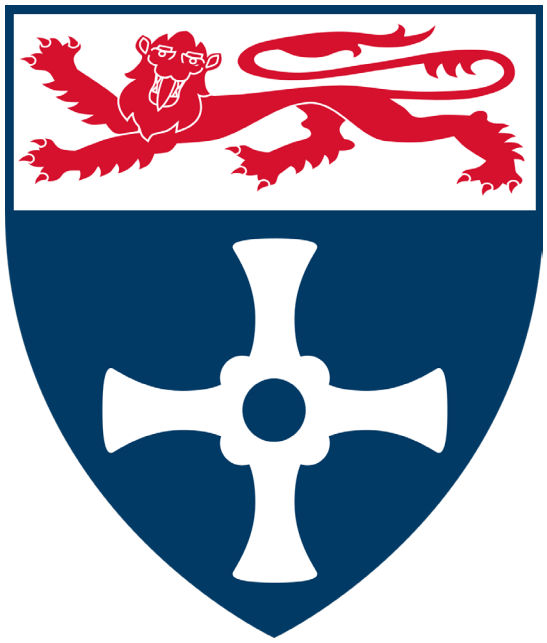
Exploring the intricacies of Rubisco expression and evolution across marine ecosystems and its ability to enhance photosynthesis in key agricultural crops.



Doctor of Philosophy Thesis

Iain James Hope

Newcastle University

Abstract

Marine environments are responsible for 50% of the net primary fixation of carbon globally. Predominantly, the Calvin Benson Basham cycle fixes carbon with the initial enzyme being Rubisco. Rubisco has been shown to be a significant bottleneck in photosynthesis, due to a slow catalytic rate and promiscuity of the enzyme. Previous Rubisco studies have largely focused on land plants, and little is known about the diversity and abundance of Rubisco within marine environments. Through analysing publicly available metagenomes and metatranscriptomes from the Earth's seas and oceans; we have begun to paint a picture of the global abundance and variation of Rubisco alongside adjoining photosynthetic apparatus. Additionally phylogenetic and sequence analysis alongside machine learning models was used to highlight regions of the Rubisco gene under selective pressure, linking selection to metagenome environment. Finally, the big leaf photosynthesis model was applied to simulate heterogenous expression of aquatic Rubisco in wheat over an entire growing season. Transcriptomic analysis showed a significant correlation with temperature across the earth's oceans. On top of this, sequence analysis of Rubisco structures provides evidence for adaption of the large and small subunit to differing environmental temperatures, additionally demonstrating residues that diverge in warm(>20^{o}C) and cold (<10^{O}C) ecological systems. In particular the ßE-ßF loop of the form ID Rubisco small subunit was highlighted as a region under widespread positive selection across phylogenetic lineages in marine environments. Finally, photosynthesis modelling efforts showed that Rubisco from form ID Rubisco, particularly *G. monillis* could simultaneously improve carbon fixation and improve water usage in wheat. Overall, this study provides an invaluable insight into the regulation and evolution of Rubisco in the Earth's ocean, generating an impetus for further investigation. Additionally modelling efforts demonstrate the potential benefits of expressing aquatic Rubisco in economically important crop species to improve future food security.

Table of Contents

List of Figures

List of Tables

Abbreviations

ECM	Enzyme-carbon-magnesium^{2+}
Epyc1	Pyrenoidal linker protein 1
ESM	Earth system model
fba	Fructose-bisphosphate aldolase
fbp	Fructose-Bisphosphatase
ffh	Signal recognition particle subunit SRP54
Flnr	Rubisco percentage of total soluble protein
ftsY	Fused signal recognition particle receptor
gapdh	glyceraldehyde-3-phosphate dehydrogenase
Gb	Boundary layer conductance
glyA	Glycine hydroxymethyltransferase
glyk	D-glycerate 3-kinase
GOI	Genes of interest
Gs	Stomatal conductance
Ha	Heat activation energy
hla3	ABC-type transporter 3
hpr1	Glycerate dehydrogenase
hs	external relative humidity
Ib	Direct beam radiation
Ibs	Scattered beam radiation
Id	Diffuse beam radiation
Jmax	Maximum rate of electron transport
kae1	N6-L-threonylcarbamoyladenine synthase
Kb	Direct beam extiction coefficient
Kc	Rubisco Micahelis Menten constant for CO2
KEGG	Kyoto Encyclopedia of Genes and Genomes
Kn	Leaf nitrogen extinction coefficient
Ko	Rubisco Micahelis Menten constant for O2
LAI	Leaf area index
lhca1	Light harvesting complex 1
lhca2	Light harvesting complex 2
LSU	large subunit
manA	Leucyl-tRNA synthetase
MES	Mesopelagic
N	Nitrogen
Oi	Intercellular oxygen

PaCP	Dry air density
Pair	Air pressure
PCA	Principal component anaylsis
PDBP	D-Glycero-2,3-pentodiulose 1,5-bisphosphate
petB	Cytochrome b6
petF	Ferredoxin-1
pgk	Phosphoglycerate kinase
pgp	Phosphoglycolate phosphatase
pheS	Phenylalanyl-tRNA synthetase alpha chain
PMG	Phylogenetic marker genes
prk	Phosphoribulokinase
psaA	Photosystem I P700 chlorophyll a apoprotein A
psaB	Photosystem I P700 chlorophyll a apoprotein B
psbA	Photosystem II reaction center A
psbL	Photosystem II reaction center L
ptca1	*Phaeodactylum tricornutum* beta-carbonic anhydrase 1
ptslc4A1	Putative solute carrier 4A1
Pyco1	Pyrenoidal linker protein 1
Raf1	Rubisco accumulation factor 1
Raf2	Rubisco accumulation factor 1
RbcL	Rubisco large subunit
RbcS	Rubisco small subunit
RbcX	Rubisco assembly chaperone protein
rbh	leaf boundary layer resistance to heat
rbw	leaf boundary layer resistance to water vapour
Rca	Rubisco activase
Rgas	Universal gas constant
RLP	Rubisco-like-protein
Rn	sum of net radiation absorbed by sunlit and shaded leaf partitions
rpiA	Ribose 5-phosphate isomerase A
rs	Stomatal resistance
rsw	Stomatal resistance
RuBP	Ribulose 1,5-bisphosphate
sbp	Sedoheptulose-1,7-bisphosphatase
sbtA	Sodium-dependent bicarbonate transporter A
Sc/o	Specificity of Rubisco
serS	Seryl-tRNA synthetase
SFL	Surface layers

SLA	Specific leaf area
ß-ca	Beta carbonic anhydrase
SSU	Small subunit
SVP	Saturated vapour pressure
Ta	Temperature air
TIC	Total inorganic carbon
tka	Transketolase
Tp	Triose phosphate utilisation rate
valS	Valyl-tRNA synthetase
Vcmax	Maximum carboxylation rate of Rubisco
Vcmax25	Maximum carboxylation rate of Rubisco at 25o
VPD	Vapour pressure deficit
XuBP	Xylulose 1,5-bisphosphate
γ	Psychometric gas constant
ychF	Ribosome-binding ATPase
θc	Soil saturation point
θw	Field wilting point

***In this study gene names and protein names follow the bacterial genetic nomenclature rules**

Introduction

1.1 Rubisco Biogenesis and Function

1.1.1 Role of Rubisco in photosynthesis

Rubisco the primary enzyme in the Calvin Benson Basham (CBB) cycle. It is responsible for carboxylating Ribulose 1,5-bisphosphate (RuBP) to form two molecules of 3-phosphoglycerate (3PGA) (Andersson, 2008). This is a multistep reaction involving the enolization, carboxylation and cleavage of RuBP with multiple state barriers (Figure 1.1). The resulting 3PGA then proceeds through the CBB cycle, where it can be converted into more complex sugars or starch (Figure 1.1) (Andersson, 2008).

Rubisco has been a common focus in photosynthesis research for many decades as it is considered to be a catalytically inefficient enzyme in need of optimisation (Whitney et al., 2011). This is due to its slow turnover rate and its promiscuity, frequently binding O_2 which results in the formation of 2-phosphoglycolate (2PGO) (Keys, 1986); a compound which is toxic to photosynthetic organisms through the inhibition of multiple other CBB enzymes (Flügel et al., 2017). As a result, 2PGO must be processed through an energetically costly process called photorespiration to recycle 2PGO back to the usable RuBP for future carboxylation (Cavanagh et al., 2022).

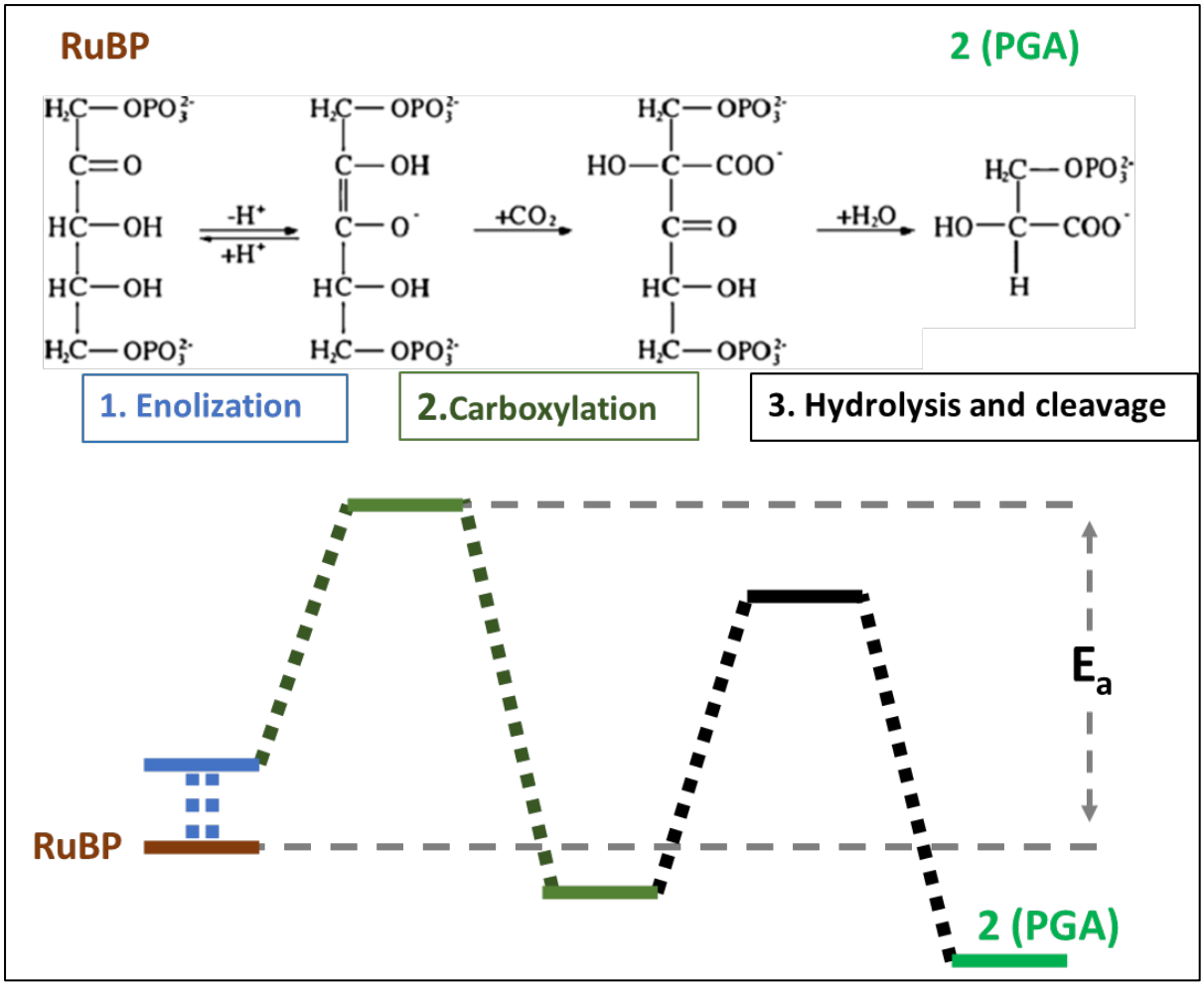


Figure 1.1 – The stepwise process for the enolization, carboxylation and hydrolysis and cleavage of RuBP by the Rubisco enzyme. A representation of the required activation energy is shown below.

