouvier JW, Emms DM, Rhodes T, Bolton JS, Brasnett A, Eddershaw A, Nielsen JR, Unitt A, Whitney SM, Kelly S. 2021. Rubisco adaptation Is more limited by phylogenetic constraint than by catalytic trade-off. Molecular Biology and Evolution 38:2880-2896.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.

Canfield DE. 2005. The early history of atmospheric oxygen: Homage to Robert M. Garrels. Annual Review of Earth and Planetary Sciences 33:1-36.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972-1973.

Cerutti H, Johnson AM, Boynton JE, Gillham NW. 1995. Inhibition of chloroplast DNA recombination and repair by dominant negative mutants of Escherichia coli RecA. Mol Cell Biol 15:3003-3011.

Chan Patricia P, Lin Brian Y, Mak Allysia J, Lowe Todd M. 2021. tRNAscan-SE 2.0: Improved detection and functional classification of transfer RNA genes. Nucleic Acids Research 49:9077-9096.

Daniell H, Lin C-S, Yu M, Chang W-J. 2016. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. Genome Biology 17:134.

Epstein CJ. 1967. Non-randomness of amino-acid changes in the evolution of homologous proteins. Nature 215:355-359.

Forsythe ES, Grover CE, Miller ER, Conover JL, Arick MA, Chavarro MCF, Leal-Bertioli SCM, Peterson DG, Sharbrough J, Wendel JF, et al. 2022. Organellar transcripts dominate the cellular mRNA pool across plants of varying ploidy levels. bioRxiv:2022.2003.2012.484027.

Grantham R. 1974. Amino acid difference formula to help explain protein evolution. Science 185:862-864.

Groemping U. 2006. Relative importance for linear regression in R: The package relaimpo. Journal of Statistical Software 17:1 - 27.

Hays JB. 2002. Arabidopsis thaliana, a versatile model system for study of eukaryotic genome-maintenance functions. DNA Repair 1:579-600.

Hedges SB, Blair JE, Venturi ML, Shoe JL. 2004. A molecular timescale of eukaryote evolution and the rise of complex multicellular life. BMC Evol Biol 4:2.

Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 89:10915-10919.

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: Improving the ultrafast bootstrap approximation. Mol Biol Evol 35:518-522.

Holland HD. 2002. Volcanic gases, black smokers, and the great oxidation event. Geochimica et Cosmochimica Acta 66:3811-3826.

Huson DH, Scornavacca C. 2012. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Syst Biol 61:1061-1067.

Jansen RK, Cai Z, Raubeson LA, Daniell H, Depamphilis CW, Leebens-Mack J, Müller KF, Guisinger-Bellian M, Haberle RC, Hansen AK, et al. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc Natl Acad Sci U S A 104:19369-19374.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods 14:587-589.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772-780.

Kelchner SA. 2000. The evolution of non-coding chloroplast DNA and its application in plant systematics. Annals of the Missouri Botanical Garden 87:482-498.

Kelly S. 2021. The economics of organellar gene loss and endosymbiotic gene transfer. Genome Biology 22:345.

Khakhlova O, Bock R. 2006. Elimination of deleterious mutations in plastid genomes by gene conversion. Plant J 46:85-94.

Kolodner R, Tewari KK. 1975. Presence of displacement loops in the covalently closed circular chloroplast deoxyribonucleic acid from higher plants. J Biol Chem 250:8840-8847.

Krishnan NM, Rao BJ. 2009. A comparative approach to elucidate chloroplast genome replication. BMC Genomics 10:237.

Krishnan NM, Seligmann H, Raina SZ, Pollock DD. 2004. Detecting gradients of asymmetry in site-specific substitutions in mitochondrial genomes. DNA Cell Biol 23:707-714.

Kunnimalaiyaan M, Nielsen BL. 1997. Fine mapping of replication origins (oriA and oriB) in Nicotiana tabacum chloroplast DNA. Nucleic Acids Research 25:3681-3686.

Larsson A. 2014. AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics 30:3276-3278.

Li FW, Kuo LY, Pryer KM, Rothfels CJ. 2016. Genes translocated into the plastid inverted repeat show decelerated substitution rates and elevated GC content. Genome Biol Evol 8:2452-2458.

Li L, Nelson CJ, Trösch J, Castleden I, Huang S, Millar AH. 2017. Protein degradation rate in Arabidopsis thaliana leaf growth and development. Plant Cell 29:207-228.

Lindahl T. 1993. Instability and decay of the primary structure of DNA. Nature 362:709-715.

Makino A, Osmond B. 1991. Effects of nitrogen nutrition on nitrogen partitioning between chloroplasts and mitochondria in pea and wheat 1. Plant Physiology 96:355-362.

Maréchal A, Brisson N. 2010. Recombination and the maintenance of plant organelle genome stability. New Phytol 186:299-317.

Martin W, Kowallik KV. 1999. Annotated English translation of Mereschkowsky's 1905 paper ‘Über Natur und Ursprung der Chromatophoren im Pflanzenreiche’. European Journal of Phycology 34:287-295.

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. Proceedings of the National Academy of Sciences 99:12246-12251.

Mogensen HL. 1996. The hows and whys of cytoplasmic inheritance in seed plants. American Journal of Botany 83:383-404.

Muller HJ. 1964. The relation of recombination to mutational advance. Mutat Res 106:2-9.

Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol 11:715-724.

Niu Z, Xue Q, Wang H, Xie X, Zhu S, Liu W, Ding X. 2017. Mutational biases and GC-biased gene conversion affect GC content in the plastomes of Dendrobium genus. Int J Mol Sci 18.

Oldenburg DJ, Bendich AJ. 2016. The linear plastid chromosomes of maize: terminal sequences, structures, and implications for DNA replication. Current Genetics 62:431-442.

Oldenburg DJ, Bendich AJ. 2004. Most chloroplast DNA of maize seedlings in linear molecules with defined ends and branched forms. J Mol Biol 335:953-970.

Pál C, Papp B, Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. Genetics 158:927-931.

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

Pond SLK, Frost SDW, Muse SV. 2004. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21:676-679.

Portman JR, Brouwer GM, Bollins J, Savery NJ, Strick TR. 2021. Cotranscriptional R-loop formation by Mfd involves topological partitioning of DNA. Proc Natl Acad Sci U S A 118.

Schwartz RM, Dayhoff MO. 1978. Origins of prokaryotes, eukaryotes, mitochondria, and chloroplasts. Science 199:395-403.

Selby CP, Sancar A. 1993. Molecular mechanism of transcription-repair coupling. Science 260:53-58.

Seward EA, Kelly S. 2016. Dietary nitrogen alters codon bias and genome composition in parasitic microorganisms. Genome Biol 17:226.

Seward EA, Kelly S. 2018. Selection-driven cost-efficiency optimization of transcripts modulates gene evolutionary rate in bacteria. Genome Biol 19:102.

Shahar N, Weiner I, Stotsky L, Tuller T, Yacoby I. 2019. Prediction and large-scale analysis of primary operons in plastids reveals unique genetic features in the evolution of chloroplasts. Nucleic Acids Research 47:3344-3352.

Smith DR. 2015. Mutation rates in plastid genomes: They are lower than you might think. Genome Biology and Evolution 7:1227-1234.

Sneath PH. 1966. Relations between chemical structure and biological activity in peptides. J Theor Biol 12:157-195.

Subramanian S, Kumar S. 2004. Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. Genetics 168:373-381.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res 34:W609-612.

Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. 2003. Random forest: A classification and regression tool for compound classification and QSAR modeling. Journal of Chemical Information and Computer Sciences 43:1947-1958.

Takeda Y, Hirokawa H, Nagata T. 1992. The replication origin of proplastid DNA in cultured cells of tobacco. Molecular and General Genetics MGG 232:191-198.

Timmis JN, Ayliffe MA, Huang CY, Martin W. 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nature Reviews Genetics 5:123-135.

Wang M, Weiss M, Simonovic M, Haertinger G, Schrimpf SP, Hengartner MO, von Mering C. 2012. PaxDb, a database of protein abundance averages across all three domains of life. Mol Cell Proteomics 11:492-500.

Wang W, Li K, Yang Z, Hou Q, Zhao WW, Sun Q. 2021. RNase H1C collaborates with ssDNA binding proteins WHY1/3 and recombinase RecA1 to fulfill the DNA damage repair in Arabidopsis chloroplasts. Nucleic Acids Research 49:6771-6787.

Wolfe KH, Li WH, Sharp PM. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc Natl Acad Sci U S A 84:9054-9058.

Wolfe KH, Morden CW, Palmer JD. 1992. Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. Proc Natl Acad Sci U S A 89:10648-10652.

Wolfsberg TG, Schafer S, Tatusov RL, Tatusov TA. 2001. Organelle genome resource at NCBI. Trends Biochem Sci 26:199-203.

Wright SI, Yau CB, Looseley M, Meyers BC. 2004. Effects of gene expression on molecular evolution in Arabidopsis thaliana and Arabidopsis lyrata. Mol Biol Evol 21:1719-1726.

Zhu A, Guo W, Gupta S, Fan W, Mower JP. 2016. Evolutionary dynamics of the plastid inverted repeat: the effects of expansion, contraction, and loss on substitution rates. New Phytol 209:1747-1756.

Chapter 3

Widespread adaptive evolution in the photosystems of angiosperms provides insight into the evolution of photosystem II repair.

Elizabeth HJ Robbins and Steven Kelly

3.1 Abstract

Oxygenic photosynthesis generates the initial energy source which fuels nearly all life on Earth. At the heart of the process are the photosystems, pigment binding multi-protein complexes that catalyse the first step of photochemical conversion of light energy into chemical energy. Here, we investigate the molecular evolution at single residue resolution of the plastid-encoded subunits of the photosystems across 773 angiosperm species. We show that despite an extremely high level of conservation, 7% of residues in the photosystems, spanning all photosystem subunits, exhibit hallmarks of adaptive evolution. Through *in silico* modelling of these adaptive substitutions we uncover the impact of these changes on the predicted properties of the photosystems, focussing on their effects on co-factor binding and the formation of inter-subunit interfaces. Through analysis of these cohorts of changes we reveal that evolution has repeatedly altered the interaction between photosystem II and its D1 subunit in a manner that is predicted to reduce the energetic barrier for D1 turn-over and photosystem repair. Together, these results provide new insight into the trajectory of photosystem evolution during the radiation of the angiosperms.

3.2 Introduction

Oxygenic photosynthesis is the process that converts sunlight and carbon dioxide into sugars and oxygen and fuels nearly all life on Earth. Although it remains unclear exactly when oxygenic photosynthesis first arose, there is evidence to suggest it had already evolved by ~3 billion years ago (Planavsky, et al. 2014; Fournier, et al. 2021), and subsequently facilitated the rise of molecular oxygen in the atmosphere and the evolution of aerobic organisms (Eigenbrode and Freeman 2006; Koch and Britton 2008). At the heart of oxygenic photosynthesis are the two photosystem complexes, photosystem I and photosystem II (Blankenship and Hartman 1998). Both photosystems are large multi-subunit protein complexes made up of two parts, the light-harvesting antenna and the reaction centre (Gao, et al. 2018). The function of the antenna complex is to scaffold pigments that absorb and funnel light energy to the reaction centre core (Gao, et al. 2018). The reaction centre uses this light energy for charge separation which ultimately fuels the generation of ATP and NADPH which in turn drive downstream cellular metabolic processes.

The first step of the light-dependent reactions is catalysed by photosystem II, whereby water is oxidised in the thylakoid lumen and the liberated electrons are shuttled through a variety of cofactors to plastoquinone on the stromal side of the thylakoid membrane (Barber 1998). Meanwhile, photosystem I catalyses a later step where reduced plastocyanin from the cytochrome b_6f complex is oxidised in the thylakoid lumen and the liberated electron is shuttled across the thylakoid membrane to reduce ferredoxin in the stroma (Chitnis 2001). During periods of high light, over-reduction of the internal electron transport chain can result in the formation of reactive oxygen species that damage the photosynthetic apparatus (Hideg, et al. 2002; Asada 2006; Pospíšil 2016). While several subunits in the photosystem are damaged by excess light (Yao, et al. 2012), a specific target of this damage is the D1 protein, conferring protective effects by reducing the extent of damage on the rest of the photosystem II complex (Mattoo, et al. 1981; Aro, et al. 1993). In particular, a frequently damaged region of D1 is the ~40 amino acid long stroma-exposed loop between transmembrane helices D and E, aptly termed the DE-loop (Greenberg, et al. 1987; Aro, et al. 1993; Frankel, et al. 2013). As a result of continuous photodamage, D1 needs to be constantly replaced to maintain photosystem function. Accordingly, D1 has the highest turn-over rate of all photosystem proteins and one of the highest expression levels found in nature (Kettunen, et al. 1997; Li, et al. 2017; Forsythe, et al. 2022).

Eukaryotes gained the capacity for photosynthesis ~1.5 billion years ago when a single celled eukaryotic organism engulfed a cyanobacterium (Schwartz and Dayhoff 1978; Martin and Kowallik 1999; Hedges, et al. 2004). As the cyanobacterium evolved into the semi-autonomous organelle known as the plastid, its genome underwent a substantial reduction, such that it now harbours <5% of the genes found in its cyanobacterial ancestor (Blanchard and Schmidt 1995; Martin, et al. 2002; Kelly 2021). Despite this large reduction, the gene content and organisation has been highly conserved across the angiosperm lineage (Palmer and Thompson 1982; Jansen, et al. 2007; Zhu, et al. 2016; Robbins and Kelly 2023). Among the ~80 protein-coding genes that have remained in the plastid genome many encode core photosystem reaction centre proteins. These are single-copy genes, and each has been subject to a low rate of evolution experienced by the plastid genome (Wolfe, et al. 1987; Smith 2015), a direct result of high copy number of the plastid genome and consequently high levels of gene conversion eliminating spontaneous mutations (Birky and Walsh 1992; Khakhlova and Bock 2006). Variation in gene expression, amino acid composition, and genome organisation have further constrained the rate of molecular evolution of genes in the plastid genome (Robbins and Kelly 2023). Together, these factors have resulted in the plastid-encoded photosystem genes being some of the slowest evolving genes found in nature (Cardona, et al. 2019; Oliver, et al. 2021; Oliver, et al. 2023; Bouvier, et al. 2024). Given the low rate of molecular evolution it is unknown whether the photosystems have been adaptively evolving during the radiation of the angiosperms, which components of the photosystems have experienced higher or lower rates of adaptive evolution, and how this change has altered the properties of the photosystems themselves.

Here, we present a molecular evolution analysis at the single residue resolution of the plastid-encoded photosystem genes from 773 angiosperm species. We have focussed on the angiosperm clade as it contains the majority of crop plants, including food plants for direct human consumption, animal feed, biofuels, textiles, building materials and pharmaceuticals (Raskin, et al. 2002; Leff, et al. 2004; Mahapatra, et al. 2021). Consequently, understanding the evolution of their photosynthetic apparatus holds most potential to inform future crop improvement strategies. We uncover a widespread occurrence of hallmarks of adaptive evolution, namely positive selection and recurrent evolution, in all photosystem genes. Through analysis of the location and type of adaptive changes that occurred, we reveal how these changes are predicted to have had a significant impact on protein-protein interfaces but not on cofactor binding within the photosystems. Moreover, we reveal a concerted suite of changes that are modelled to destabilise photosystem II interactions with the D1 protein thereby reducing the energetic barrier for D1 turn-over and photosystem repair.

3.3 Results

3.3.1 There is substantial variation in extent of evolution between photosystem subunits

The chloroplast genome in angiosperms contains 20 genes encoding photosystem proteins, 5 from photosystem I (PSI; *psaA*, *psaB*, *psaC*, *psaI* and *psaJ*) and 15 from photosystem II (PSII; *psbA*, *psbB*, *psbC*, *psbD*, *psbE*, *psbF*, *psbH*, *psbI*, *psbJ*, *psbK*, *psbL*, *psbM*, *psbN*, *psbT* and *psbZ*) (Figure 1A). To analyse the molecular evolution of these genes a robust phylogeny was built from a concatenated multiple sequence alignment of all ubiquitously conserved protein-coding sequences from 773 complete angiosperm plastid genomes (see Methods). For each photosystem gene, ancestral sequence reconstructions were performed for all 771 internal nodes in the phylogenetic tree, and all non-synonymous substitutions were identified and mapped to branches in the tree.

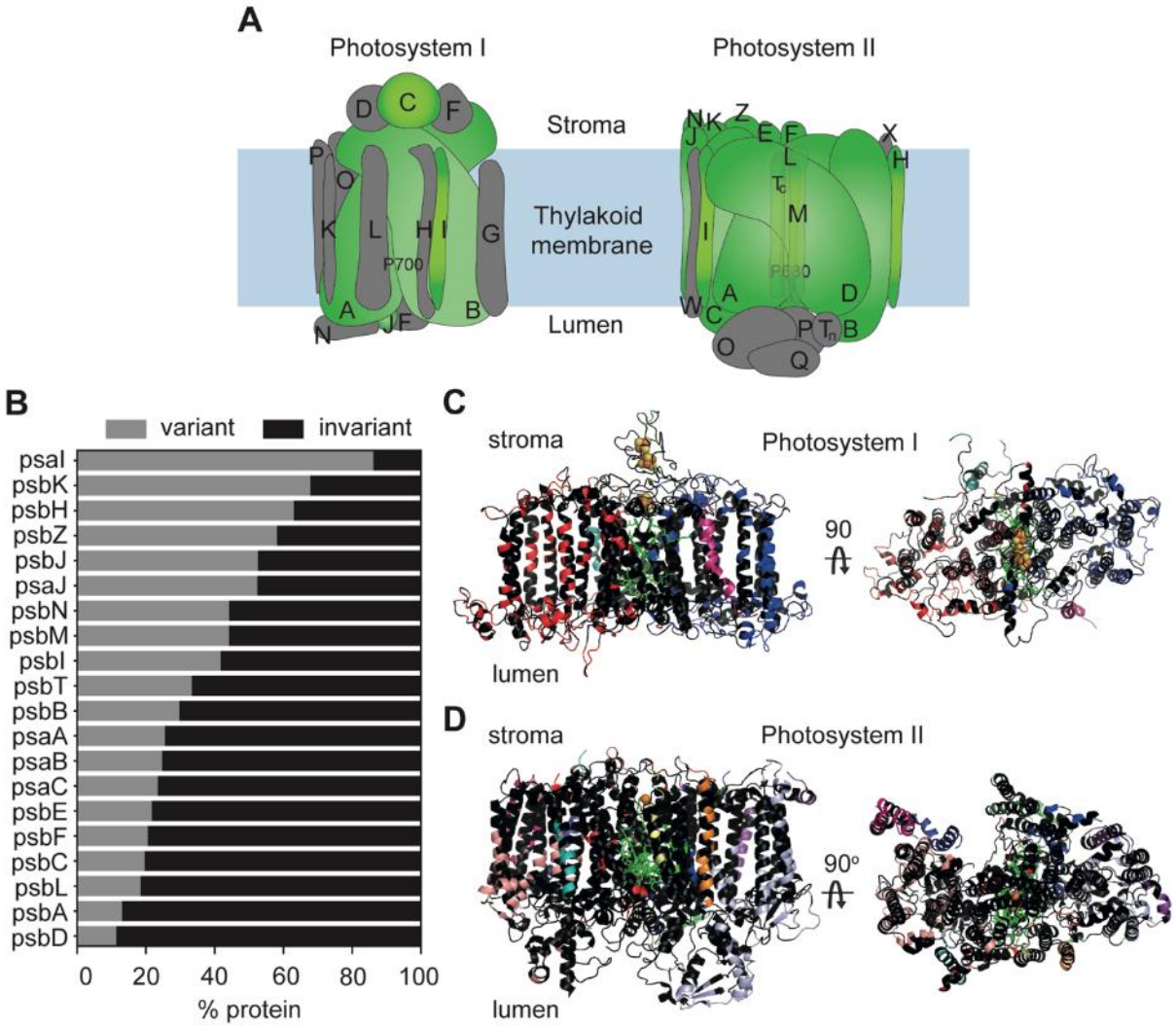


Figure 1. Highly conserved regions of the plastid-encoded photosystem proteins in angiosperms. **A**) Schematics of the photosystem I (left) and photosystem II (right) reaction centres with the proteins encoded by the 20 genes analysed in this study highlighted in green. Proteins in grey were not ubiquitously conserved in the plastid genome among the species analysed. **B**) Bar chart showing the percentage of each protein that was variant (grey) versus invariant (black) in the 773 angiosperm species dataset. Genes are ordered by decreasing percentage of protein that is variable from top to bottom. **C**) Crystal structure of the proteins analysed in photosystem I with invariant residues highlighted in black. Red, PsaA; blue, PsaB; dark green, PsaC; pink, PsaI and teal, PsaJ. For reference, the electron transfer pathway pigments are shown as green sticks and iron-sulfer centres as orange/yellow spheres. **D**) Crystal structure of the genes analysed in photosystem II with invariant residues highlighted in black. Red, PsbA; light blue, PsbB; salmon, PsbC; blue, PsbD; pale green, PsbE; dark green, PsbF; purple, PsbH; teal, psbI; dark blue, PsbK; green, PsbL; orange, PsbM; pale yellow, PsbT and pink, PsbZ. Pigments are shown as green sticks and the Fe^{2+} ion as an orange sphere.

Although 74% (2850/3870) of all residues exhibited no variation (d_N = 0, Figure 1B - D), 6732 amino acid substitutions were identified at the remaining 1020 residues, with substitutions occurring in all 20 photosystem proteins (Figure 2A). Moreover, there was a 17-fold difference between the photosystem protein with the largest and fewest amino acid substitutions per residue across the 773 species phylogeny, corresponding to 11 substitutions per residue for PsaI (encoded by *psaI*) and 0.65 substitutions per residue for D1 (*psbA*), respectively (Figure 2B). Thus, although the plastid-encoded photosystem genes have evolved slowly during angiosperm evolution, evolution has occurred and there is significant variation in evolutionary rate between photosystem components.

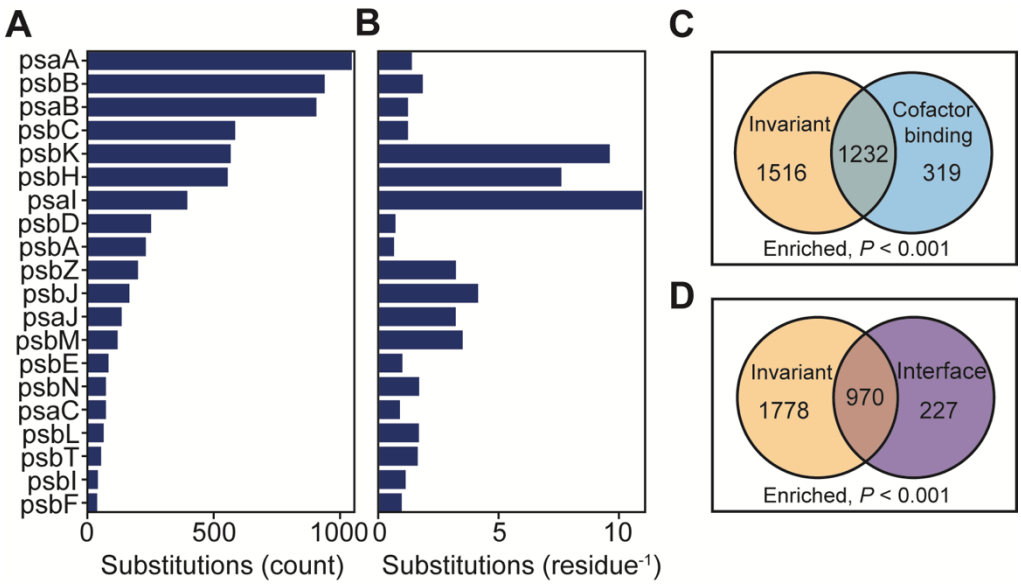


Figure 2. Amino acid substitutions in plastid-encoded photosystem genes during radiation of the angiosperms. **A**) Horizontal bar chart of the raw count of amino acid substitutions inferred to have occurred across the phylogeny for each protein. Proteins are ordered by decreasing number of amino acid substitutions from top to bottom, respectively. **B**) Horizontal bar chart of the substitution count per residue for each protein. Proteins are ordered as in **A**. **C**) Venn diagram showing the overlap between invariant residues (yellow) with cofactor binding residues (blue). **D**) Venn diagram showing the overlap between residues that are invariant (yellow) with those that are at the protein-protein interfaces between subunits (purple). *P*-values shown below Venn diagrams are the results of hypergeometric tests.

3.3.2 Despite strong functional constraint there is evidence of widespread adaptive evolution in the photosystems of angiosperms

To gain insight into the functional constraints and adaptive evolution of the photosystems in angiosperms we first analysed the sites that exhibited no variation in the 773 species dataset. This revealed that there was an enrichment of invariant sites in the cofactor binding residues of the photosystems (hypergeometric test, $P < 0.001$; Figure 2C). Furthermore, there was a significant enrichment of invariant sites at subunit interfaces in both photosystems (hypergeometric test, $P < 0.001$; Figure 2D). Thus, the requirement to bind cofactors and form inter-subunit interfaces has resulted in severe constraint on the extent of molecular evolution of the photosystem genes and can explain 58% of all invariant sites in these complexes.

Although sites involved in cofactor binding and subunit interfaces exhibited strong functional constraint, substantial variation is present. This raised the question as to whether this variation exhibits evidence of adaptive evolution, and if so, whether this change has altered the properties of the photosystems. As recurrent evolution and positive selection are widely regarded as signatures of adaptation to common selective constraints (Yang, et al. 2000; Wood, et al. 2005; Kapralov and Filatov 2007; Losos 2011; Stern 2013; Ord and Summers 2015), we sought to determine whether either of these hallmarks could be detected at variable sites in photosystem genes. Given that we had precisely mapped all amino acid substitutions to specific branches in the angiosperm phylogeny, it was possible to directly address the question as to whether any of the amino acid substitutions were recurrent in multiple branches of the phylogeny (a recurrent amino acid substitution is here defined as an amino acid substitution $X \rightarrow Y$, where $X \neq Y$, that has occurred more than once at a given site in a protein sequence at a frequency greater than expected by chance; see Methods). Remarkably, 54% (3620/6522) of all amino acid substitutions were recurrent and unlikely to have evolved by stochastic co-occurrence of random mutations in the absence of selection ($P < 0.001$, Figure 3A). These recurrent substitutions were located at 275 sites across all 20 proteins and comprised 386 unique patterns of amino acid substitution (Figure 3B, Table B.1). Similarly, there was strong evidence for pervasive positive selection acting at 35 sites in 13 genes (Figure 3C and D, Table B.2). Moreover, 30 of these 35 positively selected sites were also identified in the set of recurrent substitutions demonstrating a substantial overlap between these two hallmarks of adaptive evolution (hypergeometric test, $P < 0.001$; Figure B.1). Consistent with this, there is a significant depletion of recurrent substitutions at sites that are detected as evolving under purifying selection (hypergeometric test, $P < 0.001$; Figure B.1). Thus, there is evidence that 280 sites across all 20 plastid-encoded photosystem genes may have evolved adaptively during the radiation of the angiosperms.