1.1.2 Form III Rubisco – The Evolution of Rubisco with Carboxylation Capacity

Carboxylation capacity is thought to have evolved from an ancestral Rubisco-like-protein (RLP). These RLPs have structural commonalities with modern day Rubisco and share similarities in their preferred substrate, both binding 5-carbon sugars with a phosphorylated first carbon (Erb and Zarzycki, 2018). However, RLPs cannot carboxylate RuBP and are generally associated with methione salvaging pathways (Erb and Zarzycki, 2018). The first Rubisco form that have evolved carboxylation capacity is defined as form III Rubisco. Form III Rubisco is predominantly found in Archaea and forms an L_2 or $(L_2)_{5n}$ holoenzyme structure

(Tabita et al., 2008b) Unlike other forms of Rubisco, form III in Archaea are not involved in the CBB cycle and utilise RuBP derived from the metabolism of nucleotides(Ashida et al., 2005), (Sato et al., 2007). However, (Frolov et al., 2019) demonstrated the chemolithoautotrophic bacteria, *Thermodesulfobium acidiphilum,* possesses a form III Rubisco structure involved in the CBB cycle.

Form III Rubisco is thought to have evolved 3.5 billion years ago which has significance, as at this time the earth was encompassed in a CO_2 rich atmosphere where O_2 concentrations were minimal. Thus, the ability to discriminate between CO_2 and O_2 was not a necessity for this ancestral form of Rubisco (Tabita et al., 2008a).

1.1.3 Form II Rubisco- The dimerization of the monomeric form

Form II Rubisco is found in proteobacteria and dinoflagellates and is characterised as the first extant, ancestral Rubisco form to solely perform a role in the CBB cycle. In form II Rubisco, single subunits dimerize in a $(L_2)_{1\text{-}3n}$ configuration to form the holoenzyme structure with Rubisco sequences grouping phylogenetically based on multiples of dimers in the holoenzyme structure (Liu et al., 2022). Previous studies on the kinetic properties of form II Rubisco have shown that enzymatic turnover rates vary to a far higher degree than that of plants (Davidi et al., 2020). Certain form II species possess catalytic rates ~4 fold faster than the fastest plant Rubiscos. However, these form II Rubisco structures also are extremely poor at discriminating between CO_2 and O_2. Often proteobacterial species possessing form II Rubisco are associated with chemolithoautotrophic bacteria capable of deriving energy from sulphur oxidation to fuel Rubisco and the CBB cycle (Hanson and Tabita, 2001). In the Earth's oceans expression of sulphur oxidation genes, expression of form II Rubisco genes and depth in the water column, providing anoxic conditions have all been positively correlated (Baltar et al., 2023). This suggests that organisms possessing form II Rubisco favour environments where there is not a requirement for a high specificity of Rubisco.

1.1.4 Form I Rubisco- The incorporation of small subunits

Form I Rubisco is the most abundant protein on the planet (Raven, 2013) and differ from the ancestral form II structures due to the incorporation of small subunits (SSUs) into the holoenzyme. Although the small subunit does not interact directly with the active site of

Rubisco its considered necessary for proper catalytic function and structure (Mao et al., 2023) and plays a role in overall Rubisco regulation (Wietrzynski et al., 2021).

Across all lineages form I Rubisco is characterised the by the L_8S_8 holoenzyme structure with the exception of a small clade of bacteria, possessing Rubisco large subunits, structured as an octamer, but lacking small subunits. This is considered to be the evolutionary linker between form II and I Rubisco sequences (Banda et al., 2020). Increasing O_2 concentrations in the atmosphere necessitated increased specificity of Rubisco for intended function, this appears to be the main evolutionary driver behind the incorporation of small subunits. Small subunits of Rubisco are believed to consolidate the active site of Rubisco increasing specificity (Mao et al., 2023). Evidence for this is theory is derived from the poor specificity of form II Rubisco species (Davidi et al., 2020)as well as the form I Rubisco structure, lacking SSUs outlined above (Banda et al., 2020).

The mechanism of increased specificity by the incorporation of SSUs has shown to be further exploited by plants. Many plants possess multiple isoforms of the small subunits with subtle structural differences. Under differing environmental conditions plants will differentially express the small subunit isoforms to optimise Rubisco kinetics to the environment by modifying kinetic rate and specificity (Cavanagh et al., 2023). Green algae also possess multiple isoforms of the small subunit, however this differential expression has yet to be demonstrated (Atkinson et al., 2017).

Form I Rubisco can be divided into two groups ‘green and ‘red’ Rubisco which represent an evolutionary divergence from a proteobacterial common ancestor (Tabita et al., 2008b) These groups can be further divided into sub-forms based on phylogenetic grouping. From the ‘green’ lineage there is form IA (derived from α-cyanobacteria possessing α-carboxysomes and proteobacteria (Cabello-Yeves et al., 2022) and Form IB (derived from β-cyanobacteria possessing β-carboxysomes, green algae and plants (Whitehead et al., 2014) In the red lineage there is form IC derived from proteobacteria and form ID derived from red algae (Bracher et al., 2017).

1.1.5 Rubisco structure- The highly conserved large subunit

Across forms I and II of Rubisco the structure of the large subunit (LSU) is highly conserved. In short the LSU subunit consists of two subdomains; the N-terminal domain and the C-

terminal domain. The N-terminal domain consists of four β-sheets, interspersed with occasional helices. Whereas the larger C-terminal domain forms a barrel like structure consisting of 8 β/α parallel units (Andersson and Backlund, 2008). The conserved active site residues are located with four residues on the N-terminal domain and 6 on the C-terminal domain (Whitney et al., 2011). Visually the forms do not differ in their LSU structure, with the exception off an extended βB-βC loop in the N-terminal domain of form IC (Oh et al., 2023) and some form II organisms.

1.1.6 The Rubisco Small Subunit

Unlike the LSU there is greater variation in the structure and sequence space of the small subunit in form I organisms with only ~30% percentage identity across species (Bracher et al., 2017). The small subunit consists of two α-helices and four anti-parallel β-sheets (Knight et al., 1990). The small subunits cap each end of L_2 dimer in the holoenzyme interacting with the β/α parallel units on the C-terminal of each of the LSUs as well as loosely interacting with the two other adjacent SSUs (Knight et al., 1990). The interactions of the SSUs at the poles of the Rubisco bring the SSUs into close contact with the central pore of the Rubisco enzyme. How the SSUs interact with this central pore presents as a significant structural difference between green and red type Rubisco enzymes.

Firstly green type SSUs are characterised by a significantly extended βA–βB loop relative to that of red type SSUs (Figure 1.2). In green type organisms this can be up to 31 residues long (Goudet et al., 2020) whereas in red type SSUs this structure is generally 10 amino acids long (Joshi et al., 2015). This βA–βB loop in green type organisms lines the central pore of the enzyme in green type Rubisco. Alternatively in red type Rubisco the SSU's have a C-terminal hairpin loop formed between two beta sheets (βE–βF) (Figure 1.2). This structure is completely absent in green type rubisco (Oh et al., 2023). Like the extended βA–βB loop in green type Rubisco this C-terminal hairpin interacts with the central pore of the enzyme. However in red type rubisco this extension extends down into the solvent channel and forms a β-barrel structure with adjacent small subunits (Joshi et al., 2015).

Both the role of the extended βA–βB in green type Rubisco and the C-terminal hairpin loop in red type Rubisco are thought to play a role in the aperture diameter of the solvent channel with aperture diameter being a contributing factor to Rubisco specificity (Esquivel

et al., 2013, Poudel et al., 2020). The more invasive interaction between the red type SSU and solvent channel may explain the higher specificity ratios which are commonly found in that of red type Rubiscos , (Spreitzer et al., 2005, Oh et al., 2023).

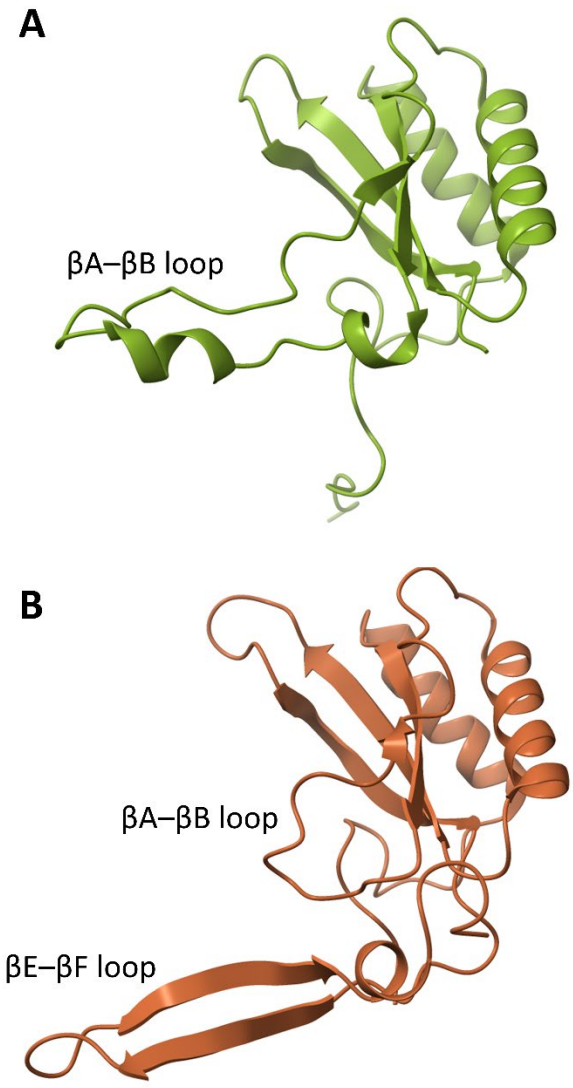


Figure 1.2- A The crystal structure of the Rubisco small subunit derived from the form IB *Chalmydomonas reinhardtii (*RCSB: 1GK8) (Taylor et al., 2001)**.** Highlighted is the extended βA–βB loop. **B** The crystal structure of the Rubisco small subunit derived from the form ID *Thalassiosira antarctica* (RCSB:5MZ2) (Valegård et al., 2018). Highlighted is the smaller βA–

βB loop and the βE–βF of the carboxy terminus which is not present in form IA and IB organisms.