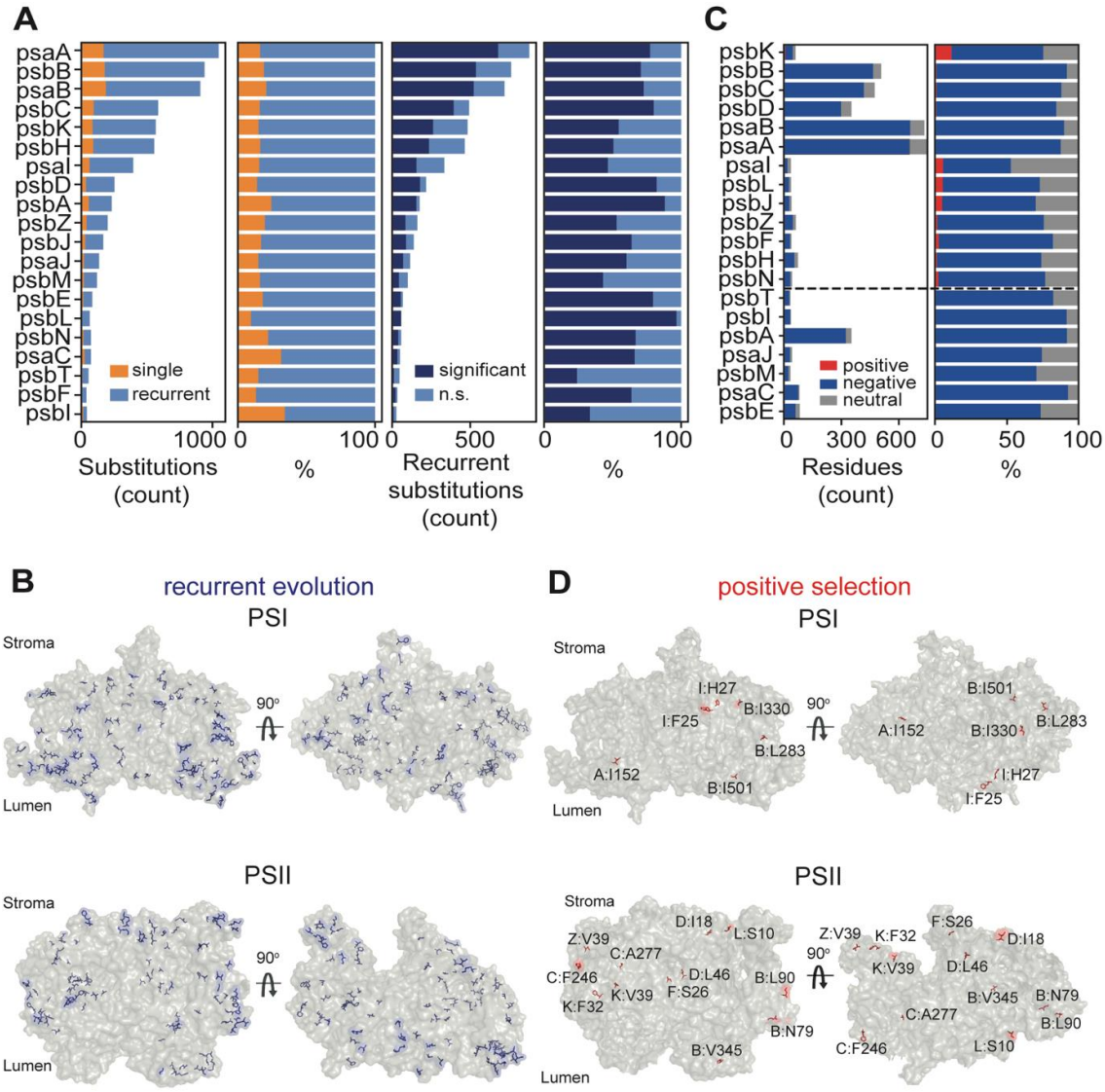


Figure 3. Adaptive evolution in the plastid-encoded photosystem proteins during the radiation of the angiosperms. **A**) Left two bar charts show amino acid substitution counts (left) and percentage of substitutions (right) that are single occurrence (orange) or recurrent (blue), i.e., a specific amino acid replacement that has occurred more than once at a given residue, respectively. Proteins are ordered by decreasing number of substitutions inferred to have occurred. Right two bar charts show the proportion of recurrent substitutions that occurred more frequently than expected by chance ($P < 0.001$, dark blue) as compared to those that did not (light blue). Raw counts are shown on the left and percentage of recurrent substitutions shown on the right. **B**) Bar charts showing the number (left) and percentage (right) of residues under positive selection (red), purifying selection (blue) or subject to neither (grey). Proteins are ordered by decreasing number of sites under positive selection from top to bottom. All proteins below the dashed line have no sites under positive selection. **C** and **D**) Structures of photosystem I and photosystem II with modelled sites of recurrent evolution highlighted in blue and sites subject to positive selection shown in red, respectively.

3.3.3 There is no significant difference in the extent of adaptation between photosystem complexes during angiosperm evolution

We next sought to determine whether adaptive evolution had been more prevalent in either the proteins of photosystem I or photosystem II during the evolution of the angiosperms. Among the 280 adaptively evolving sites, 127 were in photosystem I and 153 were in photosystem II. Given the number of adaptively evolving sites in a protein was strongly correlated to its length (R^2 = 0.67, P < 0.001, Figure B.2), we compared the number of adaptively evolving sites per residue between subunits of photosystem I versus photosystem II. Although this uncovered a 10-fold difference between the proteins with the highest and lowest levels of adaptive evolution corresponding to 0.250 adaptive sites per residue in PsaI and 0.026 adaptive sites per residue in PsbF, respectively (Figure 4A), there was no significant difference in the level of adaptive evolution between photosystem I and photosystem II subunits (t-test, P < 0.05; Figure 4B). Thus, the levels of adaptive evolution experienced by photosystem I and photosystem II has been equivalent during the radiation of the angiosperms.

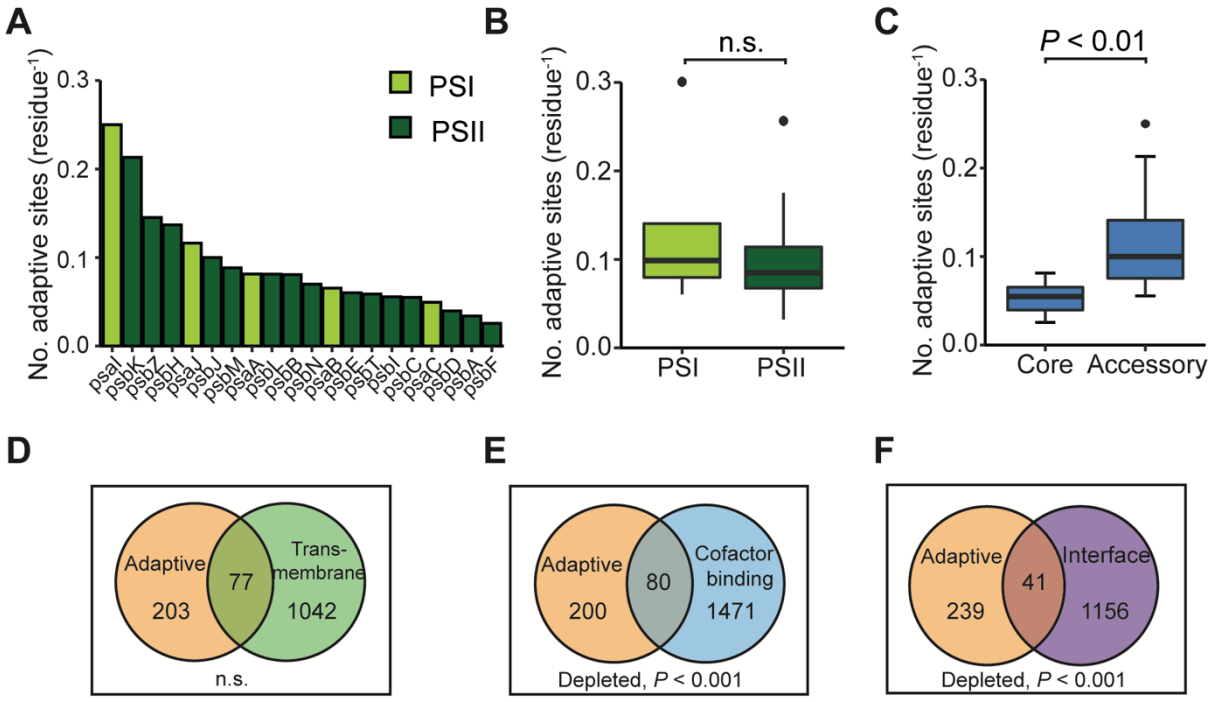


Figure 4. Analysing regional and structural enrichment of adaptive evolution in the photosystems. **A**) Bar chart showing the number of sites under adaptive evolution in the 20 photosystem proteins analysed corrected for protein length in decreasing order from left to right. Proteins in photosystem I (PSI) and photosystem II (PSII) are in light green and dark green, respectively. **B**) Boxplot showing data in (**A**) grouped by photosystem complex. Two-sample t-test, $P > 0.05$. **C**) Boxplot showing data in (**A**) grouped by photosystem region. Core proteins included PsaA, PsaB, PsaC, PsbA, PsbD, PsbC, PsbD, PsbE and PsbF. Peripheral proteins included PsaI, PsaJ, PsbH, PsbI, PsbJ, PsbK, PsbL, PsbM, PsbN, PsbT and PsbZ. *P*-value shown derived from two-sample t-test. **D-F**) Venn diagrams showing the overlap between adaptively evolving sites (orange) and transmembrane residues (green), cofactor binding residues (blue) and inter-subunit residues (purple), respectively. *P*-values shown are the results of hypergeometric tests.

3.3.4 Adaptive evolution has been limited in regions of the photosystems required for energy and electron transfer

Given that the extent of adaptive evolution was comparable between photosystem complexes, we next sought to investigate whether there was regional enrichment of adaptation within the photosystem structures. Specifically, we wanted to explore whether more adaptation had occurred in the core versus accessory proteins of the photosystem reaction centres, transmembrane versus extramembrane regions or functional regions such as cofactor binding or protein-protein interfaces. To test the former, core proteins were defined as subunits binding

the redox cofactors forming the electron transport chain (PsaA, PsaB, PsaC, D1 and D2), internal antenna proteins (PsbB and PsbC) and cytochrome b559 (PsbE and PsbF). Meanwhile, accessory proteins included the remaining 11 low molecular weight (<10 kDa) proteins that are on the periphery of the core reaction centre (PsaI, PsaJ, PsbH, PsbI, PsbJ, PsbK, PsbL, PsbM, PsbN, PsbT and PsbZ). This revealed adaptive sites were significantly more prevalent in the accessory proteins compared to the core proteins of the photosystem reaction centres, with a mean fold difference of 2.2 (t-test, $P < 0.01$; Figure 4C). Meanwhile, there was no significant enrichment of adaptive sites in either the transmembrane or extramembrane regions of the photosystem proteins (Figure 4D). Lastly, adaptive sites were significantly depleted in both cofactor binding and inter-subunit interface residues (hypergeometric test, $P < 0.001$; Figure 4E and F), making up 29% and 15% of adaptive sites, respectively. Thus, adaptive evolution has been constrained in regions of high functional constraint including the subunits responsible for energy and electron transfer and residues coordinating cofactors and protein-protein interfaces.

3.3.5 Adaptive evolution has weakened D1 protein interfaces during angiosperm evolution

Despite being depleted at cofactor binding and inter-subunit interface residues, we next sought to assess the impact of the adaptive substitutions that had occurred within these respective regions. To do this, we calculated the change in the interaction energy at the protein-cofactor and protein-protein interfaces, respectively, by analysing alterations in structural features and empirical energy terms at the interface upon substitution. We found none of the 80 adaptive substitutions at cofactor binding sites were predicted to significantly affect ligand binding (Table B.1). However, of the 41 sites at inter-subunit interfaces 10 substitutions were predicted to significantly stabilise their respective inter-subunit interface and 9 substitutions were predicted to significantly destabilise their respective inter-subunit interface (Table 1 and Table B.1). Of particular note was that all recurrent substitutions at protein-protein interfaces involving the D1 protein (*psbA*) were predicted to be destabilising (one-sample t-test, $P < 0.05$; Figure B.3 and Table B.1), weakening the association of this subunit with the remainder of the photosystem complex. Mapping these sites onto the photosystem II structure revealed that they were primarily located either in, or in contact with, the stromal DE-loop of D1 (*psbA*), a region of known importance for binding and positioning the bicarbonate ion and non-haem iron (Figure 5A). These adaptive changes weakened the interface of the core complex with D1 through a variety of mechanisms. For example, in the DE-loop the recurrent substitution E235A in D1 (*psbA*) results in the loss of polar contacts between E235 in D1 with N264 and W267 in the D2 (*psbD*) protein (Figure 5B). Similarly,

although it is not located in the DE-loop region, the L36V substitution in D1 (*psbA*) reduces hydrophobic contacts with F15 and F19 in the transmembrane helix of PsbI (*psbI*) (Figure 5C). Together, these results indicate that the photosystems in angiosperms have been adaptively evolving to make it easier to dissociate D1 from the core photosystem complex. This finding is consistent with D1 being the primary recipient of photodamage in photosystem II (Mattoo, et al. 1981; Aro, et al. 1993), having the highest turn-over rate of all subunits in photosystem II (Kettunen, et al. 1997; Li, et al. 2017; Forsythe, et al. 2022), and thus a requirement for continual removal and replacement of D1 to restore photosystem function.

Table 1. Adaptive substitutions that substantially affect protein-protein interface stability ($|\Delta\Delta G_{PPI}| \geq 0.5$ kcal.mol^{-1}).

Protein	Mutation	Protein-protein interface	DDG_{PPI} (kcal.mol^{-1})
PsaB	M640T	**B**:A	-6.15
PsbF	S26F	**F**:E	-2.7
PsaI	S25F	**I**:L	-1.67
PsbE	D68N	D:**E**:U	-1.53
PsaC	H71P	**C**:B	-1
PsbE	S72P	**E**:D	-0.66
PsbL	N10S	B:**L**:M	-0.57
PsaJ	I4L	**J**:F	-0.54
PsaB	N636T	**B**:A	-0.53
PsbB	C124T	**B**:H	-0.53
PsbL	S10N	B:**L**:M	0.57
PsbD	I18T	**D**:X	0.66
PsbE	P72S	**E**:D	0.66
PsbA	L36V	**A**:I	0.95
PsbT	K28T	A:**T**:D	1.36
PsbA	E235A	**A**:D	1.43
PsbA	E231Q	**A**:B	2.43
PsbB	A471S	**B**:D	4
PsaC	H71N	**C**:B	4.23

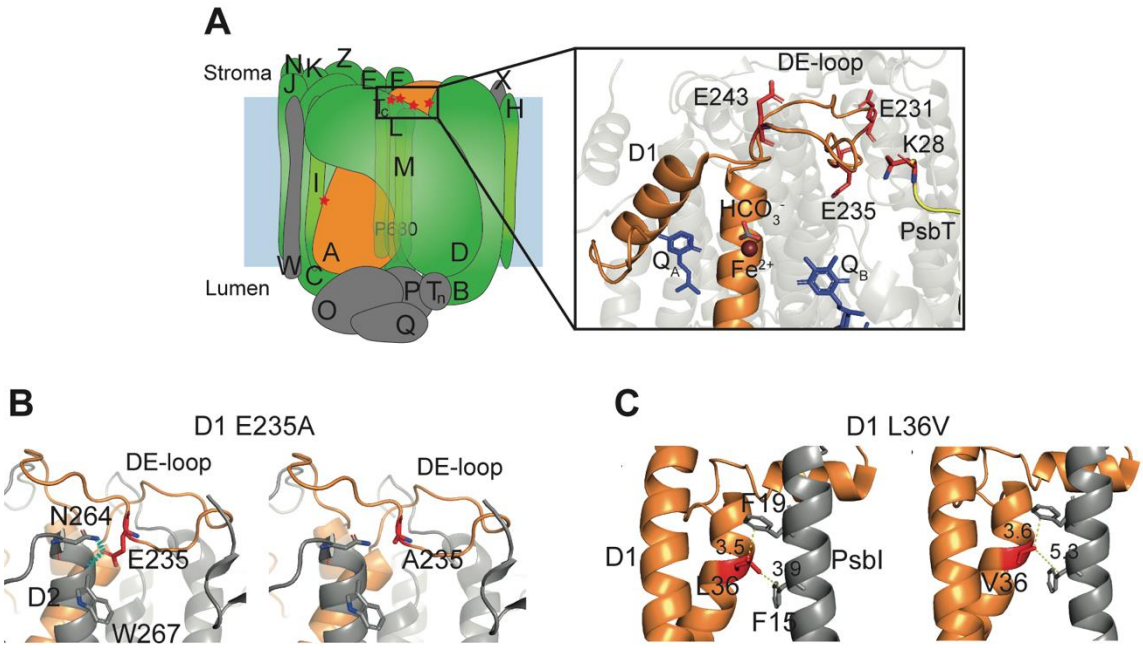


Figure 5. The position and impact of adaptive substitutions on inter-subunit interactions with the D1 (*psbA*) protein. **A**) Schematic mapping the location of substitutions that destabilise inter-subunit interactions with D1. The D1 protein is shown in orange for clarity and sites of substitutions are marked with a red star. A close-up view of the cluster of destabilising substitutions in/around the DE-loop of D1 in the 7OUI photosystem II structure is also provided. Residues at sites of destabilising substitutions are shown as red sticks. For reference, the two plastoquinones (Q_A and Q_B), Fe^{2+} and HCO_3^- involved in electron transfer are also shown. **B**) Loss of hydrogen bonds (dashed cyan lines) upon E → A substitution between D1 235 and D2 (*psbD*) N264 and W267. **C**) Reduced hydrophobic interactions upon L → V substitution at 36 in D1 with F15 and F19 in PsbI. Yellow dashed lines show distance between residues given in Å.

3.4 Discussion

Oxygenic photosynthesis underpins nearly all life on Earth. At the heart of the process are the photosystems, the molecular machines that facilitate the first step of photochemical conversion of energy from the sun into chemical energy that powers cellular metabolic processes. Here, we investigate the molecular evolution the plastid-encoded photosystem genes in angiosperms. We show that hallmarks of adaptive evolution, namely positive selection and recurrent evolution, are widespread in the photosystem subunits of angiosperms. Moreover, we reveal a concerted suite of changes are predicted to have destabilised the interaction of photosystem II with its D1 subunit reducing the energetic barrier for D1 turn-over and photosystem repair.

The finding that all adaptive substitutions at the inter-subunit interfaces with D1 (*psbA*) were predicted to destabilise the protein-protein interactions is consistent with extremely high turnover of the subunit (Kettunen, et al. 1997; Li, et al. 2017; Forsythe, et al. 2022), and the concomitant requirement to constantly remove and replace damaged D1 protein. Four of the five of these substitutions were found to be clustered in or around the stroma-exposed DE-loop in D1 (D1: E231Q, E235A, E243G and PsbT: K28T), which has been attributed two main functions. The first is that it provides two ligands (E244 and Y246) that indirectly coordinate the non-haem iron positioned between Q_A and Q_B *via* positioning a bicarbonate ion (Gao, et al. 2018). Secondly, the region between residues 238-248 in the DE-loop is a site of cleavage during D1 degradation, yielding the 23 kDa fragment (Greenberg, et al. 1987). Moreover, it has been postulated that a damage induced increase in mobility of the DE-loop may signal photosystem II repair (Forsman 2021). In this context it is interesting to note that three of the adaptive substitutions found to destabilise a D1 inter-subunit interface were located in the DE-loop region of D1, E231Q, E235A and E243G. Interestingly, it was shown that mutagenesis at two of these sites (E231D and E243Q) did not affect electron transfer but did substantially reduce the half-life of D1 (Mulo, et al. 1997). Given that electron transfer was not affected, the bicarbonate ion and non-haem iron must have still been successfully bound to the photosystem, suggesting the DE-loop has maintained a mostly correct conformation. If this is the case, the reduced half-life cannot be the result of an aberrant DE-loop conformation signalling immediate photosystem II degradation. An alternative hypothesis is that these substitutions result in less photodamage being required to initiate removal and replacement of D1, due to a reduction in the energy required to remove D1 from photosystem II. It is most parsimonious to assume that the adaptive substitutions we uncovered here act in a similar manner.

We also identified a further adaptive substitution, PsbT K28T, which is located adjacent to the DE-loop of D1 and predicted in this analysis to weaken the D1:PsbT protein-protein interface. Importantly, this adaptive substitution should not disrupt important hydrophobic interactions that occur between the neighbouring residues in PsbT (P27 and I29) with F239 in the DE-loop of D1. Without these hydrophobic interactions, the DE-loop has increased flexibility that leads to bicarbonate dissociation and slowed electron transfer between Q_A^- and Q_B (Forsman and Eaton-Rye 2021). This increased flexibility and resultant aberrant conformation of the DE-loop is thought to be the cause of an increase in turnover of D1 (Forsman and Eaton-Rye 2021). Thus, taken together the findings presented here indicate that the K28T substitution in PsbT is adaptive as it weakens the connection between PsbT and D1 without disrupting the DE-loop conformation, thereby enhancing D1 repair signalling in response to photodamage.

The question arises as to why reducing the energetic barrier for D1 turn-over and photosystem II repair would lead to an adaptive advantage. At any given time, the proportion of damaged photosystems is a function of the incident light, the activity of photoprotective mechanisms, and the rate of photosystem repair (Su, et al. 2023). Each of these factors interact to alter the proportion of photosystem II containing a non-functional D1 subunit and are thus in a non-productive photoinhibited state. Evolution by natural selection can act to modulate each of these factors to maximise the conversion of incident light energy into chemical energy and growth. As repair of photoinhibition is triggered by the detection of non-functional D1, reducing the energetic barrier for D1 turn-over would in turn increase the rate at which photoinhibited photosystem II complexes can be repaired. Furthermore, as photoinhibition only occurs when the rate of repair is unable to keep up with the rate of damage (Prasil 1992), the overall effect of reducing the energetic barrier to repair would be to have a lower proportion of non-functional photosystem II complexes, thereby increasing the conversion of incident light energy into chemical energy and growth.

The recurrent substitutions identified in this study are not uniformly distributed across the phylogeny. For example, E235A in D1 was identified in species from 15 different orders (Cornales, Arecales, Laurales, Chloranthales, Nymphaeales, Gentianales, Rosales, Poales, Ericales, Malvales, Caryophyllales, Zingiberales, Myrtales, Huerteales, and Vitales) while E243G in D1 was only observed in four orders (Brassicales, Apiales, Rosales, and Ericales). These differences in distribution may allow the effect of these substitutions to be experimentally distinguished. In this context, it will be interesting to perform a large-scale analysis of photoinhibition on species within these groups to understand how these changes have altered the function of photosystem II. Specifically, it will be interesting to determine if species harbouring these changes are less subject to photoinhibition across a range of different light intensities. Moreover, it will also be interesting to determine if introduction of these changes into species in which they have not evolved can result in a reduction in photoinhibition.

In this analysis we identified 7% of sites in the plastid-encoded photosystem proteins of angiosperms show hallmarks of adaptive evolution. Furthermore, we showed that many of these sites are located at regions of known functionality, with 29% of adaptively evolving sites involved in cofactor binding and 15% of sites at inter-subunit interfaces. However, while we provide biological explanations for substitutions at several adaptively evolving sites, the majority remain unexplained. For some of these unexplained sites, *in silico* prediction of the effect of substitution is impossible due to their location in regions that are not modelled in the

photosystem structures. Meanwhile, for the currently unexplained sites that are resolved in the photosystem structures their beneficial trait is not obvious from analysing the change to protein structure. Nonetheless, this study identifies several regions of the photosystems with hallmarks of adaptive evolution and further experimental investigation is required to elucidate the biological importance of these changes.

The analysis presented here focused on angiosperms as information gained here holds most potential to inform future crop improvement strategies. However, the analysis presented here will provide a foundation and point of comparison for investigation of adaptive evolution in photosystems of other lineages, such as algae or cyanobacteria, in future studies. In this context it is interesting to note that there are differences in the repertoire of genes that constitute the photosystems in these lineages. For example, the D1 gene is encoded by a multi-gene family in cyanobacteria (Mulo, et al. 2009). Here, different family members are differentially expressed according to environmental cues to adapt photochemical fitness to solar intensity (Vinyard, et al. 2013). Given the sub-functionalisation that has occurred, it may be that different gene family members are adapting along different evolutionary trajectories. For example, the high light adapted isoform may exhibit an analogous trajectory as described here for angiosperms, reducing the energetic barrier for D1 repair, while the low light adapted version may exhibit a different trend. Furthermore, comparison between recurrent substitutions in different Archaeplastida lineages may provide further insight into global trends not detectable through analysis of angiosperms alone.

Computational predictions of structural stability have been used repeatedly to inform the rational design and improvement of many proteins, for example (Yang, et al. 2007; Floor, et al. 2014; Goldenzweig, et al. 2016; Shirke, et al. 2016). However, these methods are not yet perfect (Marabotti, et al. 2021) and future re-evaluation of the data presented here with more sophisticated methods may uncover additional changes to the photosystems in angiosperms not reported here.

Our findings show that all plastid-encoded photosystem proteins exhibit hallmarks of adaptive evolution during the radiation of the angiosperms. By uncovering these past trajectories, and identifying key sites that have been altered by natural selection, these findings provide substantial new insight into the present and future evolution of the photosystems in angiosperms.

3.5 Methods

3.5.1 Phylogenetic tree inference

The plastid genome dataset and phylogenetic tree used in this study was the same as in (Robbins and Kelly 2023). In brief, the full set of sequenced plastid genomes was downloaded from the National Center of Biotechnology and Information (NCBI) in July 2021. Genomes were filtered so that those that did not contain the full complement of canonical plastid genes were removed, resulting in a dataset containing nucleotide, and corresponding protein sequences, for 69 plastid genes for 773 angiosperm species. For each of the 69 genes, a protein multiple sequence alignment was generated using MAFFT L-INS-I (Katoh and Standley 2013) and was used to guide an accurate nucleotide alignment using Pal2Nal (Suyama, et al. 2006). The nucleotide alignments were then trimmed to remove columns containing >90% gaps and concatenated together and used to identify the best-fit model of nucleotide evolution using IQ-TREE's inbuilt ModelFinder (Kalyaanamoorthy, et al. 2017). The best-fitting model of nucleotide evolution identified was GTR+F+R7. IQ-TREEs ultrafast bootstrapping method with 1000 replicates was then used to infer a maximum likelihood phylogenetic tree using the concatenated nucleotide sequence and GTR+F+R7 model of evolution (Hoang, et al. 2018). The tree was than manually rooted on the Nymphaeales clade, as they are a sister lineage to the rest of the sampled angiosperms in the dataset. The species tree is available to download from Zenodo at https://doi.org/10.5281/zenodo.13750849.

3.5.2 Inferring non-synonymous substitutions and calculating the rate of non-synonymous substitution

Among the 69 ubiquitously conserved genes above were 20 genes that encode proteins that participate in the photosystem reaction centres, 5 for photosystem I (*psaA*, *psaB*, *psaC*, *psaI* and *psaJ*) and 15 for photosystem II (*psbA*, *psbB*, *psbC*, *psbD*, *psbE*, *psbF*, *psbH*, *psbI*, *psbJ*, *psbK*, *psbL*, *psbM*, *psbN*, *psbT*, *psbZ*). Ancestral state reconstructions were generated for each of the 20 photosystem genes for every node in the maximum likelihood phylogenetic tree using the GTR+F+R7 model of evolution in IQ-TREE (Nguyen, et al. 2014). Non-synonymous substitutions at every internal branch in the tree were inferred by assessing all child-parent relationships in the phylogeny, excluding the Nymphaeles outgroup, using a bespoke Python script. These were translated into amino acid replacements at every residue in the protein alignment. Substitutions were tabulated in matrices where columns represent the parent residue identity and rows represent the child residue. These matrices are summarised in Table B.3. The gene trees, multiple sequence alignments and ancestral state sequences are available to download from Zenodo at https://doi.org/10.5281/zenodo.13750849. Site-wise

rates of non-synonymous mutation (d_N) were determined using the FEL method from HYPHY's software suite (Kosakovsky Pond and Frost 2005).