1.1.7 Rubisco biogenesis

Another striking difference between the Rubisco forms are the accessory proteins that are necessary for the biogenesis and function of the Rubisco enzyme. The simplest form of Rubisco associated with the CBB cycle, form II, also has the most generalised requirements for its biogenesis; simply requiring the generic GroEL/ GroES bacterial chaperones for its assembly (Figure 1.3). The GroEL/GroES complex forms a cage like structure around the form II LSU to stabilise folding and prevent aggregation, upon release LSUs will rapidly dimerise (Brinker et al., 2001).

The evolution of SSUs necessitated a more complex assembly framework. This is epitomised by the assembly of form IB Rubisco in plants and green algae. The biogenesis of Rubisco within these organisms is the most complex (Figure 1.3). Firstly Rubisco aggregation is prevented through the transient binding of a number of chloroplastic chaperonins which are homologous to the GroES/ GroEL bacterial chaperones. These include Cpn60α and Cpn60ß (Vitlin Gruber et al., 2013) which are homologous to GroES (Bracher et al., 2017), the Cpn10 chaperone which is homologous to GroEL as well as the additional Cpn20 (Tsai et al., 2012, Bracher et al., 2017) (Figure 1.3). Once subunits are folded by the chaperones, assembly factors transiently bind in succession to assemble the quaternary Rubisco complex. These assembly factors include Raf1 (Hauser et al., 2015), RbcX (Kolesinski et al., 2013), Raf2 (Salesse-Smith et al., 2018) and Bds2 (Figure 1.3) (Fracheboud et al., 2004). For form IA Rubisco in cyanobacteria and proteobacteria they have their own assembly factor, acRaf, a small, disordered protein homologous to Raf2 (Wheatley et al., 2014).

There are no known assembly factors in the red type Rubisco, form IC and ID. Within form IC organisms RbcS has been shown to mediate the assembly of the holoenzyme structure (Joshi et al., 2015). The action of the extended C-terminal hairpin loop on RbcS allows heterologous expression in Tobacco without the input of additional chaperones (Gunn et al.,

2020). Form ID Rubisco cannot assemble heterologously, suggesting the presence of currently unknown assembly factors (Oh et al., 2023).

1.1.8 Enzyme-carbon-magnesium^{2+} complex and inhibition

Rubisco requires the formation of the enzyme-carbon-magnesium^{2+} (ECM) complex for activation of enzyme activity. This involves the carbamylation of the ubiquitously conserved Lys201 residue which is subsequently stabilised by binding of the Mg^{2+} cation. Once stabilised RuBP will bind tightly to the enzyme where it can be carboxylated or oxidised (Tommasi, 2021). In the event of RuBP binding prior ECM complex formation, RuBP becomes 'caught in the Rubisco mousetrap' (Andrews, 1996) being unable to progress in the reaction. Additionally, Rubisco is capable of binding other 5-carbon sugar phosphates, namely 2-Carboxy-D-arabitinol 1-phosphate (CA1P), Xylulose 1,5-bisphosphate (XuBP) and D-Glycero-2,3-pentodiulose 1,5-bisphosphate (PDBP) with XuBP being the stereoisomer of RuBP. These sugars also have an inhibitory effect on the enzyme, strongly binding and preventing further catalysis (Orr et al., 2022).

The action of CA1P aids regulation of diurnal patterns being synthesised under low light conditions (Andralojc et al., 2012) inhibiting nighttime Rubisco action. This process of inhibition has been shown to be essential for plant growth by stabilising and maintaining high levels of Rubisco within the plant (Lobo et al., 2019). Under daylight conditions, redox regulation of carboxy-d-arabinitol-1-phosphate phosphatase (CA1Pase) promotes the dephosphorylation CA1P to CA (carboxy-d-arabinitol); preventing further binding with Rubisco.

Alternatively, XuBP and PDBP formation are not linked to diurnal cycles and are a result of Rubisco misfiring in the carboxylation (Pearce, 2006) and oxygenation of RuBP (Harpel et al., 1995) respectively. These inhibitory complexes have to be released from the Rubisco active site by the action of Rubisco activase (Rca) to prevent accumulation of inhibited Rubisco.

1.1.9 Rubisco activase

Rubisco activases are ubiquitous across all Rubisco species associated with CBB cycle; evolving at least three times through convergent evolution and being transferred across lineages by lateral gene transfer (Mueller-Cajar, 2017). The three forms can be categorised as 'green type' associated with plants, green algae and cyanobacteria possessing form IB Rubisco (Salvucci et al., 1985), 'red type'; associated with red algae, proteobacteria possessing form IC Rubisco (Loganathan et al., 2016a) (Figure 1.3) and cyanobacteria possessing form IA Rubisco and finally 'CbbQO type'; associated with proteobacteria possessing form IA Rubisco and form II Rubisco species (Tsai et al., 2015b).

Despite evolving convergently there are commonalities in their mechanistic reactivation of inhibited Rubisco and structure.

The consensus structure of the three Rca forms is a hexameric ring with a central pore which is established across the wider AAA+ ATPase superfamily (Stotz et al., 2011, Mueller-Cajar et al., 2011, Tsai et al., 2015b). The axial pore of the hexameric ring is thought to interact with sites on the Rubisco large subunit where it brings about a conformational change of holoenzyme structure by threading the large subunit through the central pore and releasing the inhibitor. Where and how Rca binds differs from form to form and has been demonstrated to be highly species specific in 'green type' Rcas preventing heterologous activity (Wachter et al., 2013). The mechanism of action in red type and CbbQO type is established acting on the C-terminus of the RbcL units (Mueller-Cajar et al., 2011, Tsai et al., 2022b). CbbQO type has the addition of a linker protein (CbbO) with VWA structure which adjoins the C-terminal of Rubisco to the active sight of the CbbQ complex (Tsai et al., 2020). The activity of red and CBBQO type of Rubisco activase appears to be significantly enhanced by Rubisco in its inactive conformation (Tsai et al., 2015a, Mueller-Cajar et al., 2011). This trait of increase activation of Rubisco activase by inhibited Rubisco is not observed in green type organsisms.

Notably the pore aperture of the hexameric green type Rubisco is larger than other Rca forms. Therefore it is thought that a more significant structural domain of the RbcL N-terminus is implicated in remodelling by activase (Scales et al., 2014) however the exact mechanism remains unknown. Both the red type and green type Rubisco activases share a

heterooligomeric structure of the holoenzyme complex. In red type this involves a plastid encoded and c-terminal extended, nuclear encoded isoform with 1:1 stoichiometry in the hexameric holoenzyme (Loganathan et al., 2016b). Plants have been shown to possess multiple isoforms of Rubisco activase which can be differentially expressed to modulate rate of activity and thermostability of the enzyme (Degen et al., 2021). This presents an engineering opportunity for photosynthetic organisms as Rubisco activase thermostability has been demonstrated to be highly variable, often being the limitation in photosynthesis at higher temperatures (Degen et al., 2021), (Loganathan et al., 2016b).

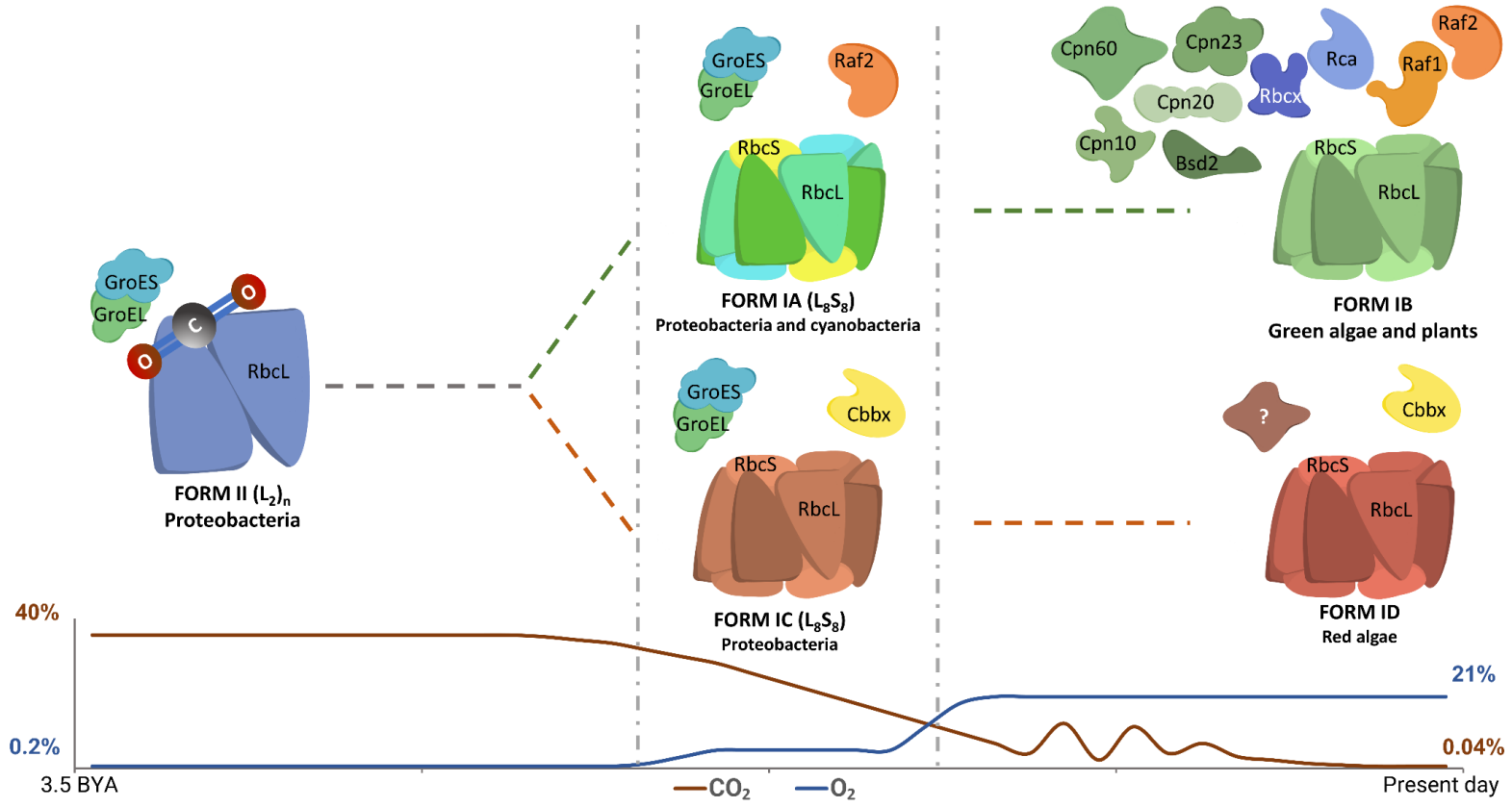


Figure 1.3 - The evolution and biogenesis requirements of each Rubisco form alongside the changing atmospheric condition. GroES and GroEL represent the generic bacterial chaperones. Cbbx denotes the red-type Rubisco activase. Raf2 is the Rubisco assembly factor 2. Raf1 is the Rubisco assembly factor 1. Rca is the green type Rubisco activase. Rbcx is an assembly factor for form IB Rubisco, Bsd2 is the intermediatory Rubisco assembly factor found in form IB organisms and the Cpn proteins are chloroplastic chaperone proteins required for form IB Rubisco assembly.

1.1.10 Positive selection of Rubisco within Form I genes

The evolution of Rubisco is clear to see between Rubisco forms with the increasing complexity of biogenesis and subtle differences in Rubisco subunit structure. However, within Rubisco forms the genes are highly conserved across phylogenies (Kapralov and Filatov, 2007). Despite this conservation there are a number of residues within the *rbcL* gene which are positively selected for (Kapralov and Filatov, 2007). This positive selection has been demonstrated as a mechanism of adaption, either to the environment (Hermida-Carrera et al., 2017) or to changes in internal CO_2, often through the incorporation of carbon concentrating mechanisms (CCMs). In form ID Haptophyta this positive selection is observed with decreasing atmospheric CO_2 concentrations and the secondary endosymbiosis event into Chromista (Young et al., 2012). In terrestrial plants positive selection of Rubisco can be observed within convergent lineages where C4 anatomy has evolved (Kapralov et al., 2012), (Parto and Lartillot, 2018) and within oak, residues under positive selection are consistent with certain climatic characteristics such as aridity or temperature (Hermida-Carrera et al., 2017). Thus it is clear that Rubisco has the capacity to evolve to changes in its environment. It is thought that these subtle changes in the Rubisco sequence space can bring about changes in the Rubisco kinetics better suited to the internal CO_2 environment (Kapralov et al., 2011) or climatic conditions (Hermida-Carrera et al., 2017).

Interestingly positive selection within the *rbcS* gene appears to be minimal with weakened signals relative to the *rbcL* positive selection (Kapralov et al., 2011). Additionally there has been no evidence to indicate residues under positive selection in algae or cyanobacteria (Kapralov and Filatov, 2007, Goudet et al., 2020). This may be a result of the increased diversity in Rubisco of marine environments and the *rbcL* gene relative to the highly conserved *rbcL* of land plants. Historically positive selection on the Rubisco gene has been measured through the application of codon-based substitution models where the ratio of synonymous to non-synonymous is measured across all codons of a gene for all sample species, irrespective of phylogeny (Yang et al., 2005). However this methodology used across previous Rubisco positive selection studies has inherent biases as positive selection is more detectable across more highly conserved, smaller alignments (Murrell et al., 2012). More recent codon models that have not been exploited in Rubisco research have increased

flexibility, allowing detection of positive selection that occurs episodically or pervasively across phylogenies (Murrell et al., 2012).

1.2 Rubisco in the marine environments

1.2.1 The importance of Rubisco within marine systems

The CBB cycle is the dominant form of carbon fixation in the marine environment (Li et al., 2020). It is thought that 50% of carbon annually is fixed through 'blue carbon cycles' and therefore represents a significant proportion of carbon fixed globally (Sabine et al., 2004).

This conversion of inorganic carbon to usable organic carbon by the CBB cycle is essential for two reasons. Firstly this cycle acts as the basis of food webs in marine systems with marine autotrophs representing a rich carbon source for heterotrophic organisms (Fry and Wainright, 1991). Secondly marine systems act as efficient sequesters of anthropogenic carbon, fixing 30% of carbon annually derived from human sources. Coastal seagrass meadows, mangroves and kelp forests are highly productive ecosystems buffering atmospheric carbon levels (Serrano et al., 2021). Additionally much of the carbon fixed in oceanic systems falls as particulate matter to deep sea sediments where it is locked away from atmospheric exchange (Krause-Jensen and Duarte, 2016).

1.2.2 Carbon-capture mechanisms in marine systems

Micronutrient availability can be highly limiting in oceanic systems with iron, zinc and magnesium varying at picomolar scale concentrations (Reinfelder, 2011). Opposingly dissolved inorganic carbon (DIC) is rarely limiting in the environment for marine (Raven and Johnston, 1991). This readily available DIC, coupled with the slow diffusive rates of gaseous CO_2 in water necessitates the evolution of a carbon capture mechanism in marine autotrophs to supply Rubisco with ample CO_2 concentrations.

There is a wide diversity of CCMs observed in marine systems but they can be divided into two discrete categories, biophysical CCMs and biochemical CCMs with the latter being comparable to C4 photosynthesis but lacking the canonical Kranz anatomy found in C4 land plants (Clement et al., 2016).

Biophysical CCMs actively import HCO_3^- from the external aqueous environment to Rubisco via the cytosol through HCO_3^- transporters. Carbonic anhydrases (CAs) then convert HCO_3^- to CO_2 in the presence of Rubisco to provide a carbon concentrate environment for carbon fixation. There is a large diversity of carbonate and bicarbonate transporter as well as CAs across marine organisms with different localisation patterns and affinities. This diversity is captured in a study by who demonstrated a wide away of CAs and solute carriers found across diatom lineages, many being the product of lateral gene transfer (Shen et al., 2017).

1.2.3 The application of Metagenomics in understanding marine systems

Marine environments are intrinsically complex systems that encompass vast 3-dimensional spaces. Efforts in recent years have been made to understand these interactions on a global scale through combining metagenomic and metatranscriptomic information with their corresponding ecological and biochemical context (Sunagawa et al., 2020). Tara Oceans have been at the forefront to this global research, conducting world-wide sampling campaigns (Pesant et al., 2015), but adjacently the uptake of smaller scale studies has expanded our knowledge of marine systems (Cao et al., 2020, Tseng and Tang, 2014).

In the context of Rubisco metagenomics of the ocean has also brought about interesting insights. Chemoautotrophic pathways coupling Rubisco carboxylation through the CBB cycle is the predominant form of carbon fixation in the deep ocean where conditions are anoxic and deplete of light (Acinas et al., 2021, Baltar et al., 2023). This process has ecological importance as carbon fixation at depths plays a role in the wider nutrient cycle, recycling sinking particulate matter into organic forms (Baltar et al., 2023). Additionally, (Pierella Karlusich et al., 2021) demonstrated the widespread prevalence of biophysical CCMs in the earth's oceans through metagenomic and metatranscriptomic analysis.

Rubisco sequences from marine metagenomes have also been exploited for their kinetic diversity (Pins et al., 2023). This study highlights the high variation of Rubisco kinetics in marine autotrophs relative to terrestrial plants and in addition shows Rubisco possessing carboxysomes have higher carboxylation rates on average (Pins et al., 2023).

One poorly understood aspect of Rubisco research is within the regulation of the genes in response to environmental stimuli. Within land plants light and temperature have been shown to be significant drivers of regulation (Zhang et al., 2002), (Ohba et al., 2000), (Cavanagh et al., 2023), (Devos et al., 1998), (Peng et al., 2021). However, little is known about the regulation of Rubisco genes within marine systems. Metagenomics and transcriptomics may help elucidate this fact.

1.2.4 Pyrenoids and Carboxysomes

Through convergent evolution, red algae, green algae, and cyano/proteobacteria have all developed means of concentrating Rubisco into a partially permeable microcompartment (Zhan et al., 2018, Kikutani et al., 2016, Ni et al., 2022). The significance of this microcompartment is that is allows for effective concentration of CO_2 around Rubisco, increasing efficiency of carboxylation.

Despite being found in disparate lineages there are commonalities in the architecture and composition of these Rubisco microcompartments. Firstly pyrenoids and carboxysomes are not homologous entities, consisting of complex matrices of interacting proteins. CA has been shown to be an intrinsic component of both pyrenoids and carboxysomes allowing for the conversion of carbonates to CO_2 in the presence of Rubisco (Adler et al., 2022, Kikutani et al., 2016, Ni et al., 2022). Secondly Rubisco activases are commonly found across pyrenoids to prevent the inhibition of Rubisco (McKay et al., 1991, Matsuda and Kroth, 2014) however it has been demonstrated that within form IA carboxysomes this is not the case (Chen et al., 2022). The form IA Rubisco activase (CbbQO complex) has been shown to be a particularly slow activase but adjacently, form IA Rubisco from *Acidithiobacillus ferrooxidans* is rarely self-inhibited, thus the activase is not an essential requirement (Tsai et al., 2022a). Differences in structure can also be observed between pyrenoids and carboxysomes. Carboxysomes have an external polyhedral shell which encapsulates Rubisco, this consists of a lattice of small shell proteins (Csos1 and Csos4) which act as semi permeable barrier (Sun et al., 2022). Rubisco is tied internally to this shell through the intrinsically disordered Csos2 shell linker protein which binds to the N and C terminal domains of Rubisco (Ni et al., 2023).

Opposingly pyrenoids do not form this shell complex observed in form IA carboxysomes. The pyrenoid more simply consists of a densely packed Rubisco matrix which is linked by small disordered proteins (Epyc1 in green algae (Mackinder et al., 2016) and Pyco1 in red algae (Oh et al., 2023)). These disordered proteins link Rubiscos via binding of the small subunits (Mackinder et al., 2016)(Oh et al., 2023). Within pyrenoids it is the chloroplastic thylakoid membrane that interlace with the pyrenoid structure where the CA is contained (Caspari et al., 2017), (Jenks and Gibbs, 2000). Despite the widespread prevalence of pyrenoid structures in both red and green algae (Goudet et al., 2020), (Oh et al., 2023) there is poor characterisation of the molecular diversity with proteomics being inferred from single species (*Chlamydomonas reinhardtii* for green algae (Mackinder et al., 2016) and *Phaeodactylum tricornutum* for red (Oh et al., 2023)).

1.2.5 Differences between Rubisco environment in terrestrial and marine organisms

The *in vivo* rate of Rubisco is 0.03 µmol CO_2 s^{-1} in land plants and 20-fold higher in marine organisms at 0.6 µmol CO_2 s^{-1} (Bar-On and Milo, 2019). This represents measurements 100-fold lower than *in vitro* measurements at 25^{o}C for land plants but only 7-fold lower for that of marine species (Bar-On and Milo, 2019). This significant discrepancy between *in vivo* marine and terrestrial rates highlights a number of interesting points.

Firstly, temperatures within the marine environment do not experience diurnal fluctuations that are observed on land with the latent heat capacity of water buffering fluctuations. Secondly due to the mixing effect of marine systems there is less spatial segregation of nutrients within the marine system. Finally, CCMs within the marine system are common place providing elevated CO_2 levels to Rubisco.

Rubisco represents a small fraction of total protein in marine phytoplankton with concentrations ranging from <2.5% (Losh et al., 2013) to 20% in some cyanobacteria (Zorz et al., 2015). Bar-On and Milo (2019) estimate the average percentage to be 3% of total protein in marine phytoplankton. This is a stark contrast to plants where Rubisco can make up to 50% of total soluble protein (Feller et al., 2008) with average concentrations estimated to be at 15% of total proteins in terrestrial plants (Bar-On and Milo, 2019). This discrepancy in Rubisco concentrations between marine and terrestrial systems highlights

the effectiveness of CCMs in marine environments. This coupled with the fact that TIC is rarely limiting in marine systems due to equilibration with the atmosphere (DeVries, 2022) means that Rubisco can be supplied with a constant supply of near saturating levels of CO_2 allowing for reduced concentrations in marine systems.