3.5.6 Identifying cofactor binding and protein-protein interface residues

Residues involved in cofactor binding and protein-protein interactions were identified using the 5L8R photosystem I from *Pisum sativum* (2.6Å resolution) (Mazor, et al. 2017) and 7OUI photosystem II from *Arabidopsis thaliana* (2.79Å resolution) (Graça, et al. 2021) structures. To identify cofactor binding residues, all pigments and cofactors associated with the 20 subunits analysed were identified in the photosystem structures. For photosystem I this included 84 chlorophyll a molecules, 2 phylloquinones, 16 beta-carotenes, 1 (3R,3'R,6S)-4,5-didehydro-5,6-dihydro-beta,beta-carotene-3,3'-diol, 4 1,2-dipalmitoyl-phosphatidyl-glycerole, 5 1,2-distearoyl-monogalactosyl-diglyceride, 3 digalactosyl diacyl glycerol, L, 1 Cl^-, 1 Ca^{2+} and 3 iron-sulfur centres. Meanwhile, for photosystem II this included 35 chlorophyll a molecules, 4 chlorophyll b molecules, 2 Ca^{2+}, 1 Cl^-, 11 beta-carotene, 1 HCO_3^-, 4 digalactosyl diacyl glycerol, 6 1,2-distearoyl-monogalactosyl-diglycerise, 2 pheophytin, 2 plastoquinone and 4 sulfaquinovosyldiacylglycerol molecules. Residues coordinating these pigments and cofactors were identified by conducting a search for residues with at least one heavy atom <4Å from an atom in a pigment or cofactor using the NeighbourSearch function in the Bio.PDB package (Hamelryck and Manderick 2003) (Figure B.4 and Table B.4). Protein-protein interface residues were identified using the EMPL-EBI PDBsum database (Figure B.5) (Laskowski, et al. 2018) (Table B.5).

3.5.7 Recurrent evolution and positive selection

Recurrent substitutions were defined as amino acid substitutions, i.e., X -> Y, where $X \neq Y$, that were inferred to have occurred more than once at a specific residue in the phylogeny. Given that recurrent substitutions could have arisen from chance a Monte Carlo approach was developed to assess whether they were likely to occur by chance given the inferred model of sequence evolution and the phylogenetic relationships between species encapsulated by the tree. To do this, the best-fitting model of protein sequence evolution was identified from the concatenated alignment of all 69 conserved plastid protein sequences and the species tree using IQ-TREE's inbuilt ModelFinder (Kalyaanamoorthy, et al. 2017). The best-fitting model identified was JTT+F+R7. Branch lengths for protein trees that were constrained to the topology of the species tree were then inferred for each of the 20 photosystem protein alignments using the JTT+F+R7 model of protein evolution. Protein sequences for extant species were then simulated for each of the 20 photosystem genes from the protein tree and ancestral node sequence using IQ-TREE's alignment simulator, with 1000 replicates. For each

of the simulated alignments, ancestral sequences were inferred, and amino acid substitutions counted as described above. Only the recurrent substitutions that had not occurred at the same or higher frequency in any of the 1000 simulations of protein evolution were deemed significant and a hallmark of adaptive evolution.

Sites under positive selection were identified using the FUBAR method from the HYPHY software package (Murrell, et al. 2013). Sites with a posterior probability that the rate of non-synonymous substitution is greater than the rate of synonymous substitution ($\beta > \alpha$) of greater than 0.9 were defined as under positive selection. Conversely, sites under purifying selection were identified as having a posterior probability that $\alpha > \beta$ greater than 0.9.

3.5.8 Modelling the effect of amino acid substitutions on cofactor binding and protein-protein interfaces

Single amino acid substitutions were introduced into the photosystem I (5L8R) and photosystem II (7OUI) structures using PyMol. To estimate the effect of a substitution on cofactor binding, the change in Gibb's free energy of binding ($\Delta G_{binding}$) between protein and ligand was calculated for the wild-type and mutated structures using the PRODIGY-LIG webserver (Vangone, et al. 2018). From these, the change in $\Delta G_{binding}$ between the mutated and wild-type structure ($\Delta\Delta G_{binding}$) can be calculated, whereby a negative value indicates an increased binding affinity, and *vice versa*. Similarly, to estimate the effect of substitution at a protein-protein interface, the change in Gibb's free energy of interaction (ΔG_{PPI}) was calculated using BAlaS for both the wild-type and mutated protein structures (Wood, et al. 2020). As before, the change in ΔG_{PPI} upon mutation ($\Delta\Delta G_{PPI}$) was then calculated. For both $\Delta\Delta G_{binding}$ and $\Delta\Delta G_{PPI}$ only values <-0.5kcal.mol-1 or >0.5kcal.mol-1 were considered to substantially alter interface interactions (Benevenuta, et al. 2022).

3.6 Acknowledgements

Work in SK's laboratory was supported by the European Union's Horizon 2020 research and innovation programme under grant agreement no. 637765 and by the Wellcome Trust under grant number 226598/Z/22/Z. E.H.J.R. is funded by the Biotechnology and Biological Sciences Research Council (BBSRC) (grant numbers BB/M011224/1 and BB/P003117/1). This research was funded in whole, or in part, by the BBSRC. For the purpose of open access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

3.7 Data Availability

All data is provided in the supplemental material or is available from Zenodo here https://doi.org/10.5281/zenodo.13750849.

3.8 Author Contributions

ER and SK conceived the study. ER conducted the analysis. ER and SK wrote the manuscript.

3.9 References

Aro E-M, Virgin I, Andersson B. 1993. Photoinhibition of photosystem II. Inactivation, protein damage and turnover. Biochimica et Biophysica Acta (BBA) - Bioenergetics 1143:113-134.

Asada K. 2006. Production and scavenging of reactive oxygen species in chloroplasts and their functions. Plant Physiology 141:391-396.

Barber J. 1998. Photosystem two. Biochimica et Biophysica Acta (BBA) - Bioenergetics 1365:269-277.

Benevenuta S, Birolo G, Sanavia T, Capriotti E, Fariselli P. 2022. Challenges in predicting stabilizing variations: An exploration. Front Mol Biosci 9:1075570.

Birky CW, Jr., Walsh JB. 1992. Biased gene conversion, copy number, and apparent mutation rate differences within chloroplast and bacterial genomes. Genetics 130:677-683.

Blanchard JL, Schmidt GW. 1995. Pervasive migration of organellar DNA to the nucleus in plants. J Mol Evol 41:397-406.

Blankenship RE, Hartman H. 1998. The origin and evolution of oxygenic photosynthesis. Trends in Biochemical Sciences 23:94-97.

Bouvier JW, Emms DM, Kelly S. 2024. Rubisco is evolving for improved catalytic efficiency and CO_2 assimilation in plants. Proceedings of the National Academy of Sciences 121:e2321050121.

Cardona T, Sánchez-Baracaldo P, Rutherford AW, Larkum AW. 2019. Early Archean origin of Photosystem II. Geobiology 17:127-150.

Chitnis PR. 2001. Photosystem I: Function and physiology. Annu Rev Plant Physiol Plant Mol Biol 52:593-626.

Eigenbrode JL, Freeman KH. 2006. Late Archean rise of aerobic microbial ecosystems. Proceedings of the National Academy of Sciences 103:15759-15764.

Floor RJ, Wijma HJ, Colpa DI, Ramos-Silva A, Jekel PA, Szymański W, Feringa BL, Marrink SJ, Janssen DB. 2014. Computational library design for increasing haloalkane dehalogenase stability. Chembiochem 15:1660-1672.

Forsman JA, Eaton-Rye JJ. 2021. The interaction between PsbT and the DE loop of D1 in photosystem II stabilizes the quinone-iron electron acceptor complex. Biochemistry 60:53-63.

Forsythe ES, Grover CE, Miller ER, Conover JL, Arick MA, 2nd, Chavarro MCF, Leal-Bertioli SCM, Peterson DG, Sharbrough J, Wendel JF, et al. 2022. Organellar transcripts dominate the cellular mRNA pool across plants of varying ploidy levels. Proc Natl Acad Sci U S A 119:e2204187119.

Fournier G, Moore K, Rangel L, Payette J, Momper L, Bosak T. 2021. The Archean origin of oxygenic photosynthesis and extant cyanobacterial lineages. Proceedings of the Royal Society B 288:20210675.

Frankel LK, Sallans L, Limbach PA, Bricker TM. 2013. Oxidized amino acid residues in the vicinity of Q_A and $Pheo_{D1}$ of the photosystem II reaction center: putative generation sites of reducing-side reactive oxygen species. PLoS One 8:e58042.

Gao J, Wang H, Yuan Q, Feng Y. 2018. Structure and function of the photosystem supercomplexes. Frontiers in Plant Science 9.

Goldenzweig A, Goldsmith M, Hill SE, Gertman O, Laurino P, Ashani Y, Dym O, Unger T, Albeck S, Prilusky J, et al. 2016. Automated structure- and sequence-based design of Proteins for high bacterial expression and stability. Mol Cell 63:337-346.

Graça AT, Hall M, Persson K, Schröder WP. 2021. High-resolution model of Arabidopsis Photosystem II reveals the structural consequences of digitonin-extraction. Scientific Reports 11:15534.

Greenberg BM, Gaba V, Mattoo AK, Edelman M. 1987. Identification of a primary in vivo degradation product of the rapidly-turning-over 32 kd protein of photosystem II. Embo j 6:2865-2869.

Hamelryck T, Manderick B. 2003. PDB file parser and structure class implemented in Python. Bioinformatics 19:2308-2310.

Hedges SB, Blair JE, Venturi ML, Shoe JL. 2004. A molecular timescale of eukaryote evolution and the rise of complex multicellular life. BMC Evolutionary Biology 4:2.

Hideg É, Barta C, Kálai T, Vass I, Hideg K, Asada K. 2002. Detection of singlet oxygen and superoxide with fluorescent sensors in leaves under stress by photoinhibition or UV radiation. Plant and Cell Physiology 43:1154-1164.

Hoang DT, Chernomor O, Von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: improving the ultrafast bootstrap approximation. Molecular biology and evolution 35:518-522.

Jansen RK, Cai Z, Raubeson LA, Daniell H, Depamphilis CW, Leebens-Mack J, Müller KF, Guisinger-Bellian M, Haberle RC, Hansen AK. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proceedings of the National Academy of Sciences 104:19369-19374.

Kalyaanamoorthy S, Minh BQ, Wong TK, Von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nature methods 14:587-589.

Kapralov MV, Filatov DA. 2007. Widespread positive selection in the photosynthetic Rubisco enzyme. BMC Evolutionary Biology 7:73.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution 30:772-780.

Kelly S. 2021. The economics of organellar gene loss and endosymbiotic gene transfer. Genome Biol 22:345.

Kettunen R, Pursiheimo S, Rintamäki E, Van Wijk KJ, Aro EM. 1997. Transcriptional and translational adjustments of *psbA* gene expression in mature chloroplasts during photoinhibition and subsequent repair of photosystem II. Eur J Biochem 247:441-448.

Khakhlova O, Bock R. 2006. Elimination of deleterious mutations in plastid genomes by gene conversion. Plant J 46:85-94.

Koch LG, Britton SL. 2008. Aerobic metabolism underlies complexity and capacity. J Physiol 586:83-95.

Kosakovsky Pond SL, Frost SDW. 2005. Not so different after all: A comparison of methods for detecting amino acid sites under selection. Molecular Biology and Evolution 22:1208-1222.

Laskowski RA, Jabłońska J, Pravda L, Vařeková RS, Thornton JM. 2018. PDBsum: Structural summaries of PDB entries. Protein Sci 27:129-134.

Leff B, Ramankutty N, Foley JA. 2004. Geographic distribution of major crops across the world. Global biogeochemical cycles 18.

Li L, Nelson CJ, Trösch J, Castleden I, Huang S, Millar AH. 2017. Protein degradation rate in Arabidopsis thaliana leaf growth and development. The Plant Cell 29:207-228.

Losos JB. 2011. Convergence, adaptation, and constraint. Evolution 65:1827-1840.

Mahapatra S, Kumar D, Singh B, Sachan PK. 2021. Biofuels and their sources of production: A review on cleaner sustainable alternative against conventional fuel, in the framework of the food and energy nexus. Energy Nexus 4:100036.

Marabotti A, Del Prete E, Scafuri B, Facchiano A. 2021. Performance of Web tools for predicting changes in protein stability caused by mutations. BMC Bioinformatics 22:345.

Martin W, Kowallik KV. 1999. Annotated English translation of Mereschkowsky's 1905 paper ‘Über Natur und Ursprung der Chromatophoren im Pflanzenreiche’. European Journal of Phycology 34:287-295.

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 U S A 99:12246-12251.

Mattoo AK, Pick U, Hoffman-Falk H, Edelman M. 1981. The rapidly metabolized 32,000-dalton polypeptide of the chloroplast is the "proteinaceous shield" regulating photosystem II electron transport and mediating diuron herbicide sensitivity. Proceedings of the National Academy of Sciences 78:1572-1576.

Mazor Y, Borovikova A, Caspy I, Nelson N. 2017. Structure of the plant photosystem I supercomplex at 2.6 Å resolution. Nature Plants 3:17014.

Mulo P, Sicora C, Aro EM. 2009. Cyanobacterial psbA gene family: optimization of oxygenic photosynthesis. Cell Mol Life Sci 66:3697-3710.

Mulo P, Tyystjärvi T, Tyystjärvi E, Govindjee, Mäenpää P, Aro EM. 1997. Mutagenesis of the D-E loop of photosystem II reaction centre protein D1. Function and assembly of photosystem II. Plant Mol Biol 33:1059-1071.

Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Kosakovsky Pond SL, Scheffler K. 2013. FUBAR: A Fast, Unconstrained Bayesian AppRoximation for Inferring Selection. Molecular Biology and Evolution 30:1196-1205.

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2014. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular Biology and Evolution 32:268-274.

Oliver T, Kim TD, Trinugroho JP, Cordón-Preciado V, Wijayatilake N, Bhatia A, Rutherford AW, Cardona T. 2023. The evolution and evolvability of photosystem II. Annual Review of Plant Biology 74:225-257.

Oliver T, Sánchez-Baracaldo P, Larkum AW, Rutherford AW, Cardona T. 2021. Time-resolved comparative molecular evolution of oxygenic photosynthesis. Biochimica et Biophysica Acta (BBA) - Bioenergetics 1862:148400.

Ord TJ, Summers TC. 2015. Repeated evolution and the impact of evolutionary history on adaptation. BMC Evolutionary Biology 15:137.

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

Planavsky NJ, Asael D, Hofmann A, Reinhard CT, Lalonde SV, Knudsen A, Wang X, Ossa Ossa F, Pecoits E, Smith AJB, et al. 2014. Evidence for oxygenic photosynthesis half a billion years before the Great Oxidation Event. Nature Geoscience 7:283-286.

Pospíšil P. 2016. Production of reactive oxygen species by photosystem II as a response to light and temperature stress. Front Plant Sci 7:1950.

Prasil O. 1992. Dynamics of photosystem II: mechanism of photoinhibition and recovery processes. Topics in photosynthesis 11:293-348.

Raskin I, Ribnicky DM, Komarnytsky S, Ilic N, Poulev A, Borisjuk N, Brinker A, Moreno DA, Ripoll C, Yakoby N, et al. 2002. Plants and human health in the twenty-first century. Trends in Biotechnology 20:522-531.

Robbins EHJ, Kelly S. 2023. The evolutionary constraints on angiosperm chloroplast adaptation. Genome Biology and Evolution 15.

Schwartz RM, Dayhoff MO. 1978. Origins of prokaryotes, eukaryotes, mitochondria, and chloroplasts. Science 199:395-403.

Shirke AN, Basore D, Butterfoss GL, Bonneau R, Bystroff C, Gross RA. 2016. Toward rational thermostabilization of Aspergillus oryzae cutinase: Insights into catalytic and structural stability. Proteins 84:60-72.

Smith DR. 2015. Mutation rates in plastid genomes: they are lower than you might think. Genome Biol Evol 7:1227-1234.

Stern DL. 2013. The genetic causes of convergent evolution. Nat Rev Genet 14:751-764.

Su J, Jiao Q, Jia T, Hu X. 2023. The photosystem-II repair cycle: updates and open questions. Planta 259:20.

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

Vangone A, Schaarschmidt J, Koukos P, Geng C, Citro N, Trellet ME, Xue LC, Bonvin AMJJ. 2018. Large-scale prediction of binding affinity in protein–small ligand complexes: the PRODIGY-LIG web server. Bioinformatics 35:1585-1587.

Vinyard DJ, Gimpel J, Ananyev GM, Cornejo MA, Golden SS, Mayfield SP, Dismukes GC. 2013. Natural variants of photosystem II subunit D1 tune photochemical fitness to solar intensity. J Biol Chem 288:5451-5462.

Wolfe KH, Li WH, Sharp PM. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc Natl Acad Sci U S A 84:9054-9058.

Wood CW, Ibarra AA, Bartlett GJ, Wilson AJ, Woolfson DN, Sessions RB. 2020. BAlaS: fast, interactive and accessible computational alanine-scanning using BudeAlaScan. Bioinformatics 36:2917-2919.

Wood TE, Burke JM, Rieseberg LH. 2005. Parallel genotypic adaptation: when evolution repeats itself. Genetica 123:157-170.

Yang DF, Wei YT, Huang RB. 2007. Computer-aided design of the stability of pyruvate formate-lyase from Escherichia coli by site-directed mutagenesis. Biosci Biotechnol Biochem 71:746-753.

Yang Z, Nielsen R, Goldman N, Pedersen A-MK. 2000. codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431-449.

Yao DCI, Brune DC, Vermaas WFJ. 2012. Lifetimes of photosystem I and II proteins in the cyanobacterium Synechocystis sp. PCC 6803. FEBS Letters 586:169-173.

Zhu A, Guo W, Gupta S, Fan W, Mower JP. 2016. Evolutionary dynamics of the plastid inverted repeat: the effects of expansion, contraction, and loss on substitution rates. New Phytologist 209:1747-1756.

Chapter 4:

RECUR: Identifying recurrent amino acid substitutions from multiple sequence alignments

Elizabeth HJ Robbins, Yi Liu and Steven Kelly

4.1 Abstract

Identifying recurrent amino acid substitutions is important to multiple aspects of biological research - from understanding the molecular basis of convergent phenotypes, to identifying the causative sequence changes that give rise to disease and antibiotic resistance. Here, we present RECUR, a method for identifying recurrent amino acid substitutions from multiple sequence alignments. It is fast, easy to use, and scalable to thousands of sequences. We demonstrate the utility and performance characteristics of the method on a large data set of SARS-CoV-2 surface glycoprotein sequences – identifying widespread recurrent evolution throughout the protein and significant enrichment of sites undergoing recurrent evolution in the region responsible for receptor binding, the S1 subunit. A standalone implementation of the algorithm is available under the GPLv3 licence at https://github.com/OrthoFinder/RECUR.

4.2 Introduction

Recurrent evolution arises when the same biological innovation evolves independently on multiple occasions. It is found across the tree of life and can be observed at the phenotypic (organismal, physiological, cellular, biochemical, or molecular) or genotypic level (Stern 2013). To date, organismal phenotypic recurrence has been the most studied (Maeso, et al. 2012). For example, the independent evolution of wings in Pterosaurs, insects, bats, and birds (Hunter 2007) and a high-resolution camera-like eye in vertebrates, cephalopods, and arthropods (Nilsson 2013). However, the increasing availability of DNA sequence data, and advanced phylogenetic software to analyse them, is making the study of recurrence at the genotypic level more tractable.

Much attention has been placed on studying recurrent protein evolution in the context of adaptation. This has uncovered a plethora of adaptive substitutions that have repeatedly evolved to alter protein function to better suit environmental conditions (Bull, et al. 1997; Christin, et al. 2007; Christin, et al. 2008; Yokoyama, et al. 2008; Christin, et al. 2009; Liu, et al. 2010; Projecto-Garcia, et al. 2013; van Ditmarsch, et al. 2013; He, et al. 2021; Robbins and Kelly 2024) and defend against toxic molecules (including antimicrobials, antivirals, herbicides and pesticides) (Dobler, et al. 2012; Feldman, et al. 2012; Toprak, et al. 2012; Zhen, et al. 2012; Brodie III and Brodie Jr 2015; Ujvari, et al. 2015; Iketani, et al. 2023). Furthermore, analyses of recurrent mutations have also helped to elucidate the molecular basis of disease (Ingram 1957; Cutting, et al. 1990; Davies, et al. 2002; Olivier, et al. 2010; Jänne, et al. 2022).

Despite the widespread research analysing recurrent substitutions in protein sequences, no automated methods exist to identify these substitutions. Here, we present RECUR, a phylogenetic tool designed to identify recurrent substitutions that have occurred in a protein or codon multiple sequence alignment. Additionally, RECUR simulates sequence evolution to assess the probability that the frequency of a recurrent substitution has not occurred by chance but rather in the presence of a selective force. We exemplify the use of RECUR on a SARS-CoV-2 surface glycoprotein multiple sequence alignment.

4.3 Results

4.3.1 Workflow and overview of RECUR.

RECUR was designed to identify recurrent amino acid substitutions that have arisen during the evolution of a set of protein sequences. RECUR requires as input a multiple sequence alignment of protein sequences or an equivalent codon alignment. The method returns 1) the complete set of substitutions observed in the evolutionary history of that alignment. 2) The

complete set of substitutions that are recurrent including whether any reversions (from the derived back to the ancestral state) have also taken place. 3) A statistical analysis of the recurrent substitutions to identify those that are unlikely to have arisen by chance given the number of sequences, their phylogenetic relationship, and the underlying model of sequence evolution.

An overview of the RECUR workflow is provided in Figure 1. In brief, a phylogenetic tree is constructed from the input multiple sequence alignment and ancestral sequences are inferred for all internal nodes of the phylogenetic tree. All amino acid substitutions are then identified by assessing changes in the protein sequence along every branch in the phylogeny. Simulated multiple sequence alignments are then generated using the inferred ancestral sequence, the inferred tree, and the best-fitting model of sequence evolution, and then evaluated to identify recurrent substitutions that are unlikely to have arisen by chance (see Methods).

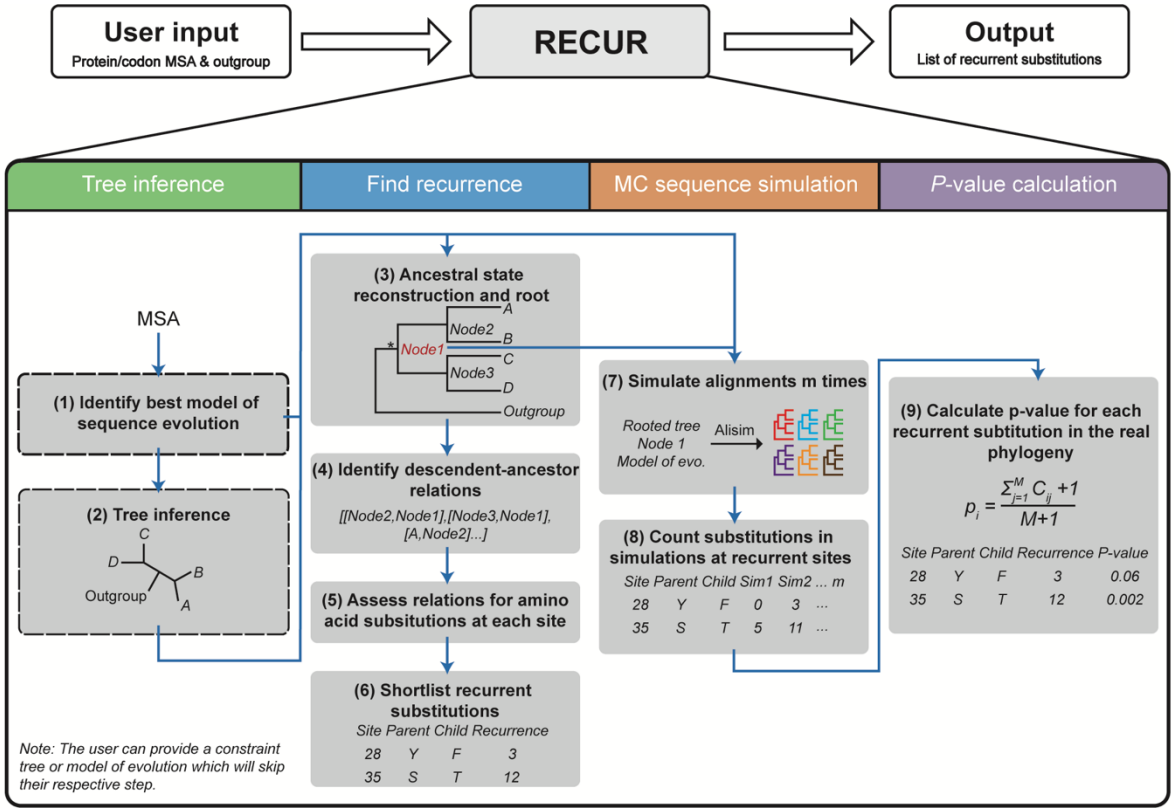


Figure 1. Overview of RECUR. The input (protein or codon) multiple sequence alignment is used to identify the best model of evolution (1), build a maximum likelihood phylogeny (2) and construct ancestral sequences (3). A user-defined outgroup is then used to root the tree (3) and assess each branch in the phylogeny for amino acid substitutions at each site in the protein sequence (4 and 5), from which recurrent substitutions are extracted (6). Sequence evolution is then simulated *M* times (default 1000) using the topology of the inferred phylogeny, best model of sequence evolution, and the root sequence of the subtree of interest (*) which excludes the outgroup sequences (7). The recurrence of substitutions in the list output in (6) is then assessed in each of the simulated alignments (8), and a *p*-value is calculated from the number of times the recurrence in simulations equals or exceeds that observed in the real alignment (9).

4.3.2 RECUR is fast and readily scalable to thousands of sequences

To demonstrate the runtime characteristics of the method we evaluated RECUR's runtime and memory usage across multiple sequence alignments ranging from 8 to 1024 sequences (Figure 2). Each alignment comprised a randomly selected set of non-redundant SARS-CoV-2 surface glycoprotein (S protein) sequences (see Methods). Each alignment was then subject to analysis with RECUR using eight threads with multiple different input options: 1) with no additional information, 2) with a pre-computed constraint tree, 3) with a pre-computed best-fitting model of sequence evolution, and 4) with both a pre-computed constraint tree and a

pre-computed best-fitting model of sequence evolution. These different input options were performed to illustrate the scalability and runtime characteristics of each step. They were also performed to illustrate different ways of running RECUR when either the tree or model of sequence evolution are known prior to running a RECUR analysis.