1.2.6 Kinetic trade-off of Rubisco in marine organisms

There is a historical theory in Rubisco research that the kinetic parameters of specificity and rate of reaction are highly constrained in a linear trade-off (Flamholz et al., 2019) (Figure 1.4). This theory is particularly true for terrestrial plants however when we consider other forms of Rubisco, there is more flexibility in this linear relationship than once considered (Flamholz et al., 2019). Additionally, (Bouvier et al., 2021) proposes a strong phylogenetic bias in Rubisco studies which may amplify the appearance of 'kinetic constraints' because of a lack in sequence diversity.

Despite this it is clear there is greater diversity in Rubisco kinetics of marine organisms relative to terrestrial plants. The fastest rate in form I Rubisco was discovered within members of the *Synnechoccus* family, although specificity does not diverge from the kinetic trade-off. Interestingly form ID Rubisco appears to break the kinetic trade-off with specificity reported to be far higher than would be expected based on the rate of Rubisco (Figure 1.4). This offers a significant engineering opportunity for photosynthesis with (Zhu et al., 2004) proposing the possibility of increasing Carbon assimilation by 30%.

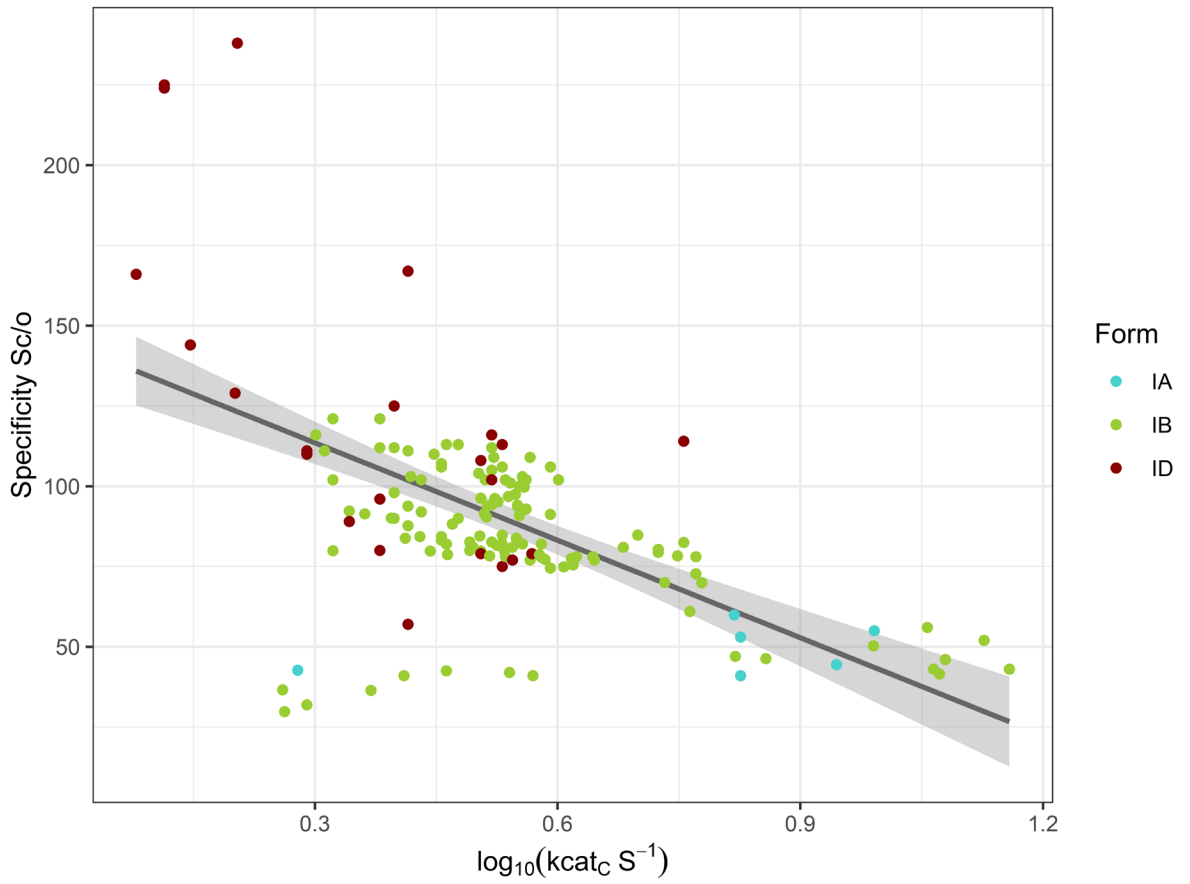


Figure 1.4- The kinetic trade-off between specificity and the $\log_{10}$ transformation of $kcat_c$ s^{-1} for form IA, IB and ID Rubisco. The linear regression represents the 95% confidence intervals.

1.2.7 Opportunities for improving photosynthesis with algal and cyanobacterial architecture

Optimising food production alongside a changing climate poses a significant engineering challenge. As highlighted by Rubisco there is often greater diversity observed in algae and cyanobacteria across all components of photosynthesis, relative to land plants. This presents engineering opportunities to optimise photosynthesis in economically important crop species.

A way to predict the effects of photosynthetic engineering efforts is through modelling. A number of studies have used modelling to predict the effects of engineering algal or

cyanobacterial photosynthetic architecture into plants. This has included efforts to include a full cyanobacterial CCM (Price et al., 2013) and improvements to the range of photosynthetic wavelengths (Chen and Blankenship, 2011). However most strikingly Zhu et al. (2004) modelling efforts demonstrated the effects of the heterogenous expression of the red type Rubisco from *Griffthisia monillis* in crop plants theoretically resulting in a net gain of 30% in carbon assimilation relative for certain crop species (Zhu et al., 2004).

However despite this significant gain in carbon assimilation,(Wu et al., 2023) gives a more balanced overview. It is suggested that a more holistic approach is required to significantly improve crop yields, simultaneously focusing on carbon delivery to Rubisco, supply of ATP through the electron transport chain and efficiency of Rubisco itself. Despite this, it is clear that photosynthetic algae and bacteria offer an under explored opportunity for engineering photosynthetic efficiency. This is a concept that will be explored further in this study.

Contrasting Photosynthetic Gene Expression Between Polar and Tropical Marine Environments

2.1 Introduction

Marine environments are responsible for 50% of the net primary fixation of carbon on Earth with the majority of marine carbon being up taken by phytoplankton (Sabine et al., 2004). The diversity of phytoplankton covers a wide breadth of phyla and environments, and their role is essential; actively taking up inorganic carbon in the form of HCO_3^- and CO_3^{2-} and converting it to organic forms which can be shared across all trophic levels (Barton et al., 2020). The specifics of the phytoplankton community dynamics and carbon uptake are complex and represent a wide range of diversity in carbon capture mechanisms and adjoining photosynthetic machinery (Fisher et al., 2020) (Pierella Karlusich et al., 2021). However, this complexity has begun to be untangled, largely in thanks to the Tara Oceans campaign (2008-2013). This study collated environmental data and genomic data from around the earth's seas and oceans presenting a vital open science research tool for further examination of large-scale community dynamics at the molecular level through metagenomics (Pesant et al., 2015).

Metagenomics is a developing field of research but has been proven to be an incredibly useful tool providing a holistic view of community dynamics or providing new sequence space for enzymatic studies without the need for culturing complex arrays of microbial organisms from environmental samples (Pereira et al., 2018). There are a number of metagenomic studies that have focused primarily on the Tara Oceans campaign data to date giving both big picture views of oceanic ecosystems and specific enzymatic dynamics. By comparing metagenomic gene abundance and transcriptomic abundance Salazar et al. (2019) demonstrated that community turnover (where the ratio of unique genes to expression is higher) in polar regions is more prevalent than that in tropical regions where additionally variance in gene expression levels was higher as highlighted by transcriptomic abundance. This means that polar regions are more sensitive to community changes with climate change (Salazar et al. 2019). Additionally, using Tara Oceans campaign data Cuadrat

et al. (2019) demonstrated the global distribution of antibiotic resistance genes highlighting that the prevalence of such genes was far more abundant in coastal samples than oceanic samples with many contigs from coastal samples containing multiple resistance genes.

Looking more granularly, Tara Oceans datasets have provided insights into sulphur oxidative pathways in deep ocean samples with the prevalence of sulphur oxidase and sulphur reductase being directly correlated with depth in water columns showing sulphur oxidative pathways to be the dominant form of autotrophy in the absence of light (Baltar et al. 2023).

A significant limitation associated with comparing metagenomic and metatranscriptomic data arises from the inability to standardise sampling. Often there is only one biological replicate and previous studies show that metagenomes rarely reach full saturation i.e they are not representative of the full microbial community at the sample site (Pereira et al., 2018). Therefore, efforts need to be made to normalise data, removing sampling variance to allow for the comparison of abundances between genes and sampling sights. There are several methods for read normalisation. Adjusting all read counts by a scaling factor, often based on the most abundant gene count across all samples or by the 50^{th}/ 75^{th} abundance quartiles is generally the most common method of normalisation (Pereira et al., 2018). A second method, used more frequently in meta-taxonomics involves assessing the degree of rarification of samples. A fully rarified sample would be a sample where the number of unique genes is fully saturated meaning that further sequencing would not yield any further gene discovery. As this is unachievable in metagenomics, samples are adjusted to match the minimum rarefication across samples (Pereira et al., 2018). A final method that is gaining prevalence in recent years is the adjustment of read counts to the prevalence of 'house keeping genes' (Milanese et al., 2019). These are genes that are ubiquitous across all phyla, exist as single genomic copies and are constitutively expressed. Normalising read counts to house-keeping genes gives per cell abundance of read counts as a result. Often Metagenomic studies will utilise multiple forms of normalisation methodologies to compile standardisation (Salazar et al. 2019).

Within marine environments there are multiple pathways for carbon fixation but the primary metabolic route is via the CBB cycle (Hügler and Sievert, 2011). For phytoplankton this requires a flux of NADPH from the photosystems to the CBB cycle allowing the central enzyme of Rubisco to fix gaseous CO_2 with RuBP into organic sugars. There are multiple

forms of Rubisco found within the marine environment with carboxylation capacity however very little is known about their abundance or diversity. Being the central enzyme to carbon fixation they are affected by changes in environmental conditions with parameters such as temperature and light intensity being shown to effect Rubisco expression at the community level and within individual species (Young et al., 2015b, Sun et al., 2014).

Of the Earth's marine environments polar environments are the most susceptible to change. With temperatures of the arctic increasing at almost four times the rate of the rest of the world since 1979 (Rantanen et al., 2022). This rate of increase has significant implications on for the coupled dynamics of metabolic rate, environmental dissolved gaseous states, iron availability and salinity. All of which effect community dynamics (Riebesell et al., 1993), (Greene et al., 1991), (Adenan et al., 2013).

Therefore, the aim of this study is to primarily investigate the diversity and abundance of Rubisco forms in the water column and how this relates to photosynthesis as a whole. We will then compare how photosynthetic expression differs between tropical and polar systems. Finally, we will assess what environmental factors are driving expression patterns of Rubisco and photosynthetic genes in the marine systems.