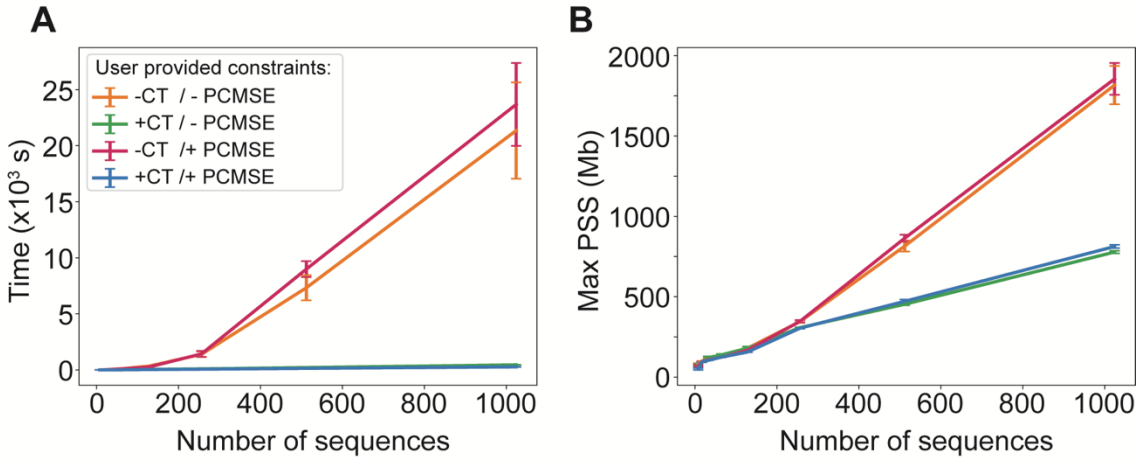


Figure 2. Analysing runtime and memory usage characteristics of RECUR. **A**) Line graph showing the relationships between runtime and number of sequences in the protein multiple sequence alignment. The different lines indicate different input variations; orange, no pre-computed constraint tree and no pre-defined model of sequence evolution (-CT / -PCMSE); pink, a pre-defined model of sequence evolution but no pre-computed constraint tree (-CT / +PCMSE); green, a pre-computed constraint tree but no pre-defined model of sequence evolution (+CT / -PCMSE) and blue, a pre-computed constraint tree and a pre-defined model of sequence evolution (+CT / +PCMSE). **B**) Line graph showing the relationship between maximum proportional set size (PSS) with the number of sequences. Lines coloured as in **A**.

The runtime is almost entirely consumed by the tree inference step in the analysis (Figure 2A). In contrast, the memory usage (proportional set size, PSS) of the internal RECUR algorithm, which excludes tree inference, increased linearly with the number of sequences analysed (Figure 2B). Furthermore, only when >250 sequences were analysed did tree inference significantly increase memory consumption.

4.3.3 RECUR identifies widespread recurrent evolution during the evolution of the surface glycoprotein of SARS-CoV-2

To demonstrate the utility of RECUR, we applied RECUR to the complete set of non-redundant SARS-CoV-2 S protein sequences derived from human hosts present in NCBI (n = 123,717, see Methods). This sequence was chosen as an illustrative example for several reasons. 1) There is general interest in the evolution of this protein sequence as it is the focus of many vaccines. 2) The surface location and biological role of the protein means that the sequence has been subject to selection for promoting transmissibility and immune system evasion (Carabelli, et al. 2023) and thus is likely to have experienced recurrent evolution (Korber, et al. 2020; Wilkinson, et al. 2022). 3) There is an abundance of sequence data readily available in NCBI which can be subject to analysis. Application of RECUR to this large sequence dataset identified 135,919 amino acid substitutions to have occurred across 78% (1267/1630) of sites during the radiation of the S protein sequences (Supplementary File C1). Of these, 98% (132,621/135,919) were recurrent substitutions that occurred across 1181 sites and involved 6,248 distinct substitution patterns (Supplementary File C2). The locations of these recurrent substitutions are mapped onto the domains of the S protein sequence (Figure 3A and B). Thus, there has been widespread occurrence of recurrent evolution along the length of the protein in all annotated protein domains.

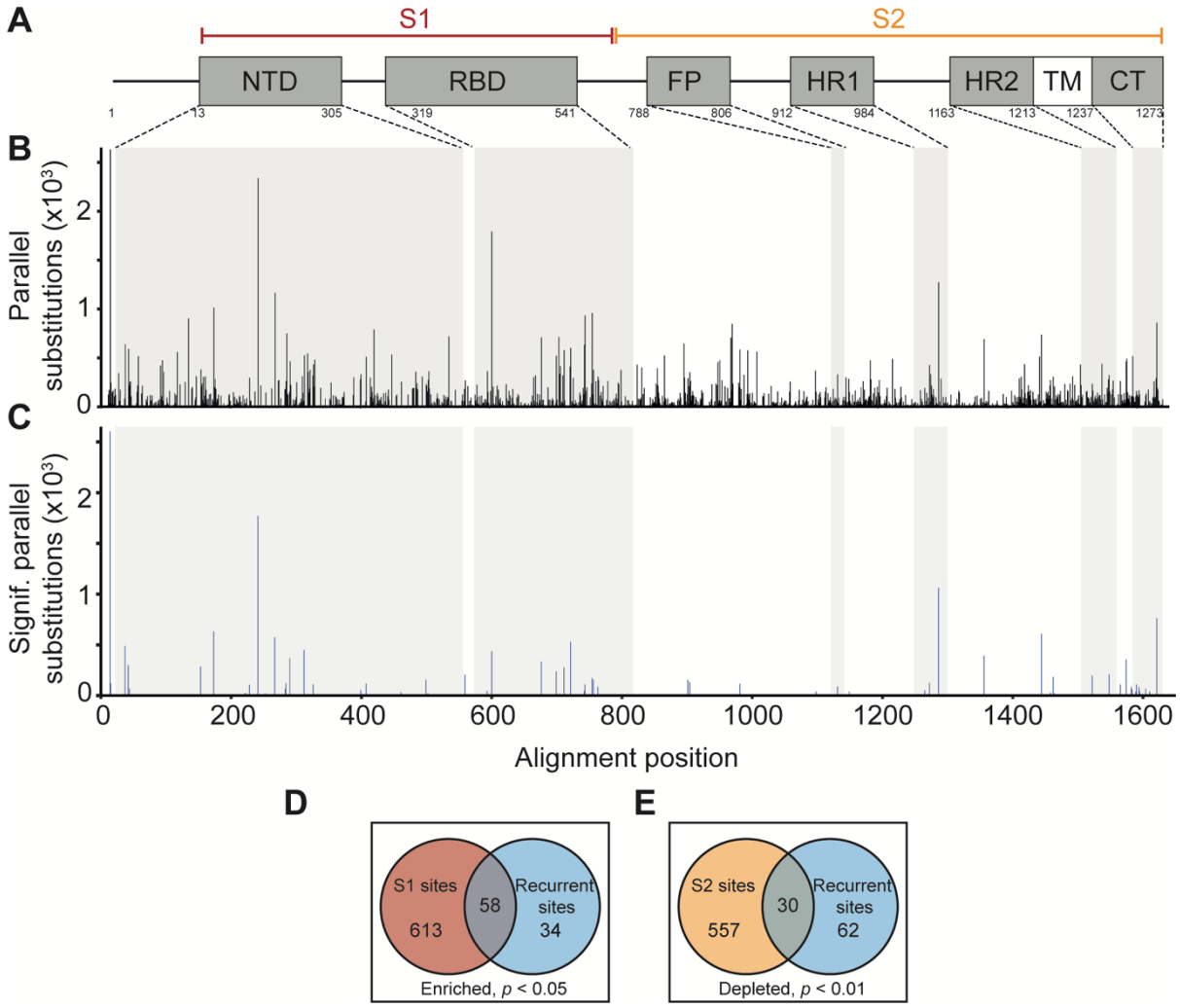


Figure 3. Analysing recurrence in the SARS-CoV-2 surface glycoprotein. **A**) Map showing positions of the protein domains (accession: QHD43416.1) and how they relate to alignment position. Domains include N-terminal domain (NTD), receptor binding domain (RBD), fusion peptide (FP), heptapeptide repeat sequence 1 (HR1), heptapeptide repeat sequence 2 (HR2), transmembrane domain (TM) and cytoplasm domain (CT). The position of the larger S1 (residues 14 - 685) and S2 (residues 686 - 1273) subunits are also shown. **B** and **C**) Bar charts showing the number of recurrent substitutions and significant recurrent substitutions (those not likely the result of stochastic mutation) at each site in the protein alignment. Protein domains are highlighted in grey. **D** and **E**) Venn diagrams showing the results of hypergeometric tests over the overlap between the sites in the S1 and S2 regions with significant recurrence sites, respectively.

To determine whether these recurrent substitutions were unlikely to occur by chance given the number of sequences, the model of sequence evolution, and the underlying phylogenetic relationship between the sequences, RECUR performs a Monte Carlo simulation of sequence evolution. This analysis revealed that 12% (16,051/132,621) of the recurrent substitutions identified above, which occurred across 92 sites and included 130 patterns of substitution, are unlikely to have occurred with such frequency by chance, as the same or larger frequency of occurrence was not replicated in any of the simulated alignments (Figure 3C).

The top 10 most recurrent significant substitutions are shown in Table 1. Six of these substitutions have occurred in the exposed S1 region of the S protein. These include D142G, T95I, Q146K, L18F and L176F in the N-terminal domain and F456L in the receptor binding domain (Table 1). All of these substitutions have previously been associated with altering viral infectivity or immune escape (Table 1). For example, the L18F substitution is known to reduce neutralization for multiple antibodies (Harvey, et al. 2021; McCallum, et al. 2021). Notably, the most recurrent substitution, L5F, was located in the signal peptide. However, it has been suggested that this specific substitution may be the result of recurring sequencing errors (De Maio, et al. 2020; Korber, et al. 2020; Grabowski, et al. 2021).

Table 1. The top 10 most significantly recurrent amino acid substitutions in the SARS-CoV-2 surface glycoprotein.

Substitution*	Recurrence	Reversion frequency**	Protein domain	Alignment position	Site annotation	Refs.
L5F	2196	409	Signal Peptide	14	May result from sequencing artefacts	(De Maio, et al. 2020; Korber, et al. 2020)
D142G	1761	443	NTD	241	Immune escape and increased viral load	(Shen, et al. 2021)
N950D	1037	150	HR1	1286	Membrane fusion	(Furusawa, et al. 2023)
V1264L	708	55	CT	1621	May affect subcellular trafficking	(Yang 2021)
T95I	630	271	NTD	173	Immune escape and increased viral load	(Shen, et al. 2021; Wilkinson, et al. 2022)
V1104L	608	38	-	1444	Protein stability	(Chand, et al. 2020)
Q146K	573	26	NTD	267	Immune escape	(McCallum, et al. 2021; Wang, et al. 2023)
F456L	529	56	RBD	721	Immune escape	(Dadonaite, et al. 2024)
L18F	490	92	NTD	37	Immune escape	(Gupta, et al. 2021; McCallum, et al. 2021)
L176F	450	31	NTD	312	Immune escape	(Wang, et al. 2023)

* Position referenced to QHD43416.1 sequence.

** Reversion frequency is the recurrence that the derived amino acid was substituted to the ancestral amino acid.

Finally, we sought to explore the position of these significant recurrent substitutions within the S protein. Specifically, we wanted to examine whether there was regional enrichment of recurrent evolution in the S1 subunit, which is responsible for receptor binding, compared to the S2 subunit, which is responsible for membrane fusion (Figure 3A). This revealed a significant enrichment in the number of sites that have undergone recurrent evolution within the S1 subunit (hypergeometric test, $p < 0.05$; Figure 3D). Consistent with this, there was a significant depletion in the number of recurrent sites located in the S2 subunit (hypergeometric test, $p < 0.01$; Figure 3E). This finding supports the proposed roles of the two subunits, with S1 subunit being more exposed on the viral surface, making it more accessible to the host immune system and thus subject to selection to adapt and change, while S2 subunit likely experiences selection to preserve function in the absence of a strong pressure to evade immune detection.

4.4 Discussion

Here, we present RECUR, a tool that automatically detects recurrent amino acid substitutions from a protein (or equivalent codon) multiple sequence alignment. We show that the method is scalable to thousands of sequences and demonstrate the utility of the method though analysis of 123,717 surface glycoprotein (S protein) sequences from SARS-CoV-2.

The RECUR method makes extensive use of the IQ-TREE 2 software package (Minh, et al. 2020). IQ-TREE 2 was chosen over other phylogenetic inference software for several reasons. First, IQ-TREE 2 is a mature, well-developed and thoroughly benchmarked software package that has the capability to carry out many steps required in the RECUR algorithms (Hoang, et al. 2017; Zhou, et al. 2017; Minh, et al. 2020; Ly-Trong, et al. 2023). These steps include model selection, tree inference, ancestral state reconstruction, and alignment simulation. It is not possible to achieve all of these steps using comparable software such as RAxML (Stamatakis 2014) or PhyML (Guindon and Gascuel 2003). Thus, use of IQ-TREE 2 reduces the number of dependencies required by RECUR. Second, IQ-TREE performs very well against all competitor methods in terms of both speed and accuracy (Zhou, et al. 2017). Third, IQ-TREE 2 can analyse alignments with thousands of sequences allowing RECUR to also scale to this number of sequences.

Topological irreproducibility in maximum likelihood phylogenetic inference is a known issue (Shen, et al. 2020). This arises from subtle differences in computational processing, such as seed number, number of threads or processor type, which can lead to variations in tree

topology between repeated runs. While RECUR exhibits this effect when these parameters change, identical outputs are consistently produced when run on the same machine with a fixed input. Alternatively, users can provide a constraint tree enabling RECUR to use a fixed topology. This provides an additional benefit of leveraging more information to infer the constraint tree than may be available in the alignment under consideration. For example, the user may wish to constrain the topology of the tree to match a known species tree when analysing orthologous sequences derived from those species. This approach was taken when analysing recurrence within the photosystem proteins encoded in the plastid genome (Robbins and Kelly 2024).

The runtime of RECUR is influenced by the number of sequences, the sequence length, and the extent of recurrence present in the sequences under consideration. Larger numbers of sequences, longer alignments, and higher rates of recurrence each result in increased computational demand. The S protein chosen to exemplify the method has a high mutation rate, is substantially longer than the average eukaryotic or bacterial protein (S protein length = ~1300, average prokaryotic protein length = ~300, average eukaryotic protein length =~ 500 (Tiessen, et al. 2012)), thus runtimes presented in this study represent the upper range of what can be expected from RECUR.

The application of RECUR to the complete set of non-redundant S protein sequences derived from a human host revealed widespread recurrent evolution across the length of the protein. Furthermore, we uncovered a significant enrichment of sites that have undergone recurrent evolution in the S1 subunit. This aligns with the functional and structural differences between these two regions of the S protein. The S1 subunit, contains the receptor binding domain and is essential for viral attachment to the angiotensin-converting enzyme 2 (ACE2) receptor, making it a hotspot for adaptive evolution as the virus evolves to concurrently optimise binding affinity and evade the immune system (Tian, et al. 2021). In addition, the S1 subunit contains the N-terminal domain, which plays a role in antigenicity, and hence drives adaptive evolution to repeatedly develop mechanisms for immune escape (McCallum, et al. 2021). Although depleted, sites under recurrent evolution were also identified in the S2 subunit. Given its role in membrane fusion, substitutions in the S2 region may impact the capacity for membrane fusion, as seen in (Kumar, et al. 2023). Additionally, these substitutions could benefit the virus through facilitating immune escape through allosteric effects on S1 epitopes (Kumar, et al. 2023).

In conclusion, RECUR is a fast, easy to use, and comprehensive tool for identifying amino acid substitutions from multiple sequence alignments. We anticipate that RECUR will serve as a valuable resource for understanding the molecular basis of convergent phenotypes as well as generating hypotheses and guiding experimental testing of protein function.

4.5 Materials and methods

4.5.1 Implementation

RECUR is a Python application that leverages IQ-TREE 2 (Minh, et al. 2020) as an external dependency and requires Python 3.9 or higher. The application utilizes DendroPy (Sukumaran and Holder 2010) to extract parent and child sequence pairs from rooted trees in Newick format. RECUR achieves high performance through the use of NumPy (Harris, et al. 2020) and multiprocessing, particularly in key steps such as the substitution analysis of Monte Carlo simulated protein sequences. In this process, parent-child sequence pairs are first converted into numerical representations and then into NumPy arrays for element-wise comparison at each residue site. Multiprocessing is employed to further speed up these operations. RECUR includes a Linux version of the IQ-TREE 2 binary, enabling it to run as a standalone application on Linux systems. For Windows and macOS users, IQ-TREE 2 must be preinstalled. Regardless of the operating system, it is recommended to run RECUR within a conda environment or a Docker container for compatibility. Full instructions on the installation and implementation of RECUR can be found at https://github.com/OrthoFinder/RECUR.

4.5.2 Tree inference and model selection

RECUR takes a protein or codon multiple sequence alignment as input. If the user does not specify a desired model of sequence evolution, then the alignment is evaluated to identify the best-fitting model of sequence evolution using ModelFinder implemented in IQ-TREE 2 (Kalyaanamoorthy, et al. 2017). Similarly, if the user does not specify a pre-computed tree, then a maximum likelihood phylogenetic tree is built using the model of sequence evolution above and IQ-TREE 2's ultrafast bootstrapping method with 1000 replicates (Hoang, et al. 2017). The tree is then rooted on a user-defined outgroup which can comprise a single sequence or a list of outgroup sequences. In the case where the supplied outgroup sequences are not monophyletic in the tree, the smallest possible clade comprising the full set of outgroup sequences will be selected as the outgroup.

4.5.3 Identifying recurrent amino acid substitutions

Ancestral sequence reconstructions are inferred for every node in the phylogenetic tree using the using the specified or computed (as appropriate) model of sequence evolution and the specified or inferred (as appropriate) maximum likelihood tree using the ancestral state reconstruction method implemented in IQ-TREE 2 (Minh, et al. 2020). To correctly infer gaps in ancestral sequences (which IQ-TREE 2 cannot currently implement) the multiple sequence alignment is converted into binary sequences, with 0 and 1 representing gapped and non-gapped sites, respectively. As with the biological sequences, the ancestral reconstructions of the gapped state for each alignment position is inferred for every node in the phylogenetic tree using the specified or inferred (as appropriate) maximum likelihood tree and the general time reversible model for binary data (GTR2) which allows for unequal state frequencies. The positions of the inferred gaps, in the ancestral sequences are then mapped onto the biological sequences to complete ancestral sequence reconstruction with indel inference.

Once the ancestral sequence reconstruction is complete, if a codon alignment was provided, the ancestral and extant sequences are then translated into protein sequences using the standard genetic code as default. The user can also specify alternative NCBI genetic codes to translate DNA sequences if required. Amino acid substitutions at each site in the alignment are then tabulated by analysing the residue identity for all parent-child transitions at every branch in the subtree that excludes outgroup sequences. A list of recurrent substitutions, i.e., substitutions that occurred more than once at a given site during the evolution of the sequence set, is then compiled.

4.5.4 Identification of significant recurrent sites

The phylogeny, model of sequence evolution, and root sequence of the subtree of interest are then used to simulate sequence evolution *M* times (default 1000) using the *alisim* function in IQ-TREE 2 (Ly-Trong, et al. 2023). As above, the gaps are mapped onto each simulated alignment. This prevents the overestimation of substitution recurrence due to incorrect placement of indels. The amino acid substitutions that have occurred in each of the simulated sequence alignments are then identified using the method outlined above. For each of the recurrent substitutions identified in the real alignment, RECUR assesses the number of times that the same substitution has occurred at the same or greater frequency in the set of simulated alignments. A *p*-value describing the probability that the recurrence of that site specific amino acid substitution having occurred by chance (p_i) is then assigned using the following equation:

$$p_i = \frac{\sum_{j=1}^{M} C_{ij} + 1}{M + 1}, \qquad i = 1,2, \ldots, N$$

$$C_{ij} = \begin{cases} 1 & R^{mcs}_{i(x \to y)j} \geq R_{i(x \to y)} \\ 0 & R^{mcs}_{i(x \to y)j} \geq R_{i(x \to y)} \end{cases}, \qquad \text{where } R_i > 1$$

where *i* is the *i*th residue, *N* is the number of residues in the protein alignment, *j* is *j*th sequence simulation, *M* is the number of sequence simulations run, C_{ij} is the piecewise function, $R_{i(x \to y)}$ is the recurrence of amino acid substitution $x \to y$ at the *i*th site in the real phylogeny, and $R^{mcs}_{i(x \to y)j}$ is the recurrence of amino acid substitution $x \to y$ at the *i*th site in the *j*th simulation.

4.5.5 Data set retrieval and curation

A dataset comprising all available SARS-CoV-2 surface glycoprotein (S protein) sequences derived from human hosts were downloaded from the National Center for Biotechnology Information on 29th July 2024 (n = 3,934,602). The dataset was filtered to remove redundant sequences, sequences shorter than 3,800 bp, sequences with ambiguous characters, and sequences without start or stop codons. This resulted in a dataset of 123,717 full-length non-redundant sequences. The most closely related S protein sequence derived from a bat host to that found in a human host was added to the dataset as an outgroup (accession: MZ937000.1:21512-25321). The coding sequences were then translated into protein using the standard genetic code.

4.5.6 Computing runtime and memory usage

To generate input datasets to demonstrate the runtime characteristics of RECUR, the 123,717 non-redundant sequences above were aligned using FAMSA (Deorowicz, et al. 2016) (Supplementary File C3). Ten-fold replicated subset multiple sequence alignments were then randomly sampled from this complete alignment, with each subset containing either 8, 16, 32, 64, 128, 256, 512 or 1024 sequences including the outgroup sequence. RECUR was run on each sampled alignment with four different input variations: no pre-inferred constraint tree and no pre-computed model of sequence evolution, a pre-inferred constraint tree but no pre-computed model of sequence evolution, a pre-computed model of sequence evolution but no pre-inferred constraint tree, and both a pre-inferred constraint tree and a pre-computed model

of sequence evolution. In analyses where constraints were provided, the models of sequence evolution and constraint trees used were those inferred by RECUR when no constraints were provided to prevent differences in tree topology or model of evolution influencing runtime or memory use. Each run was timed, and the maximum proportional set size (PSS) was recorded. These tests were conducted on an AMD EPYC 7742 64-Core Processor (2.25 GHz, x86_64 architecture) using eight threads.

4.5.7 Analysing the complete 123,717 SARS-CoV-2 surface glycoprotein sequence dataset

IQ-TREE 2 is unable to infer a phylogenetic tree from an alignment comprising 123,717 sequences. Thus, a maximum likelihood constraint tree was inferred using VeryFastTree (Piñeiro, et al. 2020) which is designed for phylogenetic tree inference from large alignments (Supplementary File C4). In addition, the best fitting model of sequence evolution and its averaged model parameters, HIVw+F+I{0.24}+G4{0.59}, were computed from the runtime and memory usage runs for the ten replicates of 1024 sequences above. By constraining the tree and the model of sequence evolution RECUR is thus able to analyse the complete sequence dataset.

4.6 Acknowledgements

This work was funded by the Royal Society and the European Union's Horizon 2020 research and innovation program under grant agreement number 637765 and by the Wellcome Trust under grant number 226598/Z/22/Z. ER is funded by the Biotechnology and Biological Sciences Research Council (BBSRC) [grant numbers BB/M011224/1 and BB/P003117/1]. This research was funded in whole, or in part, by the BBSRC. For the purpose of open access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

4.7 Data Availability

The source code is available on the RECUR GitHub repository (https://github.com/OrthoFinder/RECUR). All data used in this study is provided in the supplementary material.

4.8 Author Contributions

ER and SK conceived the study. ER and YL wrote the software. ER and YL carried out the analysis. ER and SK wrote the manuscript. All authors edited and approved the final manuscript.

4.9 References

Brodie III ED, Brodie Jr ED. 2015. Predictably convergent evolution of sodium channels in the arms race between predators and prey. Brain Behavior and Evolution 86:48-57.

Bull J, Badgett M, Wichman HA, Huelsenbeck JP, Hillis DM, Gulati A, Ho C, Molineux I. 1997. Exceptional convergent evolution in a virus. Genetics 147:1497-1507.

Carabelli AM, Peacock TP, Thorne LG, Harvey WT, Hughes J, de Silva TI, Peacock SJ, Barclay WS, de Silva TI, Towers GJ, et al. 2023. SARS-CoV-2 variant biology: Immune escape, transmission and fitness. Nature Reviews Microbiology 21:162-177.

Chand GB, Banerjee A, Azad GK. 2020. Identification of twenty-five mutations in surface glycoprotein (Spike) of SARS-CoV-2 among Indian isolates and their impact on protein dynamics. Gene Rep 21:100891.

Christin P-A, Salamin N, Muasya AM, Roalson EH, Russier F, Besnard G. 2008. Evolutionary switch and genetic convergence on rbcL following the evolution of C4 photosynthesis. Molecular Biology and Evolution 25:2361-2368.

Christin P-A, Salamin N, Savolainen V, Duvall MR, Besnard G. 2007. C4 Photosynthesis evolved in grasses via parallel adaptive genetic changes. Current Biology 17:1241-1247.

Christin PA, Samaritani E, Petitpierre B, Salamin N, Besnard G. 2009. Evolutionary insights on C4 photosynthetic subtypes in grasses from genomics and phylogenetics. Genome Biol Evol 1:221-230.

Cutting GR, Kasch LM, Rosenstein BJ, Zielenski J, Tsui L-C, Antonarakis SE, Kazazian Jr HH. 1990. A cluster of cystic fibrosis mutations in the first nucleotide-binding fold of the cystic fibrosis conductance regulator protein. Nature 346:366-369.

Dadonaite B, Brown J, McMahon TE, Farrell AG, Figgins MD, Asarnow D, Stewart C, Lee J, Logue J, Bedford T, et al. 2024. Spike deep mutational scanning helps predict success of SARS-CoV-2 clades. Nature 631:617-626.

Davies H, Bignell GR, Cox C, Stephens P, Edkins S, Clegg S, Teague J, Woffendin H, Garnett MJ, Bottomley W, et al. 2002. Mutations of the BRAF gene in human cancer. Nature 417:949-954.

De Maio N, Walker C, Borges R, Weilguny L, Slodkowicz G, Goldman N. 2020. Issues with SARS-CoV-2 sequencing data.

Deorowicz S, Debudaj-Grabysz A, Gudyś A. 2016. FAMSA: Fast and accurate multiple sequence alignment of huge protein families. Scientific Reports 6:33964.

Dobler S, Dalla S, Wagschal V, Agrawal AA. 2012. Community-wide convergent evolution in insect adaptation to toxic cardenolides by substitutions in the Na,K-ATPase. Proceedings of the National Academy of Sciences 109:13040-13045.

Feldman CR, Brodie ED, Brodie ED, Pfrender ME. 2012. Constraint shapes convergence in tetrodotoxin-resistant sodium channels of snakes. Proceedings of the National Academy of Sciences 109:4556-4561.

Furusawa Y, Kiso M, Iida S, Uraki R, Hirata Y, Imai M, Suzuki T, Yamayoshi S, Kawaoka Y. 2023. In SARS-CoV-2 delta variants, Spike-P681R and D950N promote membrane fusion, Spike-P681R enhances spike cleavage, but neither substitution affects pathogenicity in hamsters. eBioMedicine 91:104561.

Grabowski F, Preibisch G, Giziński S, Kochańczyk M, Lipniacki T. 2021. SARS-CoV-2 variant of concern 202012/01 has about twofold replicative advantage and acquires concerning mutations. Viruses 13.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52:696-704.

Gupta D, Sharma P, Singh M, Kumar M, Ethayathulla AS, Kaur P. 2021. Structural and functional insights into the spike protein mutations of emerging SARS-CoV-2 variants. Cellular and Molecular Life Sciences 78:7967-7989.

Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, et al. 2020. Array programming with NumPy. Nature 585:357-362.