2.2 Methodology

2.2.1 Collection of Genetic material and Environmental data during the Tara Oceans Expeditions (2009-2013)

Methods are described in full in (Pesant et al. 2015) however in brief; Metagenomic, metatranscriptomic and environmental data was collected from 180 sample sites encompassing the Earth's seas and oceans. Environmental data used in this study consists of temperature (°C), oxygen concentrations (µmol/L), total inorganic carbon (TIC) calculated as $\sum CO_2 = [H_2CO_3] + [CO_2] + [HCO_3^-]$ (Edmond 1970), salinity g kg^{-1}, iron concentrations (µmol/L), total NO_2 and NO_3^- concentration (µmol/L) as well as chlorophyll A concentrations (mg/m^3).

Within these sites, sampling was conducted at a range of depths defined as Surface Layer (5-10M), deep chlorophyll maximum (20-200 m) and Mesopelagic (200-1000 m) with environmental and genetic data being collected at each depth.

The subsequent genetic material was size fractioned for microbial enriched samples (0.22um-3um) and sequenced as outlined in (Alberti et al. 2017). In short environmental samples were transported from port to EMBL labs at 6 week intervals, ensuring cold storage throughout. Mechanical cryogenic grinding was used for cell lysation followed by NucleoSpin kit extraction for RNA and Macherey-Nagel DNA elution for DNA extraction. For library preparation DNA / cDNA for RNA was fragmented to ~300bp and size selected on agarose gel after amplification. For sequencing reads illumina ligation primers were attached and pair-end sequencing was conducted on the Illumina HiSeq2500 (Alberti et al. 2017).

Within this study tropical sample sites were defined as sample sites between 23.5^{o} north and south of the equator. Polar sample sites were deemed as sites that were 60^{o} north and south of the equator. For all sample sites used in the comparison of photosynthetic genes, only samples from surface layers (SFL) and deep chlorophyll maximum (DCM) layers were used. Mesopelagic (MES) and mixed samples were discounted.

2.2.2 Collation of the Ocean Microbial Reference Gene Catalogue v2 (OM-RGC_v2) and gene abundance profiles for each sample site

A reference catalogue called the OM-RGC_v2 consisting of microbial genes from marine environments was constructed by Salazar et al. (2019) consisting of over 47 million unique protein encoding genes. These were assigned Kyoto Encyclopedia of Genes and Genomes (KEGG) and Clusters of Orthologous Genes (COG) IDs and used as a reference in this study.

Additionally, Salazar et al. (2019) assembled scaffolds from metagenomic and metatranscriptomic data from Tara Oceans aligning them to the OM-RGC_v2 based on sequence homology >95%. This allowed Salazar et al. (2019) to calculate gene abundance profiles for each sample site, normalising gene count to gene length. These gene abundance

counts are used as the basis of this study. More specific details on cleaning, assembly and software used can be found in (Salazar et al. 2019)

2.2.3 Extracting Rubisco and associated proteins from the OM-RGC_v2

A list was curated of proteins corresponding to genes of interest (GOI), examined in this study. These genes corresponded to Rubisco associated proteins such as chaperones and activases as well as genes associated with carbon capture.

A local BLAST+ database was created for the OM-RGC_v2. tblastn search was used to extract unique identifiers for genes homologous to GOIs within the OM-RGC_v2. The threshold defined as homology was an evalue cutoff of 1e-10, percentage identity of 40% and a percentage cover of 75% between OM-RGC_v2 gene and the GOI.

For each GOI the total copy number was calculated by summing read count of each homologous gene extracted from the OM-RGC_v2 within each sample site.This process was done for both metagenomic and metatranscriptomic read counts.

2.2.4 Predicting *rbcL, rbcS* and *cbbX* forms in OM-RGC-v2 and calculating copy abundance

In addition to calculating read counts of GOI in sample sites, read counts for *rbcL* (encoding the Rubisco large subunit)*, rbcS* (Encoding the Rubisco small subunit) and *cbbX* (Encoding the Rubisco activase found in cyanobacteria, proteobacteria and red algae) genes were calculated by their individual forms. Forms can be determined of these three genes by their phylogenetic origin.

rbcL, rbcS and *cbbX* candidate sequences were extracted from the OM-RGC_v2 using tblastn homology search and bait sequences (threshold, e-value 1e-10, 30% query coverage and 40% percentage similarity to bait protein). The bait sequences used were RbcL of known forms including form IA from Proteobacteria and Cyanobacteria, form IB from Eukaryotes and Cyanobacteria, form IC from Proteobacteria, form ID from Eukaryotes and form II from Prokaryotes. The same principal was used for RbcS and CbbX using RbcS sequences from cyanobacteria (form IA and IB), form IA and IC from Proteobacteria and form IB and ID from

Eukaryotes. The CbbX sequences used were from cyanobacteria (form IA), proteobacteria (form IC) and red algae (form ID). The resulting sequences of the tblastn search were cleaned for duplicates.

Form designation for each of the resulting sequences was achieved by a two-step process. Firstly protein sequences extracted from the OM-RGC_v2 using tblastn were annotated using DIAMOND (v2.0.15, BLOSUM62,NCBI 2021 database), the highest e-value sequence was used. This gave taxonomic context to sequences. This combined with the construction a maximum-likelihood phylogenetic tree using a Dayhoff matrix model was used to confirm form designation of *rbcL* and *rbcS* sequences. *cbbX* sequences do not group discretely into phylogenetic clades based on form therefore the form was solely inferred from the DIAMOND annotation.

To calculate relative form abundance within a sample normalised count data for each sequence from the OM-RGC_v2, corresponding to a specific form was summed.

2.2.5 Normalisation of read counts for metagenomic and metatranscriptomic gene profiles

Phylogenetic marker genes (PMGs) are conserved across taxa, constitutively expressed and are only found as single copies within cells. Therefore each gene read count for the GOI genes, *rbcL*, *rbcS* and *cbbx* forms as well as read counts of all KEGG genes annotated by Salazar et al. 2019 were divided by the median read count of 10 PMG genes within each sample to give 'read count per cell' for both metagenomic and metatranscriptomic counts. The genes used for this normalisation step were K01409, KO1869, KO1873, KO1875, KO1883, KO1887, KO1889, KO3106, KO3110 and KO6942 (Table 2.1), which have been demonstrated as an effective means of normalisation to basal gene levels (Milanese et al.2019) (Salazar et al. 2019).

Table 2.1- Single copy, constitutively expressed KEGG orthologue genes used for gene count normalisation

KEGG orthologue No.	Enzyme	Symbol	Role
K01409	N6-L-threonylcarbamoyladenine synthase	*kae1*	tRNA Modification factor
K01869	leucyl-tRNA synthetase	*manA*	tRNA biosynthesis
K01873	valyl-tRNA synthetase	*valS*	tRNA biosynthesis
K01875	seryl-tRNA synthetase	*serS*	tRNA biosynthesis
K01883	cysteinyl-tRNA synthetase	*cysS*	tRNA biosynthesis
K01887	arginyl-tRNA synthetase	*argS*	tRNA biosynthesis
K01889	phenylalanyl-tRNA synthetase alpha chain	*pheS*	tRNA biosynthesis
K03106	signal recognition particle subunit SRP54	*ffh*	Secretion system
K03110	fused signal recognition particle receptor	*ftsY*	Secretion system
K06942	ribosome-binding ATPase	*ychF*	Ribosome biogenesis

Following this a pseudocount converted the read counts for subsequent transformations. The 'read counts per cell' were transformed as shown below:

$$\frac{read\ count\ per\ cell\ for\ gene\ in\ sample\ site}{maximum\ read\ count\ per\ cell\ for\ gene\ in\ sample} \times 10^9$$

Dividing by the maximum read count within the sample was used to scale the read counts relative to the sample as additional means of correcting sampling biases.

Following this the pseudocount was transformed using a variance stabilising transformation (Love et al., 2014). This $\log_2$ transforms read counts for each gene to have comparable variance across all sample sites irrespective of the mean abundance of the sample site. Thus making the genes approximately homoscedatic with abundances comparable between genes and across sample sites. Finally read counts were calculated relative to the median PMG counts giving $\log_2$ transformed profiles per cell.

2.2.6 Calculating relative expression of genes between polar and tropical sample sites

Relative expression was calculated by dividing the $\log_2$ transformed metatranscriptomic count by the $\log_2$ transformed metagenomic read count. For comparison of genes between tropical and Polar sample sites the mean expression was calculated and difference between

medians was defined as the difference in expression. Significance was assessed with the non-parametric Mann Whitney U-test assuming unequal variance. An additional Holm correction was applied to minimise false-discovery rate.

2.3 Results

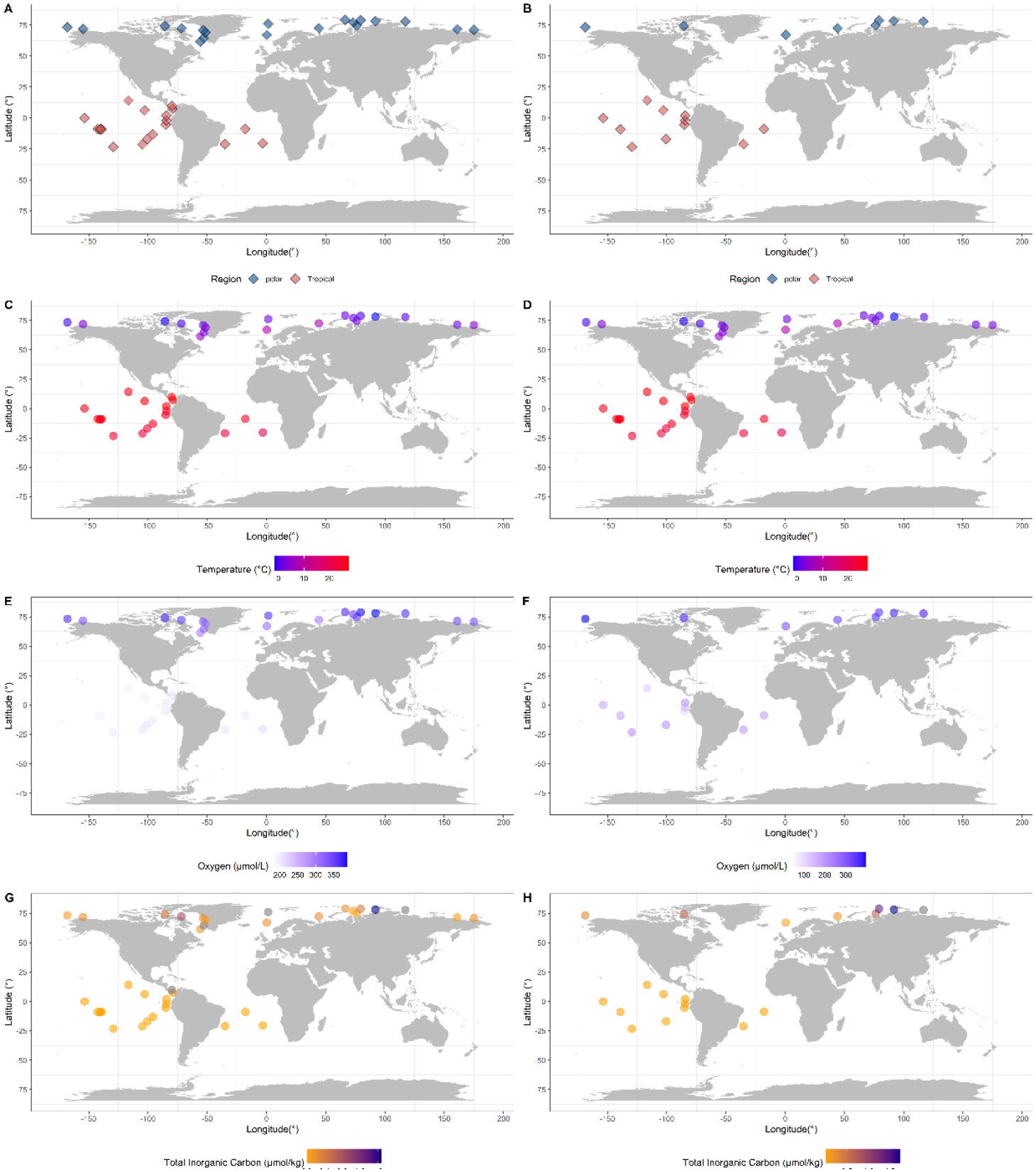
2.3.1 Overview of expression levels between polar and tropical sample sites