Harvey WT, Carabelli AM, Jackson B, Gupta RK, Thomson EC, Harrison EM, Ludden C, Reeve R, Rambaut A, Peacock SJ, et al. 2021. SARS-CoV-2 variants, spike mutations and immune escape. Nature Reviews Microbiology 19:409-424.

He K, Eastman TG, Czolacz H, Li S, Shinohara A, Kawada SI, Springer MS, Berenbrink M, Campbell KL. 2021. Myoglobin primary structure reveals multiple convergent transitions to semi-aquatic life in the world's smallest mammalian divers. Elife 10.

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2017. UFBoot2: Improving the ultrafast tootstrap approximation. Molecular Biology and Evolution 35:518-522.

Hunter P. 2007. The nature of flight. The molecules and mechanics of flight in animals. EMBO Rep 8:811-813.

Iketani S, Mohri H, Culbertson B, Hong SJ, Duan Y, Luck MI, Annavajhala MK, Guo Y, Sheng Z, Uhlemann A-C, et al. 2023. Multiple pathways for SARS-CoV-2 resistance to nirmatrelvir. Nature 613:558-564.

Ingram VM. 1957. Gene mutations in human hæmoglobin: the chemical difference between normal and sickle cell hæmoglobin. Nature 180:326-328.

Jänne PA, Riely GJ, Gadgeel SM, Heist RS, Ou S-HI, Pacheco JM, Johnson ML, Sabari JK, Leventakos K, Yau E. 2022. Adagrasib in non–small-cell lung cancer harboring a KRASG12C mutation. New England Journal of Medicine 387:120-131.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nature Methods 14:587-589.

Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, Hengartner N, Giorgi EE, Bhattacharya T, Foley B, et al. 2020. Tracking changes in SARS-CoV-2 spike: Evidence that D614G increases infectivity of the COVID-19 virus. Cell 182:812-827.e819.

Kumar S, Delipan R, Chakraborty D, Kanjo K, Singh R, Singh N, Siddiqui S, Tyagi A, Jha V, Thakur KG, et al. 2023. Mutations in S2 subunit of SARS-CoV-2 Omicron spike strongly influence its conformation, fusogenicity, and neutralization sensitivity. J Virol 97:e0092223.

Liu Y, Cotton JA, Shen B, Han X, Rossiter SJ, Zhang S. 2010. Convergent sequence evolution between echolocating bats and dolphins. Curr Biol 20:R53-54.

Ly-Trong N, Barca GMJ, Minh BQ. 2023. AliSim-HPC: parallel sequence simulator for phylogenetics. Bioinformatics 39.

Maeso I, Roy SW, Irimia M. 2012. Widespread recurrent evolution of genomic features. Genome Biol Evol 4:486-500.

McCallum M, De Marco A, Lempp FA, Tortorici MA, Pinto D, Walls AC, Beltramello M, Chen A, Liu Z, Zatta F, et al. 2021. N-terminal domain antigenic mapping reveals a site of vulnerability for SARS-CoV-2. Cell 184:2332-2347.e2316.

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Molecular Biology and Evolution 37:1530-1534.

Nilsson DE. 2013. Eye evolution and its functional basis. Vis Neurosci 30:5-20.

Olivier M, Hollstein M, Hainaut P. 2010. TP53 mutations in human cancers: origins, consequences, and clinical use. Cold Spring Harb Perspect Biol 2:a001008.

Piñeiro C, Abuín JM, Pichel JC. 2020. Very Fast Tree: speeding up the estimation of phylogenies for large alignments through parallelization and vectorization strategies. Bioinformatics 36:4658-4659.

Projecto-Garcia J, Natarajan C, Moriyama H, Weber RE, Fago A, Cheviron ZA, Dudley R, McGuire JA, Witt CC, Storz JF. 2013. Repeated elevational transitions in hemoglobin function during the evolution of Andean hummingbirds. Proceedings of the National Academy of Sciences 110:20669-20674.

Robbins EH, Kelly S. 2024. Widespread adaptive evolution in the photosystems of angiosperms provides new insight into the evolution of photosystem II repair. bioRxiv:2024.2005.2023.595573.

Shen L, Triche TJ, Bard JD, Biegel JA, Judkins AR, Gai X. 2021. Spike Protein NTD mutation G142D in SARS-CoV-2 Delta VOC lineages is associated with frequent back mutations, increased viral loads, and immune evasion. medRxiv:2021.2009.2012.21263475.

Shen X-X, Li Y, Hittinger CT, Chen X-x, Rokas A. 2020. An investigation of irreproducibility in maximum likelihood phylogenetic inference. Nature Communications 11:6096.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312-1313.

Stern DL. 2013. The genetic causes of convergent evolution. Nature Reviews Genetics 14:751-764.

Sukumaran J, Holder MT. 2010. DendroPy: a Python library for phylogenetic computing. Bioinformatics 26:1569-1571.

Tian F, Tong B, Sun L, Shi S, Zheng B, Wang Z, Dong X, Zheng P. 2021. N501Y mutation of spike protein in SARS-CoV-2 strengthens its binding to receptor ACE2. Elife 10:e69091.

Tiessen A, Pérez-Rodríguez P, Delaye-Arredondo LJ. 2012. Mathematical modeling and comparison of protein size distribution in different plant, animal, fungal and microbial species reveals a negative correlation between protein size and protein number, thus providing insight into the evolution of proteomes. BMC Res Notes 5:85.

Toprak E, Veres A, Michel J-B, Chait R, Hartl DL, Kishony R. 2012. Evolutionary paths to antibiotic resistance under dynamically sustained drug selection. Nature Genetics 44:101-105.

Ujvari B, Casewell NR, Sunagar K, Arbuckle K, Wüster W, Lo N, O'Meally D, Beckmann C, King GF, Deplazes E. 2015. Widespread convergence in toxin resistance by predictable molecular evolution. Proceedings of the National Academy of Sciences 112:11911-11916.

van Ditmarsch D, Boyle KE, Sakhtah H, Oyler JE, Nadell CD, Déziel É, Dietrich LE, Xavier JB. 2013. Convergent evolution of hyperswarming leads to impaired biofilm formation in pathogenic bacteria. Cell Rep 4:697-708.

Wang Q, Ye SB, Zhou ZJ, Song AL, Zhu X, Peng JM, Liang RM, Yang CH, Yu XW, Huang X, et al. 2023. Key mutations in the spike protein of SARS-CoV-2 affecting neutralization resistance and viral internalization. J Med Virol 95:e28407.

Wilkinson SAJ, Richter A, Casey A, Osman H, Mirza JD, Stockton J, Quick J, Ratcliffe L, Sparks N, Cumley N, et al. 2022. Recurrent SARS-CoV-2 mutations in immunodeficient patients. Virus Evolution 8.

Yang X-J. 2021. Delta-1 variant of SARS-CoV-2 acquires spike V1264L and drives the pandemic in Indonesia, Singapore and Malaysia. In: Research Square.

Yokoyama S, Tada T, Zhang H, Britt L. 2008. Elucidation of phenotypic adaptations: Molecular analyses of dim-light vision proteins in vertebrates. Proceedings of the National Academy of Sciences 105:13480-13485.

Zhen Y, Aardema ML, Medina EM, Schumer M, Andolfatto P. 2012. Parallel molecular evolution in an herbivore community. Science 337:1634-1637.

Zhou X, Shen X-X, Hittinger CT, Rokas A. 2017. Evaluating fast maximum likelihood-based phylogenetic programs using empirical phylogenomic data sets. Molecular Biology and Evolution 35:486-503.

Chapter 5

General Discussion

The plastid genome originated from a free-living cyanobacterium. Despite significant reduction, it has retained essential genes for oxygenic photosynthesis, a process that powers nearly all life on Earth. In this thesis, I have explored the evolution of this compact yet crucial piece of DNA within the clade of plants that contains the majority of crops, the angiosperms. In Chapter 2, I investigated the evolutionary rates of the highly conserved plastid-encoded protein-coding genes, revealing that photosynthetic genes evolve more slowly than those involved in transcription and translation. Furthermore, I identified three key factors that have significantly constrained the evolvability of plastid-encoded genes: gene position, level of gene expression, and the amino acid composition of the encoded protein. In Chapter 3 I searched for molecular signatures of adaptive evolution, namely positive selection and recurrent evolution, in the plastid-encoded photosystem genes, uncovering hallmarks of adaptation at 7% of residues. Finally, in Chapter 4 I introduced RECUR, a software package developed to detect recurrent evolution based on the methodology I established in Chapter 3. Each chapter contains its own dedicated discussion, while this general discussion places the work presented in this thesis in the context of the wider literature and outlines directions for future work.

5.1 Constraints on the evolvability of plastid-encoded genes

Prior to the work presented in this thesis, studies on plastid genome evolution primarily focused on comparing structural features and gene content of newly sequenced plastid genomes to those of closely related species (Saski, et al. 2005; Chumley, et al. 2006; Daniell, et al. 2006; Jansen, et al. 2006; Lee, et al. 2006; Saski, et al. 2007; Mardanov, et al. 2008; Park, et al. 2018). These studies were instrumental in advancing our understanding of plastid genome dynamics, such as the stability conferred by the inverted repeat (which has been lost in some legumes) (Palmer and Thompson 1982; Choi, et al. 2019), resolving phylogenetic relationships (Jansen, et al. 2006), and understanding the transition from autotrophic to heterotrophic lifestyles observed in non-photosynthetic parasitic plants (Funk, et al. 2007; Logacheva, et al. 2016).

Despite these advances, the relatively slow rate of evolution in the plastid genome, as reported in early studies, has limited interest in investigating the molecular evolution of plastid-encoded genes. One notable exception has been the large subunit of rubisco (*rbcL*), where significant effort has been directed toward identifying signals of adaptive evolution, specifically positive selection (Kapralov and Filatov 2007; Kapralov, et al. 2010; Hermida-Carrera, et al. 2017), to inform rational protein design in the hopes of engineering a better rubisco. This thesis aimed to investigate some of the under explored areas of plastid gene evolution, starting by uncovering the relative rates of evolution of these genes and identifying factors that may have constrained their evolution. To do this, I compiled a large dataset of plastid genomes from angiosperms and analysed both the background substitution rate that occurred within genes and the rate at which the encoded proteins had evolved. My findings revealed that over 50% of the variation in rates of nucleotide and protein evolution could be explained by a combination of production requirements and compositional factors of individual genes (Robbins and Kelly 2023). Specifically, I demonstrated that genes positioned further away from an inverted repeat border had lower rates of molecular evolution, consistent with a bidirectional dual displacement loop mechanism of plastid genome replication (Kolodner and Tewari 1975c; Robbins and Kelly 2023). Furthermore, I discovered that genes with higher expression levels had lower rates of molecular evolution. Notably, this trend could not be attributed to selection acting to maintain a low transcript biosynthesis cost or a high translational efficiency, suggesting a potential role for transcription-dependent DNA repair mechanisms (Robbins and Kelly 2023). Finally, I found that the amino acid composition of the encoded protein had a significant impact on the rate of molecular evolution of plastid-encoded genes, with genes encoding more hydrophobic proteins having evolved slower due to the compositional constraints of maintaining hydrophobicity. However, although these factors

could explain ~50% of the variation in evolutionary rate between plastid-encoded genes, this still left nearly half of the variation unexplained, and thus there is much still to be discovered about the evolutionary constraints and evolvability of the plastid genome.

One likely unexplored source of variation that could affect the rate of evolution of the plastid-encoded genes are requirements of the encoded protein sequences to maintain structural features. For example, this could be the percentage of the protein that has an ordered secondary structure, e.g., alpha helices or beta-sheets, or the requirement for the encoded protein to form interfaces between cofactors or subunits. The latter is likely given that in Chapter 3 invariant residues were enriched at both cofactor binding sites and inter-subunit interfaces. Furthermore, an analysis like the one suggested above would be possible for the majority of plastid-encoded genes given the availability of the associated protein complexes in the protein data bank. These include structures of photosystem I at 2.6 Å from *Pisum sativum* (5L8R) (Mazor, et al. 2017), photosystem II at 2.8 Å from *Arabidopsis thaliana* (7OUI) (Graça, et al. 2021), cytochrome b_6f complex at 3.6 Å from *Spinacia oleracea* (6RQF) (Malone, et al. 2019), NAD(P)H dehydrogenase-like complex at 3.7 Å from *Hordeum vulgare* (7EU3) (Shen, et al. 2022), plastid-encoded ATP synthase at 3.2 Å in *Spinacia oleracea* (6FKF) (Hahn, et al. 2018), 70S ribosome at 3.0 Å in *Spinacia oleracea* (6ERI) (Perez Boerema, et al. 2018), plastid-encoded RNA polymerase at 2.5 Å from *Sinapis alba* (8R5O) (Vergara-Cruces, et al. 2024) and the rubisco holoenzyme at 1.5 Å from *Arabidopsis thaliana* (5IU0) (Valegård, et al. 2018). Since different structures can capture different interactions, e.g., inter-complex interactions in the PSI-NDH structure (7F9O) (Shen, et al. 2022), it would be beneficial to consult more than one structure for each complex to get a more complete picture of residues at interfaces. Given the strong co-occurrence of invariant sites at positions of structural importance, the requirement to maintain structural properties essential to functionality is likely to explain the majority of the as-yet-unexplained variation in the rate of molecular evolution between plastid-encoded genes.

Another potential unexplored source of variation in the rates of molecular evolution between plastid-encoded genes is genetic redundancy. Specifically, instances where plastid-encoded genes have nuclear-encoded paralogs whose encoded proteins are targeted to the plastid. This allows a plastid-encoded gene to be functionally compensated or even replaced by a pre-existing nuclear-encoded gene in the event it acquires an N-terminal chloroplast transit peptide. Once replaced the selection pressure acting on the plastid gene would decrease significantly allowing the accumulation of substitutions without detrimental effects on organismal fitness. Since there are preexisting copies of functionally redundant ribosomal genes encoded elsewhere in the cell (mostly in the nuclear genome and a few in the

mitochondrial genome) this may contribute to these genes having elevated rates of molecular evolution that we observed in Chapter 2. An example where this has occurred in angiosperms is the silencing of plastid-encoded *rps16* in favour of a mitochondrially derived nuclear-encoded copy in *Arabidopsis thaliana*, *Lycopersicon esculentum* and *Oryza Sativa* (Ueda, et al. 2008). In more extreme cases the import of the nuclear-encoded *rps16* has resulted in loss of the plastid-encoded *rps16*, which has been observed in *Populus alba* and *Medicago truncatula* (Ueda, et al. 2008). Similarly, plastid-encoded *accD* (ATCG00500 in *A. thaliana*) has a nuclear-encoded paralog called ACC2 (AT1G36180 in *A. thaliana*). These genes both contribute to the production of fatty acids in plastids, with enzyme encoded by either gene capable of carrying out the required acetyl-CoA carboxylase reaction. However, the presence of the nuclear-encoded ACC2 has enabled monocots to lose *accD* from the plastid genome (Harris, et al. 2013).

In summary, while we have demonstrated several factors, namely gene position, gene expression and the encoded protein composition, have imparted strong constraints on the evolvability of the plastid-encoded genes, other factors remain to be explored. The future investigations outlined above will be made easier as more structures are resolved and more whole plant genomes are sequenced, allowing intricate probing of the effects of protein structural requirements and the interplay of evolution between the plastid genome and the nuclear genome, respectively.

5.2 Incorporating possible adaptive alleles into crop plastid genomes.

The work presented in Chapter 3 of this thesis focussed on computational analysis to identify likely adaptive substitutions in the plastid-encoded genes of angiosperms. Specifically, I uncovered several amino acid substitutions (D1: L36V, E231Q, E235A, E243G and PsbT: K28T) that may speed up recovery from photoinhibition. It is worth mentioning here that while photoinhibition is often considered detrimental as it reduces photosynthetic efficiency, protective roles have also been proposed following the observation that photoinhibition is accelerated under stress conditions, e.g., high light, high and low temperatures, high salinity and drought (Powles 1984; Murata, et al. 2007; Nishiyama and Murata 2014). Interestingly, this accelerated photoinhibition seems to come from supressed D1 repair, specifically inhibition of *psbA* translation, rather than an increased rate of photodamage (Nishiyama, et al. 2001; Nishiyama, et al. 2004). Potential benefits include minimizing wastage on futile and energetically expensive D1 repair cycles (which include migration of damaged photosystem II to thylakoid lamellae, proteolytic digestion of D1, *de novo* D1 synthesis, incorporation of functional D1 into the photosystem II complex and migration of the repaired photosystem II

complex back into the thylakoid grana (Theis and Schroda 2016)) under persistent stress. Additionally, photoinhibition in a highly reduced environment helps protect the irreversible oxidative damage to photosystem I which can take weeks to recover from (Tikkanen, et al. 2014). Given that the balance between the harmful and beneficial effects of photoinhibition is dependent on the environmental conditions, further investigation into whether the D1 substitutions identified in this thesis correlate with species adapted to distinct ecological niches could provide valuable insights to benefits they may confer.

With the putative adaptive substitutions I identified in photosystem II, the next step would be to experimentally validate these hypothesise by introducing these substitutions *in planta* and see how they affect photosystem function. Although the phenotype would be relatively straightforward to evaluate using a chlorophyll fluorescence-based screen under a light gradient treatment to assess whether the introduced substitutions have altered photoinhibition (Guidi, et al. 2019), the major challenge would be editing the plastid genome. Plastid genome engineering is impeded by two major obstacles. The first is the high copy number of the plastid genome, which can exceed 10,000 copies per cell (Bendich 1987), and all the copies must be edited to reach homoplasy. The second problem is the relative impermeability of the plastid to nucleic acids, which are essential components of widely used CRISPR-based systems.

To circumvent the requirement for delivering nucleic acids to the plastid via a biological approach, various chloroplast transformation methods have been developed. Notably, these transformation methods have been used for the introduction of point mutations (Martin Avila, et al. 2016) and transgenes, with the latter being desirable owing to several favourable characteristics of gene expression in plastids, which are nicely summarised in (Fuentes, et al. 2018). Additionally, alternative strategies that bypass the requirement for getting nucleic acids into the plastid include the use of nuclear-engineered programmable protein-based DNA editors, which are targeted to the plastid by attaching an N-terminal chloroplast transit peptide (Kang, et al. 2021). Moreover, nuclear-engineered RNA-binding pentatricopeptide repeat (PPR) proteins have emerged as another innovative tool which can facilitate sequence-specific RNA edits (Yagi, et al. 2014; Lesch, et al. 2022). This enables desired alterations to the plastid proteome without the need to directly alter the plastid genome.

Alternatively, another approach is to incorporate a copy of the gene containing the desired edits into the nuclear genome with an N-terminal chloroplast transit peptide for targeting to the plastid compartment. This approach was undertaken recently, where a copy of *psbA* (encoding the D1 protein) with a rubisco small subunit (*rbcS*) chloroplast transit peptide was engineered into the nucleus of *Arabidopsis thaliana*, *Nicotiana benthamiana* and *Oryza sativa* (Chen, et

al. 2020). Although the aim of the study was to examine whether heat controlled nuclear origin supplemented D1 could increase tolerance to heat stress, the study demonstrates the successful nuclear expression, import into the plastid and incorporation of D1 into photosystem II complexes. Given that this approach is effectively continuing the process of endosymbiotic gene transfer, as long as no foreign DNA is used for the chloroplast transit peptide or promoter, there is a possibility that if used to enhance crop productivity they would be classified as *cis*-genic rather than genetically modified. This may help reduce regulatory hurdles making it easier to bring such an engineered plant to market and achieve widespread agricultural impact.

While this thesis identifies potential editing targets with potential for enhancing crop productivity, and outlines methods in which to identify precision edits, the full realization of this work's impact will depend on continued advancements in genome editing technologies, particularly in the area of organellar editing. These developments will be crucial for validating and then translating these potential photosynthesis enhancing substitutions into practical agricultural solutions.

5.3 RECUR: A new method for identifying adaptive evolution?

The advent of next generation sequencing has resulted in an explosion of publicly available sequencing data. This influx of data has been accompanied by concurrent advancements in analytical tools, such as phylogenetic software, to extract meaningful insights from these large datasets. A keen area of interest for many molecular evolutionary biologists is identifying meaningful change in proteins, i.e., adaptive evolution. In Chapter 4, I presented RECUR, a new phylogenetic software package inspired by the methodology developed in Chapter 3 for identifying recurrent amino acid substitutions, a hallmark of adaptive evolution, within protein coding sequences.

Existing methods for detecting adaptive evolution within protein sequences focus on identifying positive selection, specifically sites where the d_N/d_S ratio (rate of non-synonymous substitution over synonymous substitution) is significantly greater than 1 (Kreitman and Akashi 1995; Eyre-Walker 2006). However, there are notable limitations of the method including concerns surrounding false positives and the provision of limited biological insight into phenotypic adaptations (Chamary, et al. 2006). Specifically, positive selection only analyses rates of change within a protein sequence but provides no functional significance of that change. In this regard, an advantage of RECUR is that it highlights specific amino acid substitutions that may be adaptive. These substitutions can then be used as a guide for

computational investigation into the impact of evolution on protein structure and function. This approach to uncover molecular adaptation is being made easier with the continual development of bioinformatic tools to study the effects of substitutions on proteins, which includes, but is not limited to, protein folding (Abramson, et al. 2024), molecular dynamics (Abraham, et al. 2015; Rodrigues, et al. 2018), protein-protein interfaces (Wood, et al. 2020) and ligand binding (Vangone, et al. 2018). Although not discussed in this thesis, a specific example where this additional information was important for understanding the underlying mechanism of adaptation is site 1 in the photosystem II subunit PsbL. This site was identified as under positive selection in Chapter 3 and RECUR uncovered significant recurrence at this site where a threonine was substituted to a methionine 23 times. Interestingly, this is a RNA editing site (Kudla, et al. 1992), and the cytosine to thymine substitution in the DNA (ACG $\rightarrow$ ATG) removes the requirement to edit the RNA to produce an efficiently translated start codon.

Nevertheless, RECUR has its own limitations. A primary limitation of RECUR is that it can be affected by gene tree species tree discordance, where the gene tree topology (inferred by RECUR) does not perfectly match that of a species tree (Degnan and Rosenberg 2009). This discordance can prevent the accurate interpretation of substitution recurrence. Thus, in Chapter 3 when analysing recurrence in the photosystem genes we used the species tree built from the concatenated sequence alignment from the 69 highly conserved plastid-encoded genes as a constraint tree to mitigate the effects of gene tree inference error. Although, we attempted a similar approach for the analysis in Chapter 4 by trying to generate a tree from a whole SARS-CoV-2 genome alignment, the dataset was too large for successful tree construction due to computational limitations.

Overall, I hope RECUR will be a useful and easy to use tool to delve into the evolution of protein sequences. It can be employed as a standalone method for detecting recurrent evolution or used in conjunction with preexisting positive/purifying selection techniques to help pinpoint important substitutions that are likely to have occurred at sites identified as being under selection.

5.4 The big prize: engineering a better rubisco

The main engineering target for crop improvement in the plastid genome is the large subunit of rubisco (*rbcL*). This is because rubisco is responsible for carbon fixation but is a highly inefficient catalyst. For example, the maximum turnover rate reported for rubisco is <12 s^{-1} (Badger, et al. 1998). Furthermore, it catalyses a competing reaction with O_2 resulting in downstream carbon release, a process termed photorespiration, impeding plant growth

(Bowes, et al. 1971; Chollet 1977; Busch 2020). To overcome these kinetic limitations, plants have adapted to produce vast amounts of rubisco (up to 50% of soluble leaf protein (Carmo-Silva, et al. 2015)) and have also independently evolved carbon concentrating mechanisms to favour carboxylation over oxygenation (Meyer and Griffiths 2013; Schlüter, et al. 2023). But the question remains, could we engineer a superior rubisco to improve crop growth? Initially, it was thought that strong catalytic trade-offs between turnover rate and specificity were constrained the evolution of a rubisco with better kinetics (Tcherkez, et al. 2006; Savir, et al. 2010). However, recently the strength of these trade-offs has been shown to be inflated by phylogenetic signal that was not accounted for in the original analysis (Bouvier, et al. 2021). Furthermore, despite being in the slowest evolving 1% of enzymes on Earth, evolutionary change is still improving rubisco kinetics (Bouvier, et al. 2024). Thus, engineering a better rubisco is a real possibility.

For precision editing, analysing the evolution of the *rbcL* is essential for identifying viable engineering targets. This is because in addition to functional constraints, the RbcL protein is subject to strong constraints to enable correct folding given the requirement of interacting with numerous chaperones for the proper folding and assembly of the rubisco holoenzyme (Carmo-Silva, et al. 2015; Aigner, et al. 2017). An analysis similar to that presented in Chapter 3 for the photosystem genes would identify invariant sites which are likely crucial for maintaining function and/or proper assembly, and thus would not be considered as sensible engineering targets. Furthermore, the application of the RECUR method would build on previous positive selection analyses by providing further biological context and likely identify new sites that have exhibited hallmarks of adaptive evolution. Importantly, this method would highlight tolerated amino acid variants for engineering that would likely not disrupt crucial functions given their occurrence in nature. Since the plastid genome is uniparentally inherited, and thus has an inability to recombine, artificially combining multiple substitutions hypothesised to be adaptive in an interesting avenue for engineering a better rubisco. The methods and techniques developed in this thesis could contribute significantly toward achieving this goal.

5.5 Conclusion

In conclusion, this thesis posits the genes encoded in the plastid genome as exciting engineering targets for future crop improvement. These genes are crucial for photosynthesis and, by extension, growth. Despite their functional importance, the evolvability of these plastid-encoded genes have been significantly constrained by factors unrelated to their primary functions, included being trapped in the slowly evolving plastid genome in the absence of sexual recombination, and thus preventing natural selection from combining adaptive alleles

across individuals even within a species. Furthermore, the work in this thesis has revealed that the majority of the variation in rate of molecular evolution between plastid-encoded genes is explained by compositional features and production requirements rather than functional optimization. This provides a strong basis for the argument that plastid-encoded genes are not perfectly adapted for their functions, and given their vital importance, makes them attractive engineering targets. Nevertheless, subsequent analysis of the slowest-evolving plastid-encoded genes uncovered hallmarks of adaptive evolution are widespread, suggesting that despite these constraints, adaptive evolution has occurred. Moreover, providing a structural context to these adaptive substitutions allowed mechanistic hypothesise to be put forward surrounding the benefit that several of these substitutions conferred. The best prospect for harnessing the potential of the discoveries made in this thesis will be through advances in genome editing technologies as this will allow the results of millions of years of experiments by natural selection across the plant kingdom to be harnessed in a single plastid genome.

Appendices

Appendix A

Appendix A contains the supplementary information for Chapter 2. The supplementary figures are shown first and links to supplementary tables and files provided after.

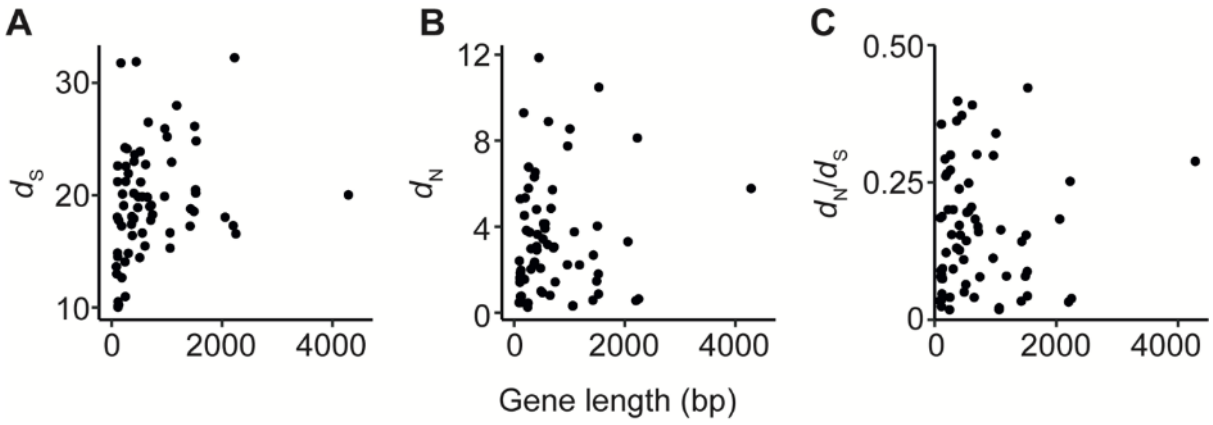


Figure A.1. Relationships between the rates of molecular evolution and sequence length for 69 plastid-encoded genes. Scatter plots of the rate of synonymous substitution (d_S) (**A**), non-synonymous substitution (d_N) (**B**) and the ratio of non-synonymous to synonymous substitution (d_N/d_S) (**C**) versus the average gene length given in base pairs (bp). d_S has the units of number of synonymous changes per synonymous sequence site and d_N has the units of the number of non-synonymous changes per non-synonymous sequence site.

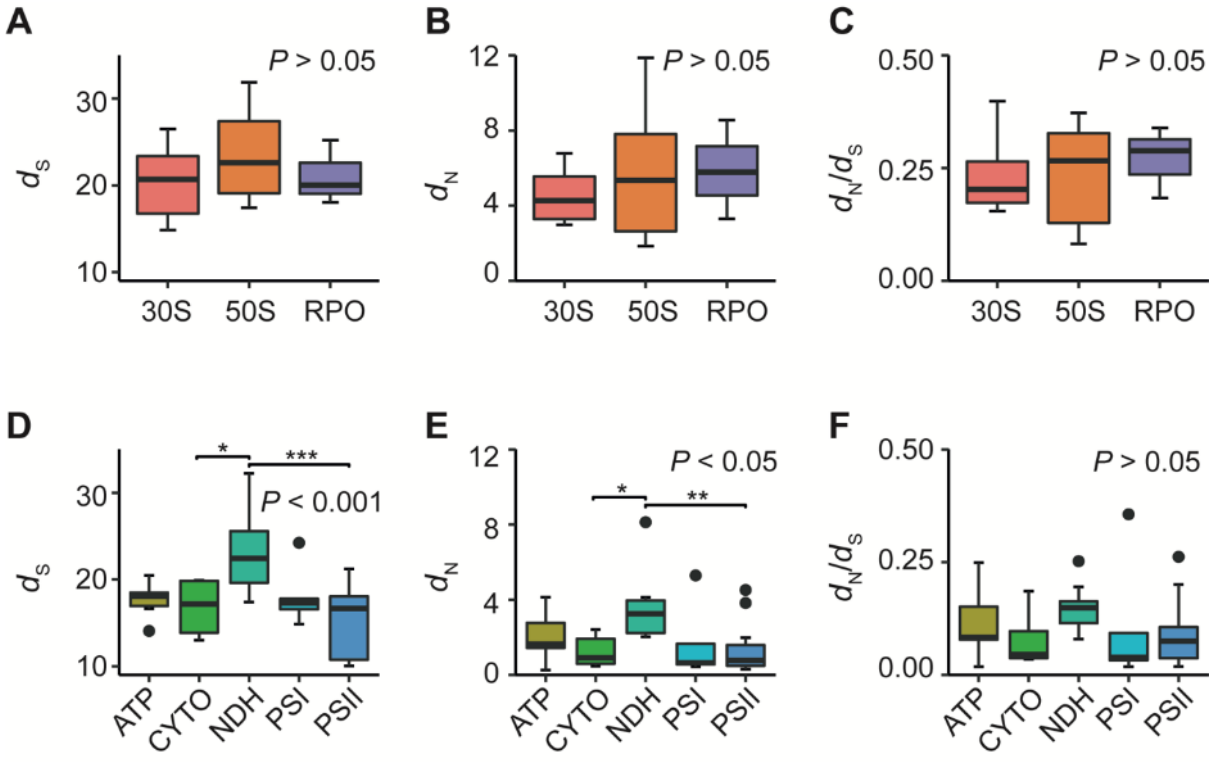


Figure A.2. Exploring the differences in rates of molecular evolution and functional constraint between the multiprotein complexes encoded by the plastid. The difference in the rate of synonymous substitution (d_S), rate of non-synonymous substitution (d_N) and functional constraint (d_N/d_S) between the major multiprotein complexes of the information processing category (**A-C**) and energy production category (**D-F**), respectively. Groups include the large ribosomal subunit (50S, n = 7), small ribosomal subunit (30S, n= 10), RNA polymerase (RPO, n = 3), photosystem I (PSI, n = 5), photosystem II (PSII, n = 15), cytochrome b_6f complex (CYTO, n = 6), ATP-synthase (ATP, n = 6) and NAD(P)H dehydrogenase-like complex (NDH, n = 10). Statistical significance was assessed using one-way ANOVAs and their associated *P*-values are displayed. Where appropriate, asterisks indicate the *P*-value significance levels of *post-hoc* Tukey tests.

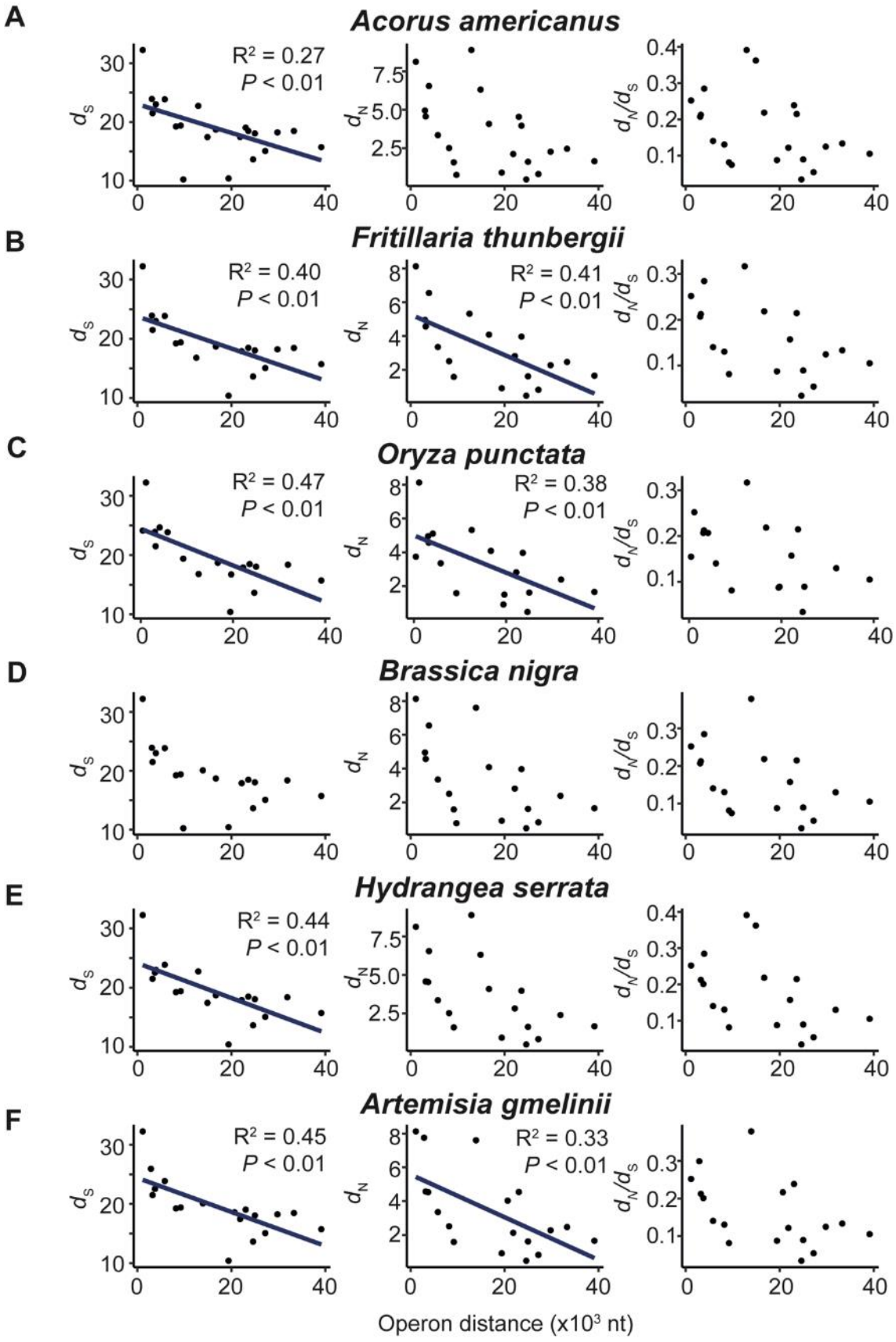

A
Acorus americanus
R² = 0.27
P < 0.01
B
Fritillaria thunbergii
R² = 0.40
P < 0.01
R² = 0.41
P < 0.01
C
Oryza punctata
R² = 0.47
P < 0.01
R² = 0.38
P < 0.01
D
Brassica nigra
E
Hydrangea serrata
R² = 0.44
P < 0.01
F
Artemisia gmelinii
R² = 0.45
P < 0.01
R² = 0.33
P < 0.01
dS
dN
dN/dS
Operon distance (x10³ nt)

Figure A.3. The relationship between the distance of an operon to the closest inverted repeat border and the rate of molecular evolution and strength of purifying selection of its encoded genes for six species from across the tree analysed in this study. These species included three monocots and three eudicots, all from different orders. The rate of synonymous substitution, non-synonymous substitution and strength of purifying selection of an operon is calculated from the average rate of molecular evolution of its encoded genes that were included in this analysis. Scatter plots showing the average distance of each operon to the closest inverted repeat border measured in nucleotides (nt) versus the rate of synonymous substitution (left), non-synonymous substitution (middle) and ratio of non-synonymous substitution to synonymous substitution (right) for *Acorus americanus* (**A**), *Fritillaria thunbergia* (**B**), *Oryza punctata* (**C**), *Brassica nigra* (**D**), *Hydrangea serrata* (**E**) and *Artemisia gmelinii* (**F**). Linear models are shown in blue with their associated R^2 and *P*-values.

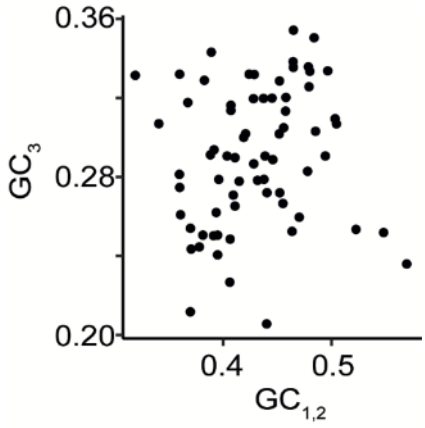


Figure A.4. The relationship between the GC content of the first and second codon positions ($GC_{1,2}$) versus the GC content of the third codon position (GC_3) for 69 plastid-encoded protein coding genes.

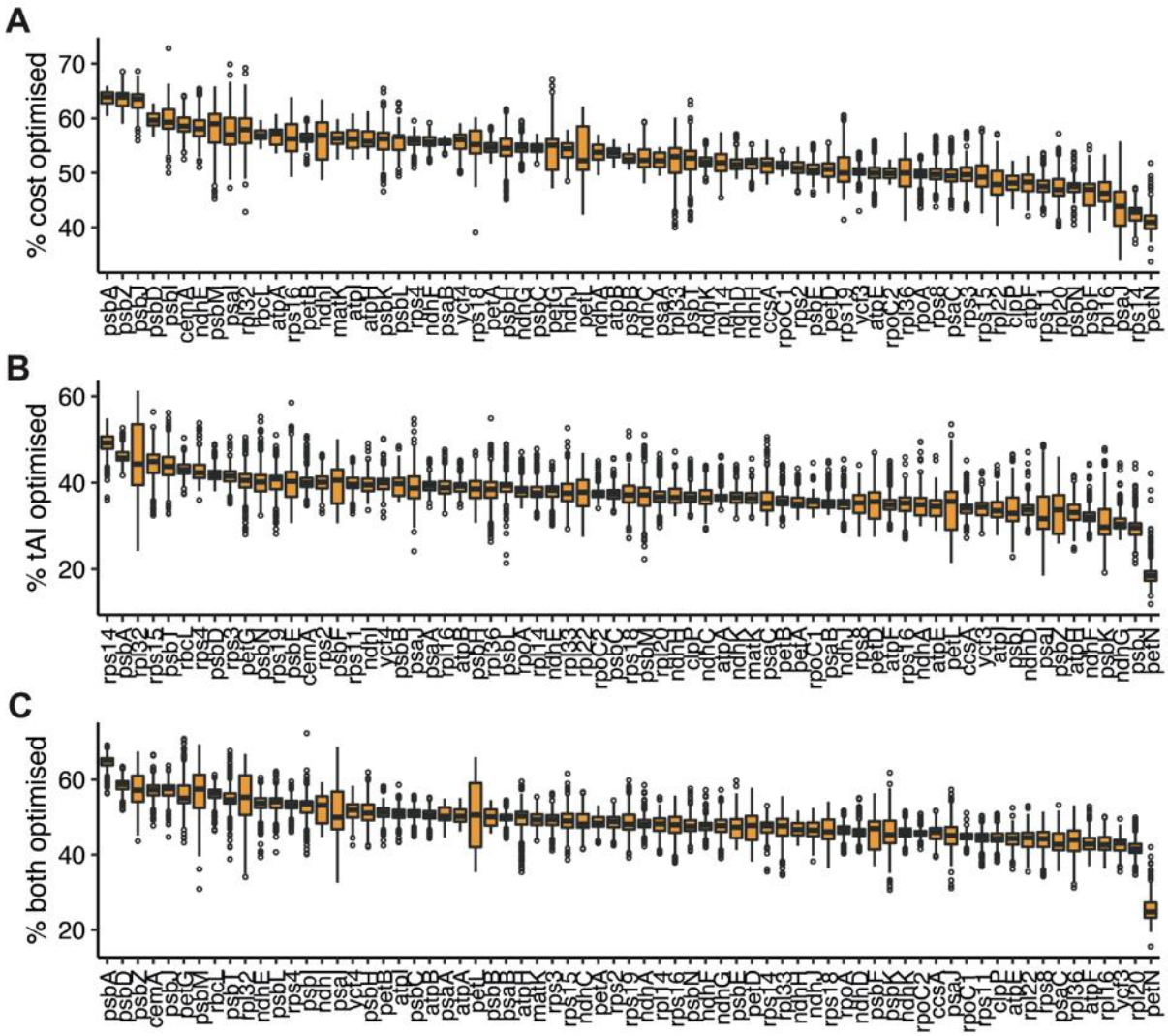


Figure A.5. Extent of gene optimisation for transcript biosynthetic cost and translational efficiency for 69 plastid-encoded genes. For each gene, CodonMuse was used to calculate percent optimisation values for transcript biosynthetic cost (**A**), translational efficiency (**B**) and the trade-off between these two evolutionary forces (**C**) for 773 species which are represented as boxplots. Genes are arranged with decreasing median values from left to right.

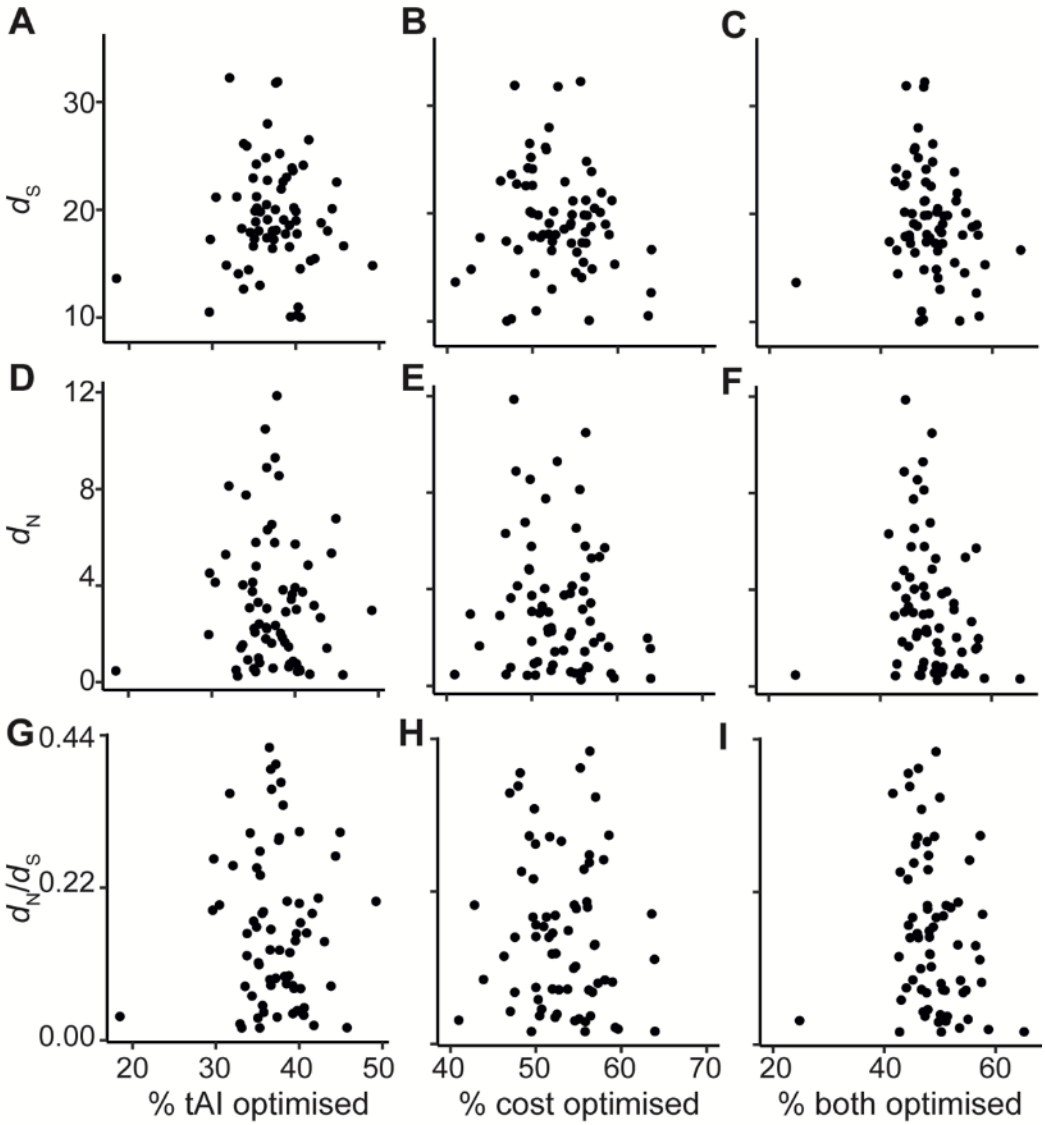


Figure A.6. The relationship between the extent of optimisation of transcript cost and translational efficiency with the rate of molecular evolution. Optimisation values were calculated using CodonMuSe and are given as a percentage, where 100% is the most optimised nucleotide sequence for a given protein sequence. The translational efficiency optimisation score is based on the tRNA adaptation index (tAI) and the cost optimisation scores are based on codon nitrogen cost. Relationships between the median of gene optimisation for translational efficiency, transcript cost and the trade-off with the rate of synonymous substitution (**A-C**), the rate of non-synonymous substitution (**D-F**) and the strength of selection pressure (**G-I**), respectively.

A

Model Index	Predictors
1	Distance
2	Distance + ln(mRNA abundance)
3	Distance + ln(mRNA abundance) + GC3
4	Distance + ln(mRNA abundance) + GC3 + % tAI optimised
5	Distance + ln(mRNA abundance) + GC3 + % tAI optimised + % both optimised
6	Distance + ln(mRNA abundance) + GC3 + % tAI optimised + % both optimised + % cost optimised
7	Distance + ln(mRNA abundance) + GC3 + % tAI optimised + % both optimised + % cost optimised + substitution tolerance

B

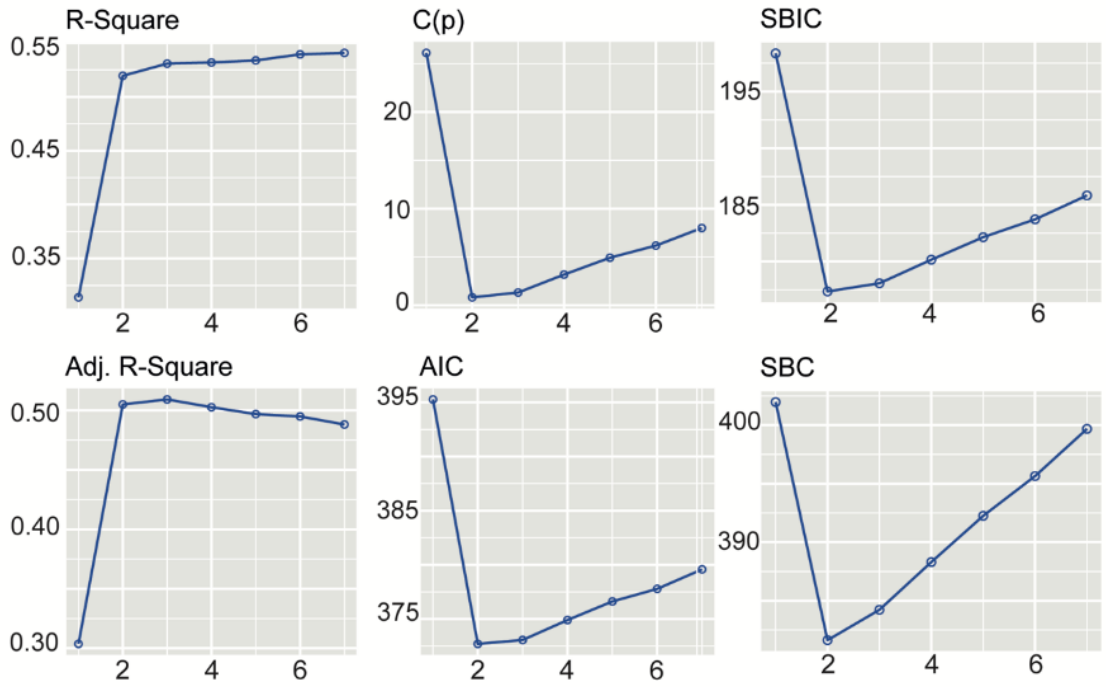


Figure A.7. Variable selection for the best-fitting linear model explaining the maximum variation in the rate of synonymous substitution. **A**) Table showing the subset of variables used for the different models tested (indexed 1-8). Variables include average gene distance to the inverted repeat, Distance; the natural logarithm of transcript abundance, ln(mRNA abundance); GC content of the wobble position, GC_3, mean percent optimisation for translational efficiency, % tAI optimised; mean percent optimisation for transcript cost, % cost optimised; mean trade-off for optimisation of both translational efficiency and transcript cost, % both optimised and protein tolerance to substitution, substitution tolerance. **B**) Various objective criteria values for the models described in A. R-Square, R-squared of the model; Adj. R-Square, R-squared after adjusting for the number of parameters; C(p), Mallow's Cp; AIC, Akaike information criterion; SBIC, Sawa's Bayesian information criteria; SBC, Schwarz Bayesian information criteria.

A

Model Index	Predictors
1	ln(mRNA abundance)
2	ln(mRNA abundance) + substitution tolerance
3	ln(mRNA abundance) + substitution tolerance + Distance
4	ln(mRNA abundance) + substitution tolerance + Distance + % tAI optimised
5	ln(mRNA abundance) + substitution tolerance + Distance + % tAI optimised + GC3
6	ln(mRNA abundance) + substitution tolerance + Distance + % tAI optimised + GC3 + % cost optimised
7	ln(mRNA abundance) + substitution tolerance + Distance + % tAI optimised + GC3 + % cost optimised + % both optimised

B

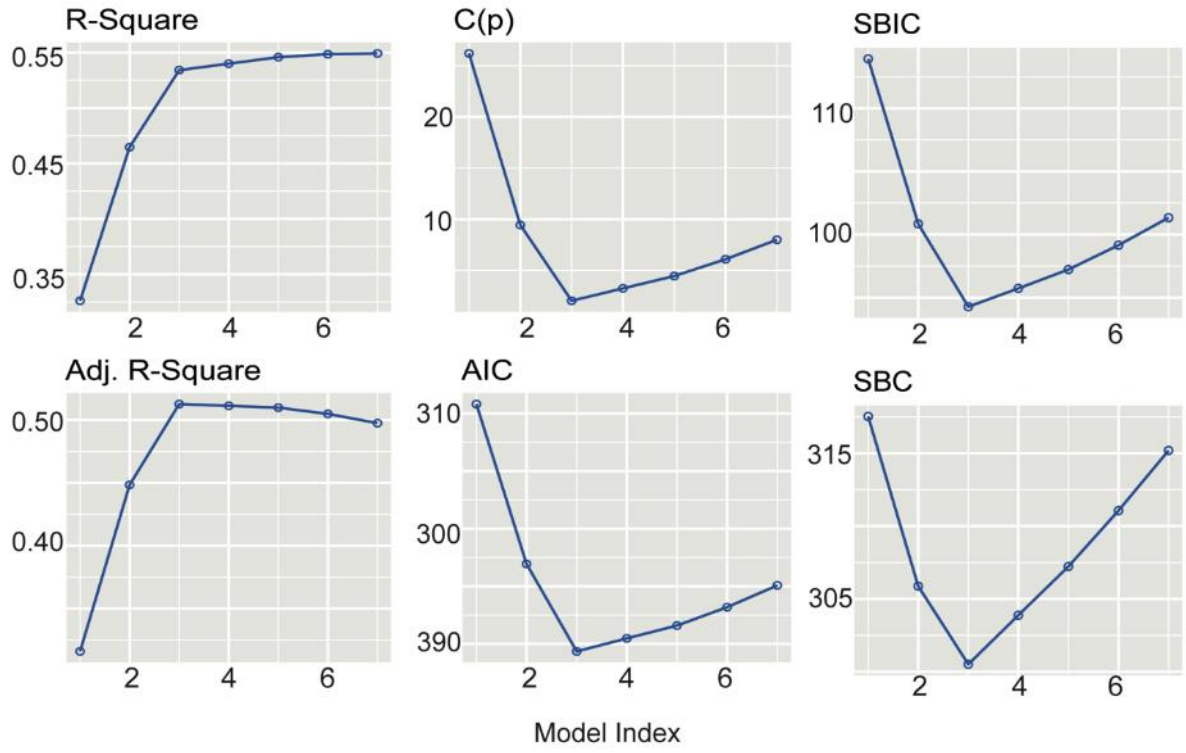


Figure A.8. Variable selection for the best-fitting linear model explaining the maximum variation in the rate of non-synonymous substitution. **A**) Table containing the subset of predictors for each model tested. **B**) Various object criterion values for the models described in A.