During the Tara Oceans campaigns (2008-2013) genomic and the environmental data of temperature (°C), oxygen concentrations (µmol/L), total inorganic carbon (TIC) salinity g kg^{-1}, iron concentrations (µmol/L), total NO_2 and NO_3^- concentration (µmol/L) and chlorophyll A concentrations (mg/m^3) were taken from a of range sample sites across the Earth's seas and oceans. In this study we compare expression levels of photosynthetic genes between polar and tropical samples in surface water (SRF) and the deep chlorophyll maximum (DCM) water layer (Figure 2.1). In total there were 19 SRF sample sites and 11 DCM sites categorised as tropical sites. In addition, there were 19 polar SRF sample sites and 8 DCM polar sample sites (Figure 2.1). Within the polar sample sites the water temperature did not exceed 8.47°C with an average of 1.55°C. For tropical sites, water samples were not below 17.29°C (Figure 2.3), with an average temperature of 24.11°C, this represented a significant difference for both SRF and DCM sample sites. Oxygen levels are tightly correlated with temperature levels (Figure 2.4) presenting as a significant divide between tropical and polar in both SRF and DCM sample sites. Average oxygen levels in polar waters were 346.70 µmol/kg and tropical waters being 184.02 µmol/kg (Figure 2.3). Additionally, TIC levels and salinity represent an environmental divide between polar and tropical sites with 0.43 µmol/kg TIC on average in polar sites and 32.13g/kg of salinity, in tropical sites there is 0.0089µmol/kg of TIC in tropical sites and 35.16g/kg of salinity. Once again this was a significant difference in SRF and DCM sample sites (Figure 2.3).

There is more localised variation for nitrate and nitrite levels in tropical sites however this did not represent a significant difference between polar and tropical sites with Nitrogen concentrations being 2.08 µmol/L and 4.24 µmol/L respectively (Figure 2.1, Figure 2.3). Iron levels are distinct between polar and tropical sample sites with an average of 0.00087 µmol/L in polar sites and 0.00022 µmol/L in tropical sites. This represented a significant

difference between SRF and DCM sample sites (Figure 2.1, Figure 2.3). Significance was assessed using a Mann Whitney U-test assuming unequal variance.

Within Polar sample sites environmental variation was shown to be far greater with iron salinity and TIC varying to a greater extent in both SRF and DCM Polar sample sites compared to that of Tropical sample sites (Figure 2.3). This environmental variation is reflected in the increased variance of Polar sample sites observed when analysed by a PCA . However despite this high level of variance within Polar sample sites both Tropical and Polar sample sites exhibit discrete environmental conditions overall (Figure 2.3).


A
B
C
D
E
F
G
H
Latitude (°)
Longitude(°)
Region
polar
Tropical
Temperature (°C)
Oxygen (µmol/L)
Total Inorganic Carbon (µmol/kg)

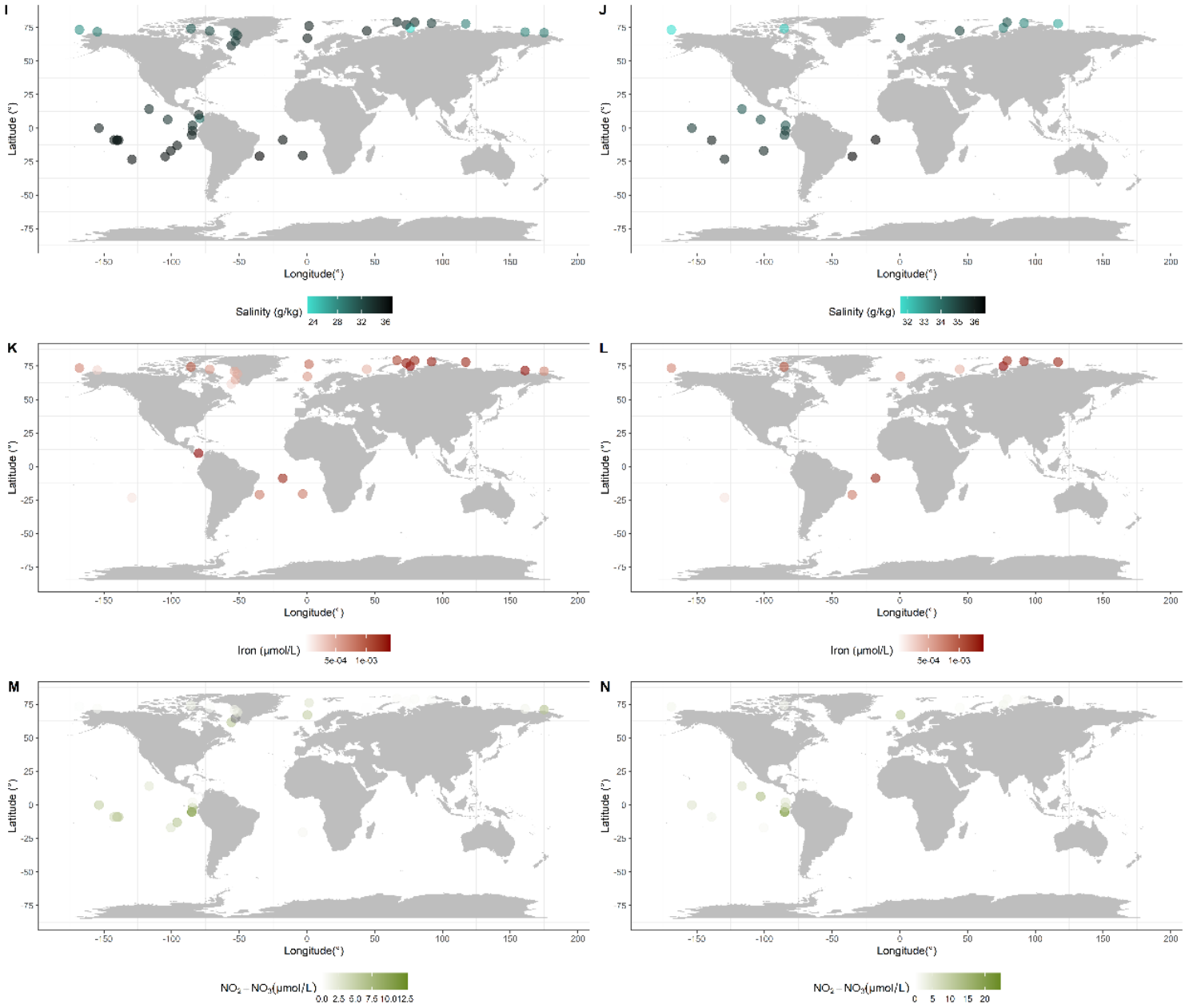


Figure 2.1- **(A)** Shows the geographic location of SRF samples taken **(B)** Shows the geographic location of the DCM water samples. The light blue makers indicate the geolocation of the Polar sample sites (SRF n=19, DCM n=8). The orange diamonds indicate the geolocation of the tropical sample sites (SRF n=19, DCM=11). Tropical sites were defined as sites less than 23.5° north and south of the equator. For each sample 100L of water was taken from the water column (Pesant et al. 2015) and the genomic data was subsequently sequenced and quantified (Salazar et el. 2019). **(C)** Indicates the temperature (°C) of the SRF sites. **(D)** Indicates the temperature (°C) of the DCM sites. **(E)** Represents the oxygen

concentration within SRF sites (µmol/L). **(F)** Represents the oxygen concentration within DCM sites (µmol/L) **(G)** Represents the total inorganic carbon (TIC) in the SRF samples calculated as $\sum CO_2 = [H_2CO_3] + [CO_2] + [HCO_3^-]$ (Edmond 1970) **(H)** Represents the total inorganic carbon (TIC) in DCM samples. **(I)** Represents the salinity of the SRF sites measured as g kg^{-1} of water **(J)** Represents the salinity of the DCM sites (g kg^{-1}) **(K)** Represents the iron concentrations in SRF sites (µmol/L). **(L)** Represents the iron concentrations in DCM sites (µmol/L). **(M)** Represents the total NO_2 and NO_3^- concentration (µmol/L) in SRF samples. **(N)** Represents the total NO_2 and NO_3^- concentration (µmol/L) in DCM samples. The environmental context to the water samples was all collated by (Pesant et al. 2015).