A

Model Index	Predictors
1	ln(mRNA abundance)
2	ln(mRNA abundance) + substitution tolerance
3	ln(mRNA abundance) + substitution tolerance + Distance
4	ln(mRNA abundance) + substitution tolerance + Distance + % tAI optimised
5	ln(mRNA abundance) + substitution tolerance + Distance + % tAI optimised + GC3
6	ln(mRNA abundance) + substitution tolerance + Distance + % tAI optimised + GC3 + % both optimised
7	ln(mRNA abundance) + substitution tolerance + Distance + % tAI optimised + GC3 + % both optimised + % cost optimised

B

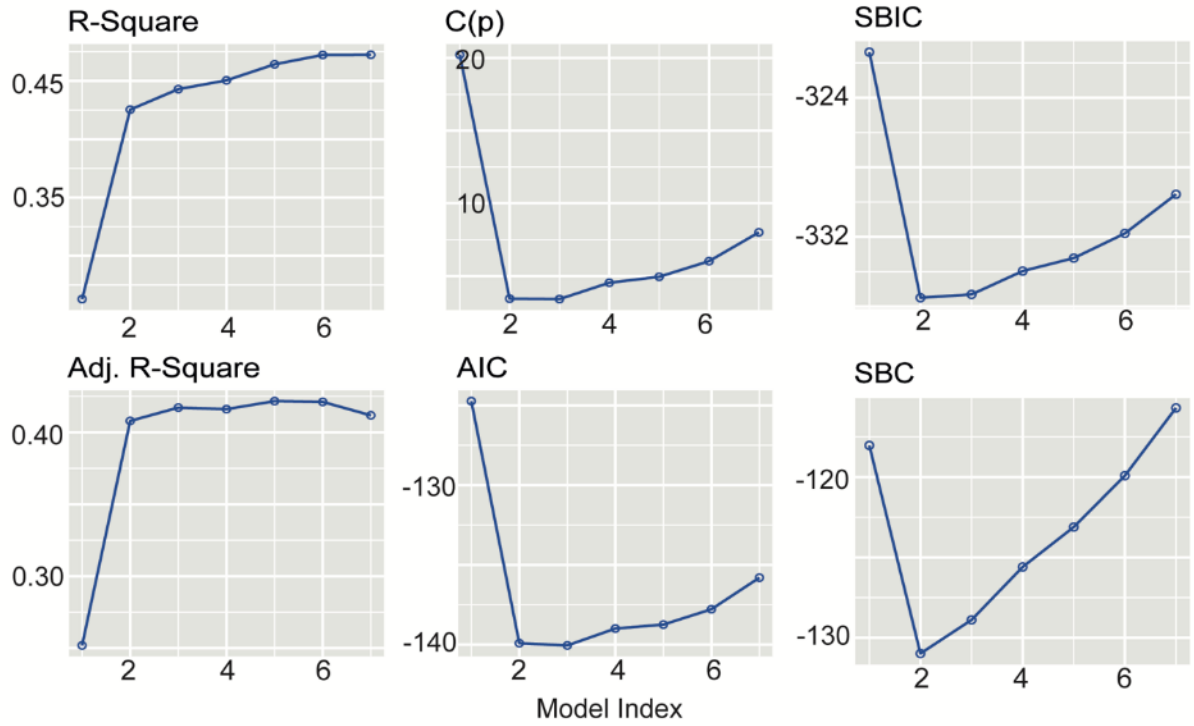


Figure A.9. Variable selection for the best-fitting linear model explaining the maximum variation in the ratio of the rate of non-synonymous substitution to the rate of synonymous substitution. **A**) Table containing the subset of predictors for each model tested. **B**) Various object criterion values for the models described in A.

Tables A.1
Link: https://doi.org/10.1093/gbe/evad101

Table A.2
Link: https://doi.org/10.1093/gbe/evad101

Supplemental File A.1
Link: https://doi.org/10.1093/gbe/evad101

Supplemental File A.2
Link: https://doi.org/10.1093/gbe/evad101

Supplemental File A.3
Link: https://doi.org/10.1093/gbe/evad101

Appendix B

Appendix B contains the supplementary information for Chapter 3. The supplementary figures are shown first and links to supplementary tables and files provided after.

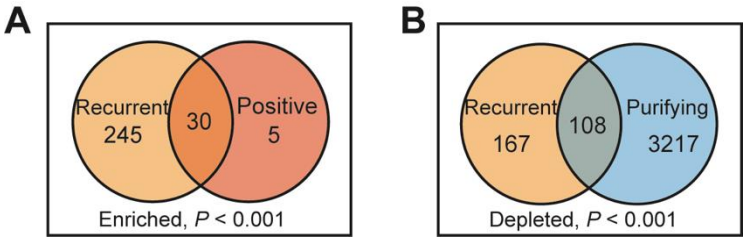


Figure B.1. Venn diagrams showing the overlap between sites with recurrent substitutions (orange) and positive selection (red) (**A**), and purifying selection (**B**). *P*-values shown are the results of hypergeometric tests.

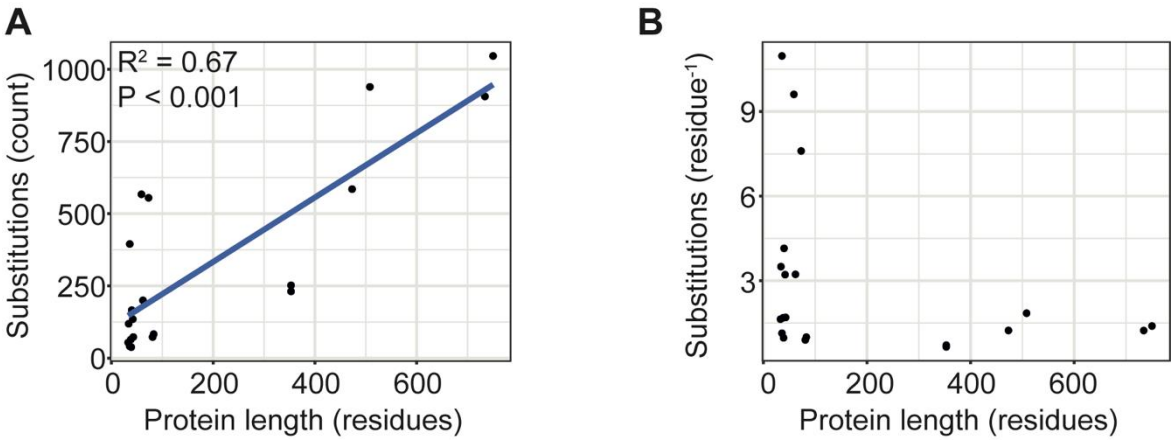


Figure B.2. Relationships between the number of inferred non-synonymous substitutions and protein length. **A**) Scatter plot of the non-synonymous substitution count per protein with the length of the protein given in residues. A linear regression line is shown in blue with the associated R^2 and *P* values given. **B**) Scatter plot of the non-synonymous substitutions per residue versus protein length.

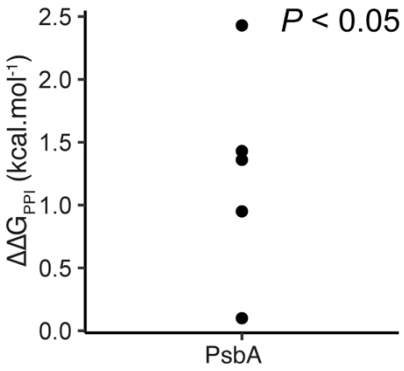


Figure B.3. Plot showing the change in inter-subunit interaction energy resulting from adaptive substitutions at a PsbA (D1) interface (PsbA: L36V, E231Q, E235A, E243G; PsbT K28T). *P*-value indicates the result of a two-sided one-sample t-test.

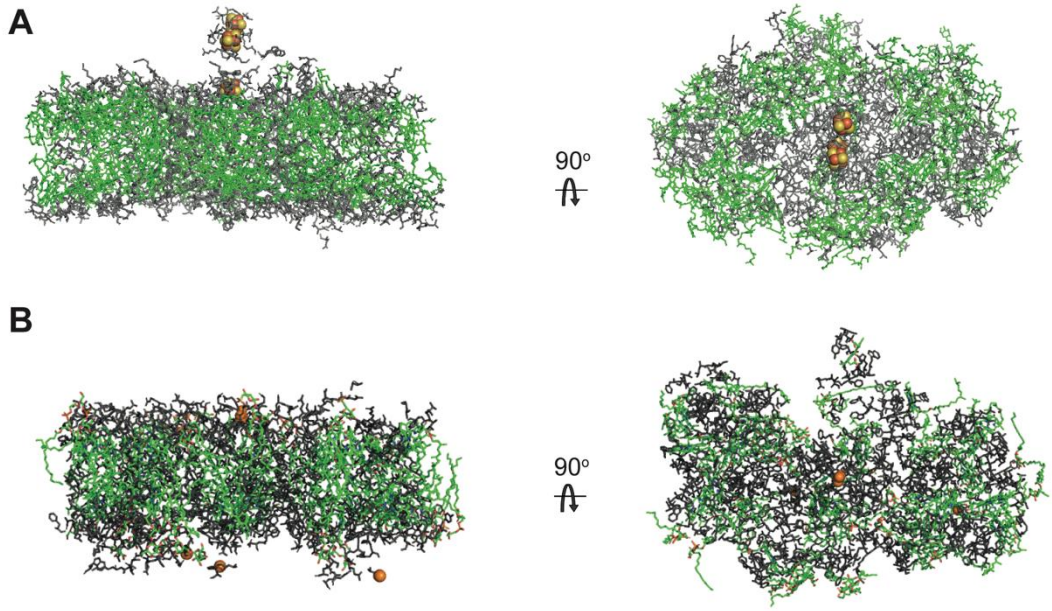


Figure B.4. Protein structures showing the cofactor interacting residues in the plastid-encoded photosystem proteins. **A**) Analysis of the 5L8R structure of photosystem I. Cofactors associated with PsaA, PsaB, PsaC, PsaI or PsaJ are shown in green. The 832 residues in these proteins that have an atom within 4Å of a cofactor atom are shown in grey. **B**) Analysis of the 7OUI structure of photosystem II. Cofactors associated with PsbA, PsbB, PsbC, PsbD, PsbE, PsbF, PsbH, PsbI, PsbK, PsbL, PsbM, PsbT or PsbZ are shown in green or as orange spheres for small molecules and atoms. The 719 residues in these proteins that have an atom within 4 Å of a cofactor atom are shown in grey.

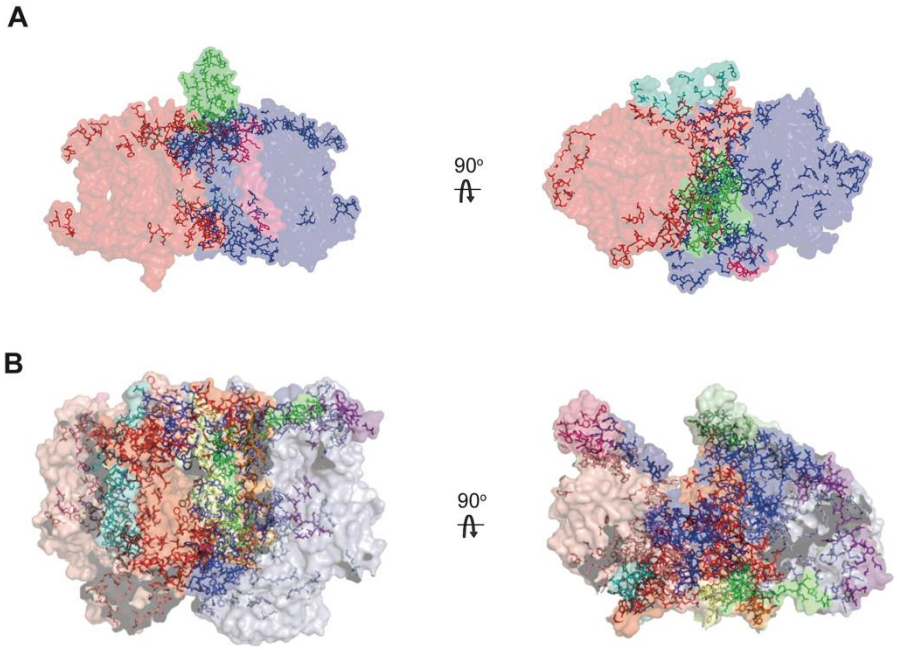


Figure B.5. Protein structures showing the inter-subunit interface residues in the plastid-encoded photosystem proteins. Only proteins analysed in this study are shown. Residues identified at inter-subunit interfaces shown as sticks. **A**) Analysis of the 5L8R structure of photosystem I. Red, PsaA; blue, PsaB; green, PsaC; pink, PsaI and teal, PsaJ. **B**) Analysis of the 7OUI structure of photosystem II. Red, PsbA; light blue, PsbB; salmon, PsbC; blue, PsbD; pale green, PsbE; dark green, PsbF; purple, PsbH; teal, PsbI; dark blue, PsbK; green, PsbL; orange, PsbM; pale yellow, PsbT and pink, PsbZ.

Tables B.1

Link: https://doi.org/10.5281/zenodo.13889880

Tables B.2

Link: https://doi.org/10.5281/zenodo.13889880

Tables B.3

Link: https://doi.org/10.5281/zenodo.13889880

Tables B.4

Link: https://doi.org/10.5281/zenodo.13889880

Tables B.5

Link: https://doi.org/10.5281/zenodo.13889880

Appendix C

Supplementary File C.1

Link: https://doi.org/10.5281/zenodo.13889880

Supplementary File C.2

Link: https://doi.org/10.5281/zenodo.13889880

Supplementary File C.3

Link: https://doi.org/10.5281/zenodo.13889880

Supplementary File C.4

Link: https://doi.org/10.5281/zenodo.13889880

References

Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, Lindahl E. 2015. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX 1-2:19-25.

Abrahamsen MS, Templeton TJ, Enomoto S, Abrahante JE, Zhu G, Lancto CA, Deng M, Liu C, Widmer G, Tzipori S. 2004. Complete genome sequence of the apicomplexan, Cryptosporidium parvum. Science 304:441-445.

Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, Ronneberger O, Willmore L, Ballard AJ, Bambrick J, et al. 2024. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 630:493-500.

Adl SM, Simpson AGB, Farmer MA, Andersen RA, Anderson OR, Barta JR, Bowser SS, Brugerolle G, Fensome RA, Frederricq S, et al. 2005. The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. Journal of Eukaryotic Microbiology 52:399-451.

Aigner H, Wilson R, Bracher A, Calisse L, Bhat J, Hartl F, Hayer-Hartl M. 2017. Plant RuBisCo assembly in E. coli with five chloroplast chaperones including BSD2. Science 358:1272-1278.

Allen JF. 1993. Control of gene expression by redox potential and the requirement for chloroplast and mitochondrial genomes. J Theor Biol 165:609-631.

Allen JF. 2017. The CoRR hypothesis for genes in organelles. Journal of Theoretical Biology 434:50-57.

Allen JF. 2003a. Cyclic, pseudocyclic and noncyclic photophosphorylation: new links in the chain. Trends Plant Sci 8:15-19.

Allen JF. 2002. Photosynthesis of ATP – electrons, proton pumps, rotors, and Poise. Cell 110:273-276.

Allen JF. 2003b. Why chloroplasts and mitochondria contain genomes. Comparative and functional genomics 4:31-36.

Allen JF. 2015. Why chloroplasts and mitochondria retain their own genomes and genetic systems: Colocation for redox regulation of gene expression. Proc Natl Acad Sci U S A 112:10231-10238.

Andersson B, Anderson JM. 1980. Lateral heterogeneity in the distribution of chlorophyll-protein complexes of the thylakoid membranes of spinach chloroplasts. Biochimica et Biophysica Acta (BBA)-Bioenergetics 593:427-440.

Archibald JM. 2007. Nucleomorph genomes: structure, function, origin and evolution. Bioessays 29:392-402.

Armarego-Marriott T, Sandoval-Ibañez O, Kowalewska Ł. 2020. Beyond the darkness: recent lessons from etiolation and de-etiolation studies. J Exp Bot 71:1215-1225.

Aro E-M, Virgin I, Andersson B. 1993. Photoinhibition of photosystem II. Inactivation, protein damage and turnover. Biochimica et Biophysica Acta (BBA) - Bioenergetics 1143:113-134.

Asada K. 2006. Production and Scavenging of reactive oxygen species in chloroplasts and their functions. Plant Physiology 141:391-396.

Ayliffe MA, Scott NS, Timmis JN. 1998. Analysis of plastid DNA-like sequences within the nuclear genomes of higher plants. Molecular Biology and Evolution 15:738-745.

Badger MR, Andrews TJ, Whitney S, Ludwig M, Yellowlees DC, Leggat W, Price GD. 1998. The diversity and coevolution of Rubisco, plastids, pyrenoids, and chloroplast-based CO2-concentrating mechanisms in algae. Canadian Journal of Botany 76:1052-1071.

Barry BA, Babcock GT. 1987. Tyrosine radicals are involved in the photosynthetic oxygen-evolving system. Proceedings of the National Academy of Sciences 84:7099-7103.

Bekker A, Holland H, Wang P-L, Rumble Iii D, Stein H, Hannah J, Coetzee L, Beukes N. 2004. Dating the rise of atmospheric oxygen. Nature 427:117-120.

Bendich AJ. 2004. Circular chloroplast chromosomes: the grand illusion. The Plant Cell 16:1661-1666.

Bendich AJ. 1987. Why do chloroplasts and mitochondria contain so many copies of their genome? Bioessays 6:279-282.

Bernhard Teicher H, Lindberg Møller B, Vibe Scheller H. 2000. Photoinhibition of photosystem I in field-grown barley (Hordeum vulgare L.): Induction, recovery and acclimation. Photosynthesis Research 64:53-61.

Bhattacharya D, Medlin L. 1995. Phylogeny of plastids A review based on comparisons of small subunit ribosomal RNA coding regions. J Phycol 31:487-496.

Birky CW, Jr., Walsh JB. 1992. Biased gene conversion, copy number, and apparent mutation rate differences within chloroplast and bacterial genomes. Genetics 130:677-683.

Björkholm P, Ernst AM, Hagström E, Andersson SGE. 2017. Why mitochondria need a genome revisited. FEBS Letters 591:65-75.

Bock R, Kössel H, Maliga P. 1994. Introduction of a heterologous editing site into the tobacco plastid genome: the lack of RNA editing leads to a mutant phenotype. The EMBO journal 13:4623-4628.

Bölter B, Soll J, Schulz A, Hinnah S, Wagner R. 1998. Origin of a chloroplast protein importer. Proceedings of the National Academy of Sciences of the United States of America 95:15831-15836.

Borg M, Krueger-Hadfield SA, Destombe C, Collén J, Lipinska A, Coelho SM. 2023. Red macroalgae in the genomic era. New Phytologist 240:471-488.

Bouvier JW, Emms DM, Kelly S. 2024. Rubisco is evolving for improved catalytic efficiency and CO_2 assimilation in plants. Proceedings of the National Academy of Sciences 121:e2321050121.

Bouvier JW, Emms DM, Rhodes T, Bolton JS, Brasnett A, Eddershaw A, Nielsen JR, Unitt A, Whitney SM, Kelly S. 2021. Rubisco adaptation is more limited by phylogenetic constraint than by catalytic trade-off. Molecular Biology and Evolution 38:2880-2896.

Bowes G, Ogren W, Hageman R. 1971. Phosphoglycolate production catalyzed by ribulose diphosphate carboxylase. Biochemical and biophysical research communications 45:716-722.

Boyer PD. 1997. The ATP synthase—a splendid molecular machine. Annual review of biochemistry 66:717-749.

Burki F, Kaplan M, Tikhonenkov DV, Zlatogursky V, Minh BQ, Radaykina LV, Smirnov A, Mylnikov AP, Keeling PJ. 2016. Untangling the early diversification of eukaryotes: a phylogenomic study of the evolutionary origins of Centrohelida, Haptophyta and Cryptista. Proceedings of the Royal Society B: Biological Sciences 283:20152802.

Busch FA. 2020. Photorespiration in the context of Rubisco biochemistry, CO2 diffusion and metabolism. The Plant Journal 101:919-939.

Canfield DE, Poulton SW, Narbonne GM. 2007. Late-Neoproterozoic deep-ocean oxygenation and the rise of animal life. Science 315:92-95.

Carmo-Silva E, Scales JC, Madgwick PJ, Parry MA. 2015. Optimizing Rubisco and its regulation for greater resource use efficiency. Plant, cell & environment 38:1817-1832.

Cavalier-Smith T. 2003. Genomic reduction and evolution of novel genetic membranes and protein-targeting machinery in eukaryote-eukaryote chimaeras (meta-algae). Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 358:109-134.

Cavalier-Smith T. 1982. The origins of plastids. Biological Journal of the Linnean Society 17:289-306.

Chamary JV, Parmley JL, Hurst LD. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. Nature Reviews Genetics 7:98-108.

Chen JH, Chen ST, He NY, Wang QL, Zhao Y, Gao W, Guo FQ. 2020. Nuclear-encoded synthesis of the D1 subunit of photosystem II increases photosynthetic efficiency and crop yield. Nat Plants 6:570-580.

Choi H, Yi T, Ha S-H. 2021. Diversity of Plastid Types and Their Interconversions. Frontiers in Plant Science 12.

Choi IS, Jansen R, Ruhlman T. 2019. Lost and Found: Return of the Inverted Repeat in the Legume Clade Defined by Its Absence. Genome Biol Evol 11:1321-1333.

Chollet R. 1977. The biochemistry of photorespiration. Trends in Biochemical Sciences 2:155-159.

Chumley TW, Palmer JD, Mower JP, Fourcade HM, Calie PJ, Boore JL, Jansen RK. 2006. The complete chloroplast genome sequence of Pelargonium x hortorum: organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. Molecular Biology and Evolution 23:2175-2190.

Correa-Galvis V, Poschmann G, Melzer M, Stühler K, Jahns P. 2016. PsbS interactions involved in the activation of energy dissipation in Arabidopsis. Nature Plants 2:15225.

Croce R, van Amerongen H. 2014. Natural strategies for photosynthetic light harvesting. Nature Chemical Biology 10:492-501.

Crumpton-Taylor M, Grandison S, Png KMY, Bushby AJ, Smith AM. 2011. Control of starch granule numbers in Arabidopsis chloroplasts. Plant Physiology 158:905-916.

DalCorso G, Pesaresi P, Masiero S, Aseeva E, Schünemann D, Finazzi G, Joliot P, Barbato R, Leister D. 2008. A complex containing PGRL1 and PGR5 is involved in the switch between linear and cyclic electron flow in Arabidopsis. Cell 132:273-285.

Daniell H, Lee S-B, Grevich J, Saski C, Quesada-Vargas T, Guda C, Tomkins J, Jansen RK. 2006. Complete chloroplast genome sequences of Solanum bulbocastanum, Solanum lycopersicum and comparative analyses with other Solanaceae genomes. Theoretical and Applied Genetics 112:1503-1518.

Daniell H, Lin C-S, Yu M, Chang W-J. 2016. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. Genome Biology 17:134.

Degnan JH, Rosenberg NA. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends in Ecology & Evolution 24:332-340.

Delaye L, Valadez-Cano C, Pérez-Zamorano B. 2016. How really ancient Is Paulinella chromatophora? PLoS Curr 8.

Demmig-Adams B. 1990. Carotenoids and photoprotection in plants: a role for the xanthophyll zeaxanthin. Biochimica et Biophysica Acta (BBA)-Bioenergetics 1020:1-24.

Deng X-W, Wing RA, Gruissem W. 1989. The chloroplast genome exists in multimeric forms. Proceedings of the National Academy of Sciences 86:4156-4160.

Dodge JD, Crawford RM. 1969. Observations on the fine structure of the eyespot and associated organelles in the dinoflagellate Glenodinium foliaceum. Journal of Cell Science 5:479-493.

Erling Tjus S, Lindberg Møller B, Vibe Scheller H. 1998. Photosystem I is an early target of photoinhibition in barley illuminated at chilling temperatures. Plant Physiology 116:755-764.

Eyre-Walker A. 2006. The genomic rate of adaptive evolution. Trends in Ecology & Evolution 21:569-575.

Ferreira KN, Iverson TM, Maghlaoui K, Barber J, Iwata S. 2004. Architecture of the photosynthetic oxygen-evolving center. Science 303:1831-1838.

Fuentes P, Armarego-Marriott T, Bock R. 2018. Plastid transformation and its application in metabolic engineering. Curr Opin Biotechnol 49:10-15.

Funk HT, Berg S, Krupinska K, Maier UG, Krause K. 2007. Complete DNA sequences of the plastid genomes of two parasitic flowering plant species, Cuscuta reflexa and Cuscuta gronovii. BMC Plant Biol 7:45.

Gabruk M, Mysliwa-Kurdziel B. 2015. Light-dependent protochlorophyllide oxidoreductase: phylogeny, regulation, and catalytic properties. Biochemistry 54:5255-5262.

Gao C, Li T, Zhao X, Wu C, Zhang Q, Zhao X, Wu M, Lian Y, Li Z. 2023. Comparative analysis of the chloroplast genomes of Rosa species and RNA editing analysis. BMC Plant Biol 23:318.

Gilmore AM, Yamamoto HY. 1993. Linear models relating xanthophylls and lumen acidity to non-photochemical fluorescence quenching. Evidence that antheraxanthin explains zeaxanthin-independent quenching. Photosynthesis Research 35:67-78.

Graça AT, Hall M, Persson K, Schröder WP. 2021. High-resolution model of Arabidopsis Photosystem II reveals the structural consequences of digitonin-extraction. Scientific Reports 11:15534.

Greiner S, Lehwark P, Bock R. 2019. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. Nucleic Acids Res 47:W59-w64.

Groth-Malonek M, Wahrmund U, Polsakiewicz M, Knoop V. 2007. Evolution of a pseudogene: exclusive survival of a functional mitochondrial nad7 gene supports Haplomitrium as the earliest liverwort lineage and proposes a secondary loss of RNA editing in Marchantiidae. Molecular Biology and Evolution 24:1068-1074.

Guergova-Kuras M, Boudreaux B, Joliot A, Joliot P, Redding K. 2001. Evidence for two active branches for electron transfer in photosystem I. Proceedings of the National Academy of Sciences 98:4437-4442.

Guidi L, Lo Piccolo E, Landi M. 2019. Chlorophyll fluorescence, photoinhibition and abiotic stress: Does it make any dfference the fact to be a C3 or C4 species? Frontiers in Plant Science 10.

Gururani Mayank A, Venkatesh J, Tran LSP. 2015. Regulation of photosynthesis during abiotic stress-induced photoinhibition. Molecular Plant 8:1304-1320.

Hahn A, Vonck J, Mills DJ, Meier T, Kühlbrandt W. 2018. Structure, mechanism, and regulation of the chloroplast ATP synthase. Science 360:eaat4318.

Harris ME, Meyer G, Vandergon T, Vandergon VO. 2013. Loss of the acetyl-coA carboxylase (accD) gene in Poales. Plant Molecular Biology Reporter 31:21-31.

Hazen RM. 2010. How old is Earth, and how do we know? Evolution: Education and Outreach 3:198-205.

Hedges SB, Blair JE, Venturi ML, Shoe JL. 2004. A molecular timescale of eukaryote evolution and the rise of complex multicellular life. BMC Evol Biol 4:2.

Hermida-Carrera C, Fares MA, Fernández Á, Gil-Pelegrín E, Kapralov MV, Mir A, Molins A, Peguero-Pina JJ, Rocha J, Sancho-Knapik D, et al. 2017. Positively selected amino acid

replacements within the RuBisCO enzyme of oak trees are associated with ecological adaptations. PLoS One 12:e0183970.

Hohmann-Marriott MF, Blankenship RE. 2011. Evolution of photosynthesis. Annual Review of Plant Biology 62:515-548.

Huang CY, Ayliffe MA, Timmis JN. 2003. Direct measurement of the transfer rate of chloroplast DNA into the nucleus. Nature 422:72-76.

Huang W, Zhang S-B, Cao K-F. 2010. Cyclic electron flow plays an important role in photoprotection of tropical trees illuminated at temporal chilling temperature. Plant and Cell Physiology 52:297-305.

Jansen RK, Kaittanis C, Saski C, Lee S-B, Tomkins J, Alverson AJ, Daniell H. 2006. Phylogenetic analyses of Vitis (Vitaceae) based on complete chloroplast genome sequences: effects of taxon sampling and phylogenetic methods on resolving relationships among rosids. BMC Evolutionary Biology 6:32.

Järvi S, Suorsa M, Aro E-M. 2015. Photosystem II repair in plant chloroplasts — Regulation, assisting proteins and shared components with photosystem II biogenesis. Biochimica et Biophysica Acta (BBA) - Bioenergetics 1847:900-909.

Joët T, Cournac L, Horvath EM, Medgyesy P, Peltier G. 2001. Increased sensitivity of photosynthesis to antimycin A induced by inactivation of the chloroplast *ndhB* gene. Evidence for a participation of the NADH-dehydrogenase complex to cyclic electron flow around photosystem I. Plant Physiology 125:1919-1929.

Johnson MP. 2016. Photosynthesis. Essays Biochem 60:255-273.

Johnston IG, Williams BP. 2016. Evolutionary inference across eukaryotes identifies specific pressures favoring mitochondrial gene retention. Cell systems 2:101-111.

Joliot P, Johnson GN. 2011. Regulation of cyclic and linear electron flow in higher plants. Proceedings of the National Academy of Sciences 108:13317-13322.

Juan CA, Pérez de la Lastra JM, Plou FJ, Pérez-Lebeña E. 2021. The chemistry of reactive oxygen species (ROS) revisited: Outlining their role in biological macromolecules (DNA, lipids and proteins) and Induced Pathologies. Int J Mol Sci 22.

Kang B-C, Bae S-J, Lee S, Lee JS, Kim A, Lee H, Baek G, Seo H, Kim J, Kim J-S. 2021. Chloroplast and mitochondrial DNA editing in plants. Nature Plants 7:899-905.

Kapralov MV, Filatov DA. 2007. Widespread positive selection in the photosynthetic Rubisco enzyme. BMC Evolutionary Biology 7:73.

Kapralov MV, Kubien DS, Andersson I, Filatov DA. 2010. Changes in Rubisco Kinetics during the Evolution of C4 Photosynthesis in Flaveria (Asteraceae) Are Associated with Positive Selection on Genes Encoding the Enzyme. Molecular Biology and Evolution 28:1491-1503.

Keeling PJ. 2010. The endosymbiotic origin, diversification and fate of plastids. Philos Trans R Soc Lond B Biol Sci 365:729-748.

Keeling PJ. 2013. The number, speed, and impact of plastid endosymbioses in eukaryotic evolution. Annual Review of Plant Biology 64:583-607.

Kelly S. 2021. The economics of organellar gene loss and endosymbiotic gene transfer. Genome Biology 22:345.

Khakhlova O, Bock R. 2006. Elimination of deleterious mutations in plastid genomes by gene conversion. Plant J 46:85-94.

Kies L, Kremer B. 1979. Function of cyanelles in the thecamoeba Paulinella chromatophora. Naturwissenschaften 66:578-579.

Klukas O, Schubert WD, Jordan P, Krau N, Fromme P, Witt HT, Saenger W. 1999. Localization of two phylloquinones, QK and QK', in an improved electron density map of photosystem I at 4-A resolution. J Biol Chem 274:7361-7367.

Kolodner R, Tewari KK. 1979. Inverted repeats in chloroplast DNA from higher plants. Proc Natl Acad Sci U S A 76:41-45.

Kolodner R, Tewari KK. 1975a. The molecular size and conformation of the chloroplast DNA from higher plants. Biochim Biophys Acta 402:372-390.

Kolodner R, Tewari KK. 1975b. Presence of displacement loops in the covalently closed circular chloroplast deoxyribonucleic acid from higher plants. J Biol Chem 250:8840-8847.

Kolodner RD, Tewari KK. 1975c. Chloroplast DNA from higher plants replicates by both the Cairns and the rolling circle mechanism. Nature 256:708-711.

Kowalewska Ł, Mazur R, Suski S, Garstka M, Mostowska A. 2016. Three-dimensional visualization of the tubular-lamellar transformation of the internal plastid membrane network during runner bean chloroplast biogenesis. The Plant Cell 28:875-891.

Krause K. 2008. From chloroplasts to “cryptic” plastids: evolution of plastid genomes in parasitic plants. Current Genetics 54:111-121.

Kreitman M, Akashi H. 1995. Molecular evidence for natural selection. Annual Review of Ecology and Systematics 26:403-422.

Kudla J, Igloi GL, Metzlaff M, Hagemann R, Kössel H. 1992. RNA editing in tobacco chloroplasts leads to the formation of a translatable psbL mRNA by a C to U substitution within the initiation codon. The EMBO Journal 11:1099-1103.

Kudoh H, Sonoike K. 2002. Irreversible damage to photosystem I by chilling in the light: cause of the degradation of chlorophyll after returning to normal growth temperature. Planta 215:541-548.

Kunnimalaiyaan M, Nielsen BL. 1997. Fine mapping of replication origins (oriA and oriB) in Nicotiana tabacum chloroplast DNA. Nucleic Acids Research 25:3681-3686.

Lane N, Martin W. 2010. The energetics of genome complexity. Nature 467:929-934.

Lawrence ME, Possingham JV. 1986. Microspectrofluorometric measurement of chloroplast DNA in dividing and expanding leaf cells of Spinacia oleracea. Plant Physiol 81:708-710.

Lee S-B, Kaittanis C, Jansen RK, Hostetler JB, Tallon LJ, Town CD, Daniell H. 2006. The complete chloroplast genome sequence of Gossypium hirsutum: organization and phylogenetic relationships to other angiosperms. BMC Genomics 7:61.

Lesch E, Schilling MT, Brenner S, Yang Y, Gruss Oliver J, Knoop V, Schallenberg-Rüdinger M. 2022. Plant mitochondrial RNA editing factors can perform targeted C-to-U editing of nuclear transcripts in human cells. Nucleic Acids Research 50:9966-9983.

Lewis LA, McCourt RM. 2004. Green algae and the origin of land plants. American Journal of Botany 91:1535-1556.

Li H, Nakajima Y, Nango E, Owada S, Yamada D, Hashimoto K, Luo F, Tanaka R, Akita F, Kato K, et al. 2024. Oxygen-evolving photosystem II structures during S1–S2–S3 transitions. Nature 626:670-677.

Li M, Kim C. 2022. Chloroplast ROS and stress signaling. Plant Communications 3.

Li X-P, BjoÈrkman O, Shih C, Grossman AR, Rosenquist M, Jansson S, Niyogi KK. 2000. A pigment-binding protein essential for regulation of photosynthetic light harvesting. Nature 403:391-395.

Li X-P, Phippard A, Pasari J, Niyogi KK. 2002. Structure–function analysis of photosystem II subunit S (PsbS) in vivo. Functional Plant Biology 29:1131-1139.

Lister DL, Bateman JM, Purton S, Howe CJ. 2003. DNA transfer from chloroplast to nucleus is much rarer in Chlamydomonas than in tobacco. Gene 316:33-38.

Liu H, Zhang H, Niedzwiedzki DM, Prado M, He G, Gross ML, Blankenship RE. 2013. Phycobilisomes supply excitations to both photosystems in a megacomplex in cyanobacteria. Science 342:1104-1107.

Logacheva MD, Schelkunov MI, Shtratnikova VY, Matveeva MV, Penin AA. 2016. Comparative analysis of plastid genomes of non-photosynthetic Ericaceae and their photosynthetic relatives. Scientific Reports 6:30042.

Lyons TW, Reinhard CT, Planavsky NJ. 2014. The rise of oxygen in Earth's early ocean and atmosphere. Nature 506:307-315.

Malone LA, Qian P, Mayneord GE, Hitchcock A, Farmer DA, Thompson RF, Swainsbury DJK, Ranson NA, Hunter CN, Johnson MP. 2019. Cryo-EM structure of the spinach cytochrome b6 f complex at 3.6 Å resolution. Nature 575:535-539.

Marano MR, Serra EC, Orellano EG, Carrillo N. 1993. The path of chromoplast development in fruits and flowers. Plant Science 94:1-17.

Mardanov AV, Ravin NV, Kuznetsov BB, Samigullin TH, Antonov AS, Kolganova TV, Skyabin KG. 2008. Complete sequence of the duckweed (Lemna minor) chloroplast genome: structural organization and phylogenetic relationships to other angiosperms. Journal of molecular evolution 66:555-564.

Marin B, Nowack EC, Melkonian M. 2005. A plastid in the making: evidence for a second primary endosymbiosis. Protist 156:425-432.

Martin Avila E, Gisby MF, Day A. 2016. Seamless editing of the chloroplast genome in plants. BMC Plant Biol 16:168.

Martin W, Herrmann RG. 1998. Gene transfer from organelles to the nucleus: How much, what Happens, and why? 1. Plant Physiology 118:9-17.

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. Proceedings of the National Academy of Sciences 99:12246-12251.

Matsumoto T, Shinozaki F, Chikuni T, Yabuki A, Takishita K, Kawachi M, Nakayama T, Inouye I, Hashimoto T, Inagaki Y. 2011. Green-colored plastids in the dinoflagellate genus Lepidodinium are of core chlorophyte origin. Protist 162:268-276.

Mattoo AK, Pick U, Hoffman-Falk H, Edelman M. 1981. The rapidly metabolized 32,000-dalton polypeptide of the chloroplast is the "proteinaceous shield" regulating photosystem II electron transport and mediating diuron herbicide sensitivity. Proceedings of the National Academy of Sciences 78:1572-1576.

Mazor Y, Borovikova A, Caspy I, Nelson N. 2017. Structure of the plant photosystem I supercomplex at 2.6 Å resolution. Nature Plants 3:17014.

McFadden GI. 2014. Origin and evolution of plastids and photosynthesis in eukaryotes. Cold Spring Harb Perspect Biol 6:a016105.

McFadden GI, Reith ME. 1996. Plastid in human parasites. Nature 381.

Meyer M, Griffiths H. 2013. Origins and diversity of eukaryotic CO2-concentrating mechanisms: lessons for the future. Journal of experimental botany 64:769-786.

Millen RS, Olmstead RG, Adams KL, Palmer JD, Lao NT, Heggie L, Kavanagh TA, Hibberd JM, Gray JC, Morden CW. 2001. Many parallel losses of *infA* from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus. The Plant Cell 13:645-658.

Mitchell P. 1976. Possible molecular mechanisms of the protonmotive function of cytochrome systems. Journal of Theoretical Biology 62:327-367.

Mohanta TK, Mishra AK, Khan A, Hashem A, Abd Allah EF, Al-Harrasi A. 2020. Gene loss and evolution of the plastome. Genes (Basel) 11.

Moustakas M. 2022. Plant photochemistry, reactive oxygen species, and photoprotection. Photochem 2:5-8.

Müh F, Zouni A. 2020. Structural basis of light-harvesting in the photosystem II core complex. Protein Sci 29:1090-1119.

Muller HJ. 1964. The relation of recombination to mutational advance. Mutat Res 106:2-9.

Murata N, Takahashi S, Nishiyama Y, Allakhverdiev SI. 2007. Photoinhibition of photosystem II under environmental stress. Biochimica et Biophysica Acta (BBA)-Bioenergetics 1767:414-421.

Murchie EH, Ruban AV. 2020. Dynamic non-photochemical quenching in plants: from molecular mechanism to productivity. The Plant Journal 101:885-896.

Nelson N, Junge W. 2015. Structure and energy tansfer in photosystems of oxygenic photosynthesis. Annual Review of Biochemistry 84:659-683.

Nishiyama Y, Allakhverdiev SI, Yamamoto H, Hayashi H, Murata N. 2004. Singlet oxygen inhibits the repair of photosystem II by suppressing the translation elongation of the D1 protein in Synechocystis sp. PCC 6803. Biochemistry 43:11321-11330.

Nishiyama Y, Murata N. 2014. Revised scheme for the mechanism of photoinhibition and its application to enhance the abiotic stress tolerance of the photosynthetic machinery. Applied Microbiology and Biotechnology 98:8777-8796.

Nishiyama Y, Yamamoto H, Allakhverdiev SI, Inaba M, Yokota A, Murata N. 2001. Oxidative stress inhibits the repair of photodamage to the photosynthetic machinery. The EMBO journal.

Nitschke W, Rutherford AW. 1991. Photosynthetic reaction centres: variations on a common structural theme? Trends in Biochemical Sciences 16:241-245.

Noffke N, Christian D, Wacey D, Hazen RM. 2013. Microbially induced sedimentary structures recording an ancient ecosystem in the ca. 3.48 billion-year-old Dresser Formation, Pilbara, Western Australia. Astrobiology 13:1103-1124.

Nursall JR. 1959. Oxygen as a prerequisite to the origin of the metazoa. Nature 183:1170-1172.

Oldenburg DJ, Bendich AJ. 2015. DNA maintenance in plastids and mitochondria of plants. Frontiers in Plant Science 6.

Oldenburg DJ, Bendich AJ. 2016. The linear plastid chromosomes of maize: terminal sequences, structures, and implications for DNA replication. Current Genetics 62:431-442.

Oldenburg DJ, Bendich AJ. 2004. Most chloroplast DNA of maize seedlings in linear molecules with defined ends and branched forms. J Mol Biol 335:953-970.

Oldenkott B, Yamaguchi K, Tsuji-Tsukinoki S, Knie N, Knoop V. 2014. Chloroplast RNA editing going extreme: more than 3400 events of C-to-U editing in the chloroplast transcriptome of the lycophyte Selaginella uncinata. Rna 20:1499-1506.

Oliver N, Avramov AP, Nürnberg DJ, Dau H, Burnap RL. 2022. From manganese oxidation to water oxidation: assembly and evolution of the water-splitting complex in photosystem II. Photosynthesis Research 152:107-133.

Oliver T, Kim TD, Trinugroho JP, Cordón-Preciado V, Wijayatilake N, Bhatia A, Rutherford AW, Cardona T. 2023. The evolution and evolvability of photosystem II. Annual Review of Plant Biology 74:225-257.

Ostermeier M, Garibay-Hernández A, Holzer VJC, Schroda M, Nickelsen J. 2024. Structure, biogenesis, and evolution of thylakoid membranes. The Plant Cell.

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

Palmer JD, Osorio B, Aldrich J, Thompson WF. 1987. Chloroplast DNA evolution among legumes: loss of a large inverted repeat occurred prior to other sequence rearrangements. Current Genetics 11:275-286.

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

Park S, An B, Park S. 2018. Reconfiguration of the plastid genome in Lamprocapnos spectabilis: IR boundary shifting, inversion, and intraspecific variation. Scientific Reports 8:13568.

Park S, An B, Park S. 2020. Recurrent gene duplication in the angiosperm tribe Delphinieae (Ranunculaceae) inferred from intracellular gene transfer events and heteroplasmic mutations in the plastid matK gene. Scientific Reports 10:2720.

Perez Boerema A, Aibara S, Paul B, Tobiasson V, Kimanius D, Forsberg BO, Wallden K, Lindahl E, Amunts A. 2018. Structure of the chloroplast ribosome with chl-RRF and hibernation-promoting factor. Nature Plants 4:212-217.

Perry AS, Wolfe KH. 2002. Nucleotide substitution rates in legume chloroplast DNA depend on the presence of the inverted repeat. J Mol Evol 55:501-508.

Ponce-Toledo RI, Deschamps P, López-García P, Zivanovic Y, Benzerara K, Moreira D. 2017. An early-branching freshwater cyanobacterium at the origin of plastids. Current Biology 27:386-391.

Pospíšil P. 2012. Molecular mechanisms of production and scavenging of reactive oxygen species by photosystem II. Biochimica et Biophysica Acta (BBA) - Bioenergetics 1817:218-231.

Powles SB. 1984. Photoinhibition of photosynthesis induced by visible light. Annual review of plant physiology 35:15-44.

Prywes N, Phillips NR, Tuck OT, Valentin-Alvarado LE, Savage DF. 2023. Rubisco function, evolution, and engineering. Annual Review of Biochemistry 92:385-410.

Pyke K. 2007. Plastid biogenesis and differentiation. In: Bock R, editor. Cell and Molecular Biology of Plastids. Berlin, Heidelberg: Springer Berlin Heidelberg. p. 1-28.

Raymond J. 2008. Coloring in the tree of life. Trends in Microbiology 16:41-43.

Reumann S, Keegstra K. 1999. The endosymbiotic origin of the protein import machinery of chloroplastic envelope membranes. Trends Plant Sci 4:302-307.

Richardson LGL, Schnell DJ. 2019. Origins, function, and regulation of the TOC–TIC general protein import machinery of plastids. Journal of Experimental Botany 71:1226-1238.

Robbins EHJ, Kelly S. 2023. The evolutionary constraints on angiosperm chloroplast adaptation. Genome Biology and Evolution 15.

Rodrigues CH, Pires DE, Ascher DB. 2018. DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. Nucleic Acids Res 46:W350-w355.

Rodriguez-Concepcion M, Stange C. 2013. Biosynthesis of carotenoids in carrot: an underground story comes to light. Arch Biochem Biophys 539:110-116.

Rogers MB, Gilson PR, Su V, McFadden GI, Keeling PJ. 2007. The complete chloroplast genome of the chlorarachniophyte Bigelowiella natans: evidence for independent origins of

chlorarachniophyte and euglenid secondary endosymbionts. Molecular Biology and Evolution 24:54-62.

Rowan BA, Oldenburg DJ, Bendich AJ. 2010. RecA maintains the integrity of chloroplast DNA molecules in Arabidopsis. J Exp Bot 61:2575-2588.

Sagan L. 1967. On the origin of mitosing cells. Journal of theoretical biology 14:225-IN226.

Sánchez-Baracaldo P, Cardona T. 2020. On the origin of oxygenic photosynthesis and Cyanobacteria. New Phytol 225:1440-1446.

Saski C, Lee S-B, Daniell H, Wood TC, Tomkins J, Kim H-G, Jansen RK. 2005. Complete chloroplast genome sequence of Glycine max and comparative analyses with other legume genomes. Plant molecular biology 59:309-322.

Saski C, Lee S-B, Fjellheim S, Guda C, Jansen RK, Luo H, Tomkins J, Rognli OA, Daniell H, Clarke JL. 2007. Complete chloroplast genome sequences of Hordeum vulgare, Sorghum bicolor and Agrostis stolonifera, and comparative analyses with other grass genomes. Theoretical and Applied Genetics 115:571-590.

Savir Y, Noor E, Milo R, Tlusty T. 2010. Cross-species analysis traces adaptation of Rubisco toward optimality in a low-dimensional landscape. Proceedings of the National Academy of Sciences 107:3475-3480.

Scharff LB, Koop HU. 2006. Linear molecules of tobacco ptDNA end at known replication origins and additional loci. Plant Mol Biol 62:611-621.

Schlüter U, Bouvier JW, Guerreiro R, Malisic M, Kontny C, Westhoff P, Stich B, Weber AP. 2023. Brassicaceae display variation in efficiency of photorespiratory carbon-recapturing mechanisms. Journal of Experimental Botany 74:6631-6649.

Schnepf E, Elbrächter M. 1988. Cryptophycean-like double membrane-bound chloroplast in the dinoflagellate, Dinophysis Ehrenb.: evolutionary, phylogenetic and toxicological implications. Botanica Acta 101:196-203.

Shen L, Tang K, Wang W, Wang C, Wu H, Mao Z, An S, Chang S, Kuang T, Shen J-R, et al. 2022. Architecture of the chloroplast PSI–NDH supercomplex in Hordeum vulgare. Nature 601:649-654.

Shrestha B, Gilbert LE, Ruhlman TA, Jansen RK. 2020. Rampant nuclear Ttansfer and substitutions of plastid genes in Passiflora. Genome Biology and Evolution 12:1313-1329.

Small ID, Schallenberg-Rüdinger M, Takenaka M, Mireau H, Ostersetzer-Biran O. 2020. Plant organellar RNA editing: what 30 years of research has revealed. The Plant Journal 101:1040-1056.

Smith DR. 2015. Mutation rates in plastid genomes: They are lower than you might think. Genome Biology and Evolution 7:1227-1234.

Smith DR, Crosby K, Lee RW. 2011. Correlation between nuclear plastid DNA abundance and plastid number supports the limited transfer window hypothesis. Genome Biology and Evolution 3:365-371.

Sonoike K, Terashima I, Iwaki M, Itoh S. 1995. Destruction of photosystem I iron-sulfur centers in leaves of Cucumis sativus L. by weak illumination at chilling temperatures. FEBS Letters 362:235-238.

Spang A, Saw JH, Jørgensen SL, Zaremba-Niedzwiedzka K, Martijn J, Lind AE, Van Eijk R, Schleper C, Guy L, Ettema TJ. 2015. Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature 521:173-179.

Stegemann S, Hartmann S, Ruf S, Bock R. 2003. High-frequency gene transfer from the chloroplast genome to the nucleus. Proceedings of the National Academy of Sciences 100:8828-8833.

Steiner JM, Löffelhardt W. 2002. Protein import into cyanelles. Trends in plant science 7:72-77.

Stern DB, Goldschmidt-Clermont M, Hanson MR. 2010. Chloroplast RNA metabolism. Annual Review of Plant Biology 61:125-155.

Stern DB, Lonsdale DM. 1982. Mitochondrial and chloroplast genomes of maize have a 12-kilobase DNA sequence in common. Nature 299:698-702.

Sun T, Yuan H, Cao H, Yazdani M, Tadmor Y, Li L. 2018. Carotenoid metabolism in plants: The Role of Plastids. Molecular Plant 11:58-74.

Takenaka M, Zehrmann A, Verbitskiy D, Härtel B, Brennicke A. 2013. RNA editing in plants and its evolution. Annual review of genetics 47:335-352.

Tcherkez GG, Farquhar GD, Andrews TJ. 2006. Despite slow catalysis and confused substrate specificity, all ribulose bisphosphate carboxylases may be nearly perfectly optimized. Proc Natl Acad Sci U S A 103:7246-7251.

Tengs T, Dahlberg OJ, Shalchian-Tabrizi K, Klaveness D, Rudi K, Delwiche CF, Jakobsen KS. 2000. Phylogenetic analyses indicate that the 19′ hexanoyloxy-fucoxanthin-containing dinoflagellates have tertiary plastids of haptophyte origin. Molecular Biology and Evolution 17:718-729.

The Rice Chromosome 10 Sequencing C. 2003. In-Depth View of Structure, Activity, and Evolution of Rice Chromosome 10. Science 300:1566-1569.

Theis J, Schroda M. 2016. Revisiting the photosystem II repair cycle. Plant Signaling & Behavior 11:e1218587.

Tikkanen M, Mekala NR, Aro E-M. 2014. Photosystem II photoinhibition-repair cycle protects Photosystem I from irreversible damage. Biochimica et Biophysica Acta (BBA) - Bioenergetics 1837:210-215.

Timmis JN, Ayliffe MA, Huang CY, Martin W. 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nature Reviews Genetics 5:123-135.

Tsudzuki T, Wakasugi T, Sugiura M. 2001. Comparative analysis of RNA editing sites in higher plant chloroplasts. J Mol Evol 53:327-332.

Ueda M, Nishikawa T, Fujimoto M, Takanashi H, Arimura S, Tsutsumi N, Kadowaki K. 2008. Substitution of the gene for chloroplast RPS16 was assisted by generation of a dual targeting signal. Mol Biol Evol 25:1566-1575.

Valegård K, Hasse D, Andersson I, Gunn LH. 2018. Structure of Rubisco from Arabidopsis thaliana in complex with 2-carboxyarabinitol-1,5-bisphosphate. Acta Crystallogr D Struct Biol 74:1-9.

Vangone A, Schaarschmidt J, Koukos P, Geng C, Citro N, Trellet ME, Xue LC, Bonvin AMJJ. 2018. Large-scale prediction of binding affinity in protein–small ligand complexes: the PRODIGY-LIG web server. Bioinformatics 35:1585-1587.

Vergara-Cruces Á, Pramanick I, Pearce D, Vogirala VK, Byrne MJ, Low JKK, Webster MW. 2024. Structure of the plant plastid-encoded RNA polymerase. Cell 187:1145-1159.e1121.

Wang D, Timmis JN. 2013. Cytoplasmic Organelle DNA Preferentially Inserts into Open Chromatin. Genome Biology and Evolution 5:1060-1064.

Wang R-J, Cheng C-L, Chang C-C, Wu C-L, Su T-M, Chaw S-M. 2008. Dynamics and evolution of the inverted repeat-large single copy junctions in the chloroplast genomes of monocots. BMC Evolutionary Biology 8:36.

Wang Z, Wu M. 2015. An integrated phylogenomic approach toward pinpointing the origin of mitochondria. Scientific Reports 5:7949.

Watanabe MM, Suda S, Inouya I, Sawaguchi T, Chihara M. 1990. Lepidodium vidre gen. sp. Nov. (gymnodinaiales, Dinophyta), a green dinoglagellate with a chlorophyll a and b containing endosymbiont. Journal of Phycology 26:741-751.

Weber AP, Linka M, Bhattacharya D. 2006. Single, ancient origin of a plastid metabolite translocator family in Plantae from an endomembrane-derived ancestor. Eukaryotic Cell 5:609-612.

Wicke S, Schneeweiss GM, dePamphilis CW, Müller KF, Quandt D. 2011. The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. Plant Molecular Biology 76:273-297.

Wise RR. 2006. The diversity of plastid form and function. In: Wise RR, Hoober JK, editors. The Structure and Function of Plastids. Dordrecht: Springer Netherlands. p. 3-26.

Wolfe KH, Li WH, Sharp PM. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc Natl Acad Sci U S A 84:9054-9058.

Wood CW, Ibarra AA, Bartlett GJ, Wilson AJ, Woolfson DN, Sessions RB. 2020. BAlaS: fast, interactive and accessible computational alanine-scanning using BudeAlaScan. Bioinformatics 36:2917-2919.

Xu P, Widmer G, Wang Y, Ozaki LS, Alves JM, Serrano MG, Puiu D, Manque P, Akiyoshi D, Mackey AJ. 2004. The genome of Cryptosporidium hominis. Nature 431:1107-1112.

Yagi Y, Nakamura T, Small I. 2014. The potential for manipulating RNA with pentatricopeptide repeat proteins. Plant J 78:772-782.

Yamamoto H, Nakayama T, Chichester C. 1962. Studies on the light and dark interconversions of leaf xanthophylls. Archives of Biochemistry and Biophysics 97:168-173.

Zaremba-Niedzwiedzka K, Caceres EF, Saw JH, Bäckström D, Juzokaite L, Vancaester E, Seitz KW, Anantharaman K, Starnawski P, Kjeldsen KU. 2017. Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature 541:353-358.

Zentgraf U, Andrade-Galan AG, Bieker S. 2022. Specificity of H2O2 signaling in leaf senescence: is the ratio of H2O2 contents in different cellular compartments sensed in Arabidopsis plants? Cellular & Molecular Biology Letters 27:4.

Zhu A, Guo W, Gupta S, Fan W, Mower JP. 2016. Evolutionary dynamics of the plastid inverted repeat: the effects of expansion, contraction, and loss on substitution rates. New Phytol 209:1747-1756.