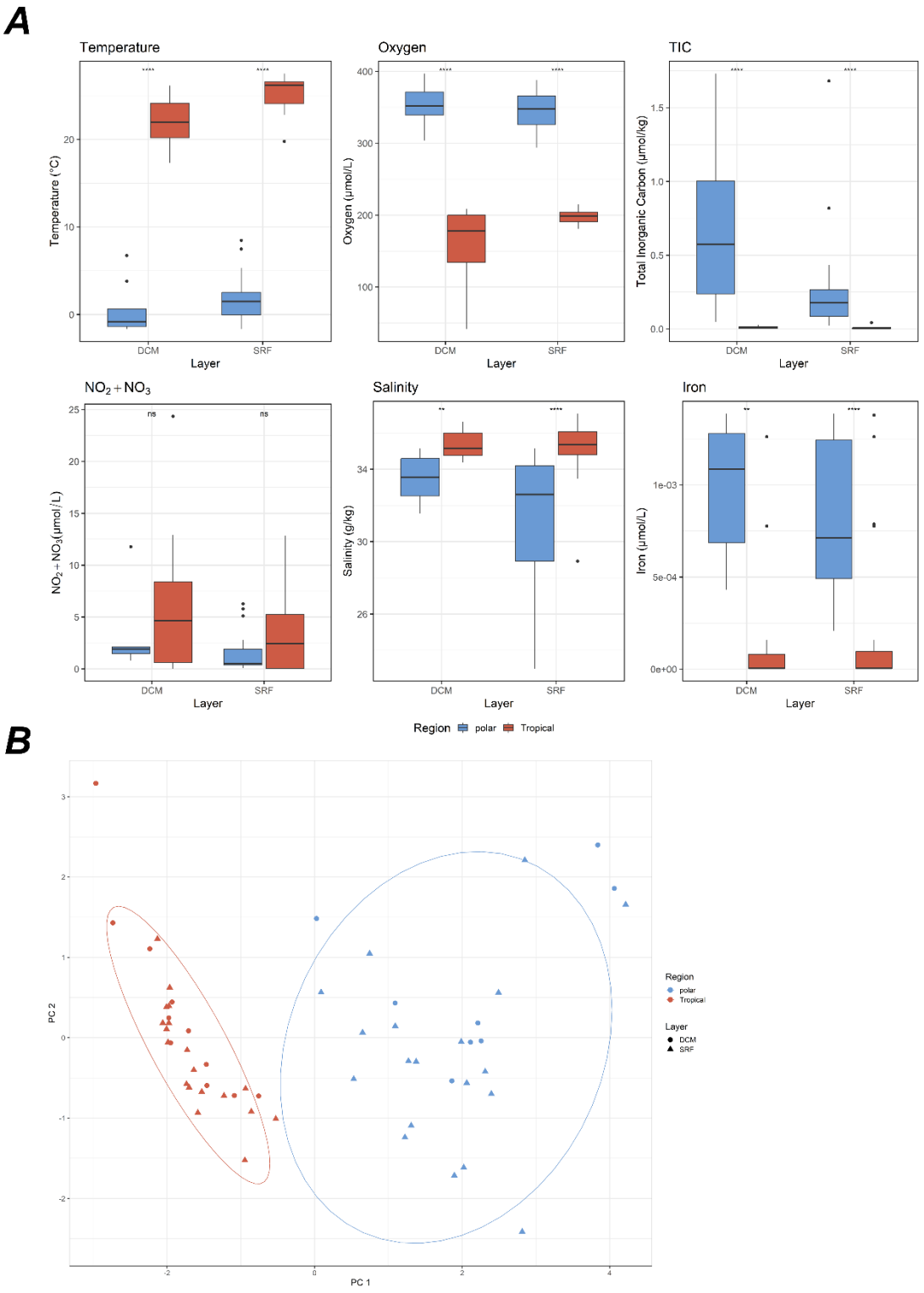


Figure 2.3- A) Shows the range of environmental parameters for temperature, oxygen, TIC, Salinity, iron and nitrate and nitrite levels for polar and tropical water samples for both SRF and DCM sample sites. B) Is a principal component analysis of the environmental parameters of the polar and tropical sample site, the point shape represents the layer at

which the sample was taken from. Ellipses represent multivariate t-distribution of samples. Stars above boxes represents significance (* = p<0.05, ** = p<0.01, *** = p<0.001).

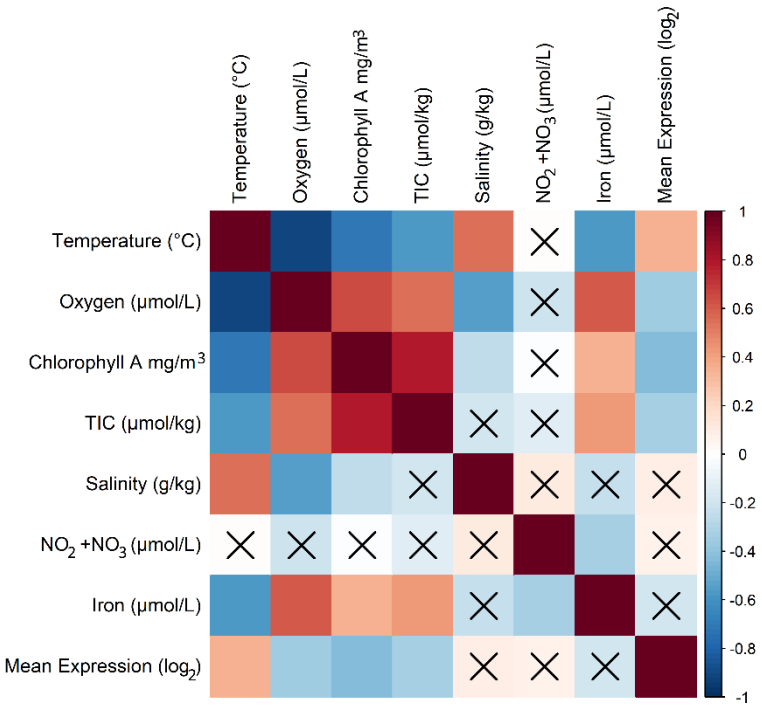


Figure 2.4 – A correlation matrix for the environmental parameters of all SRF and DCM samples sites including the additional subtropical and temperate water sites (total n=89). In addition to the mean expression (log_2) of the 8912 KEGG genes within each sample site. Correlation was assessed using Pearson's correlation coefficient with the depth of colour of the above matrix representing the strength of correlation. 'X' represent non-significant correlations (p>0.05). The environmental context was collected by Pesant et al. 2015

The relative expression of 8912 KEGG genes annotated by (Salazar et al. 2019) within the polar and tropical sample sites were calculated (methods 2.1.7). The relative expression of each of these genes was used to compare metabolic profiles within polar and tropical microbial communities. PCA analysis shows that expression profiles of KEGG orthologue genes within Polar and tropical samples are largely discrete however there is a degree of overlap at certain sample sites with certain tropical samples sites overlapping with polar sites (Figure 2.5).

Comparing the differential expression of KEGG orthologues across all polar and tropical sample sites highlighted that of the 8912 genes examined, 4428 had significantly higher expression levels in tropical environments, 1076 were significantly expressed to greater levels in polar environments and 3408 of the genes showed no significant difference in expression between the polar and tropical sites (Figure 2.5). The difference in expression for each KEGG orthologue was calculated by subtracting median log_2 expression levels in polar sample sites from the median tropical expression level (log_2), significance was calculated by a non-parametric Mann-Whitney-U on log_2 with a p-value <0.05 deemed to be significant (Figure 2.5). The gene with the greatest expression in tropical systems relative to polar systems was L-ectoine synthase (K06720)**.** The gene with the greatest expression in polar sites was that encoding 3-hexulose-6-phosphate synthase (K08903) (Figure 2.5).

Mean expression levels of all KEGG ortholog genes were compared between polar and tropical sites within both SRF and DCM layers. For tropical sites the median expression levels were found to be significantly higher than that of polar communities in SRF samples. 1.06 log_2 expression and 0.58 log_2 expression ($p=0.013$) respectively (Figure 2.5). However there was no significant difference between polar and tropical DCM samples with median expression levels being 0.59 log_2 and 0.76 log_2 respectively ($P>0.05$) (Figure 2.5).

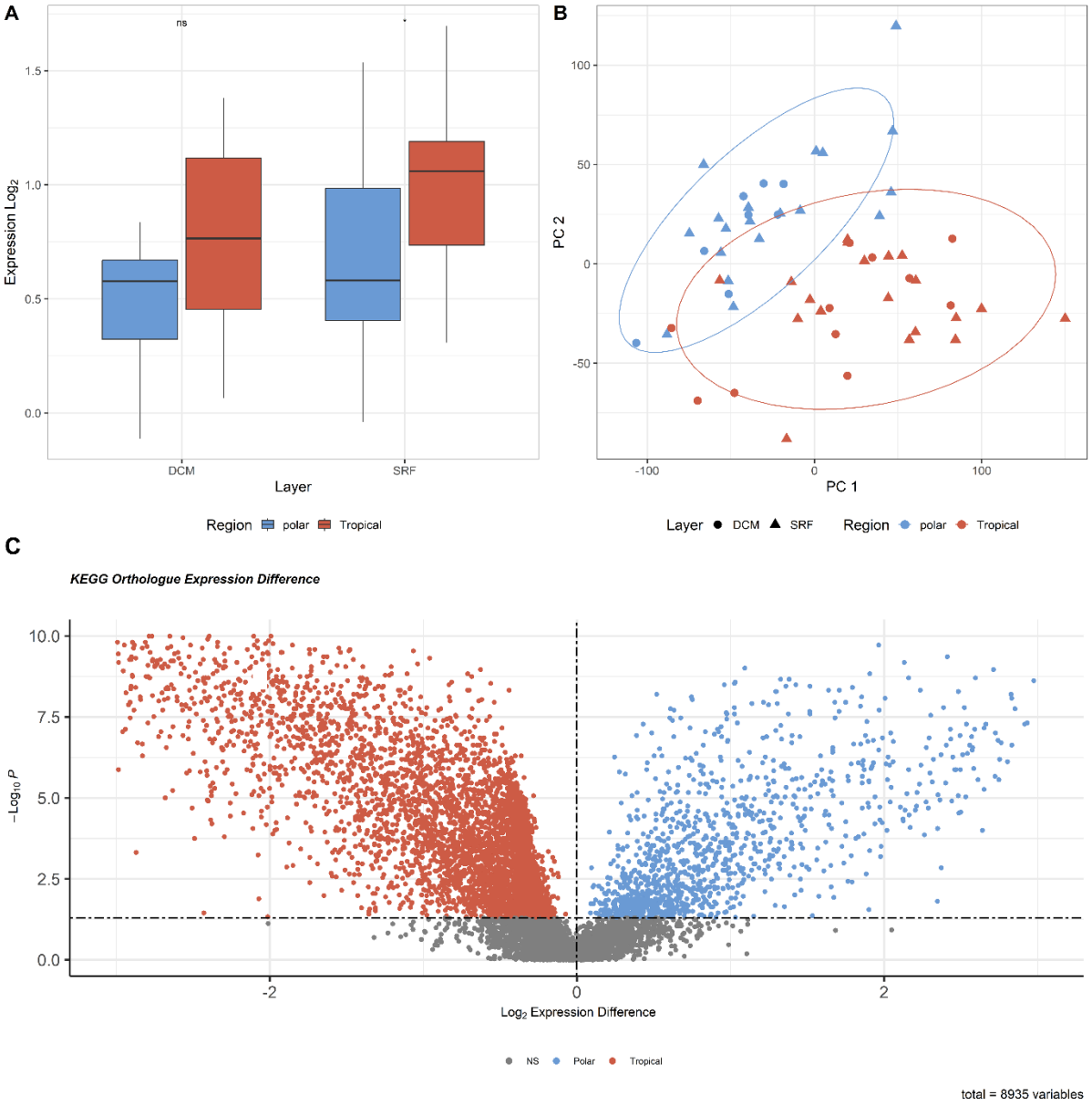


Figure 2.5 – **(A)** The mean log2 expression of all KEGG genes for polar and tropical sample sites at SRF and DCM water layers. Significant difference assessed by Mann-Whitney-U and Tropical sites Stars above boxes represents significance (* = p<0.05, ** = p<0.01, *** = p<0.001). **(B)** Principal component analysis of the relative $\log_2$ expression levels of all 8912 KEGG genes in each sample site. Each point represents a different sample site and the shape represents the water layer. **(C)** Expression difference between polar and tropical SRF and DCM sites was calculated for each KEGG orthologue annotated by (Salazar et al. 2019). Difference was calculated as $median\ polar\ expression - median\ tropical\ expression$ Significance was assessed by a Mann Whitney-U non-parametric test.

The environmental parameters and mean expression levels across all SRF and DCM sample sites were assessed for correlation. This included the addition of sub-tropical and temperate sample sites as to avoid assessing correlation on data that was clustered by polar and tropical sites (Figure 2.4).

Mean expression levels (log_2) within individual samples sites was significantly positively correlated with temperature and negatively correlated with oxygen, Chlorophyll A and TIC. The environmental parameters of temperature, salinity, oxygen, TIC, iron and chlorophyll A concentration were all ecologically linked with only correlation between TIC and salinity not being significantly correlated as well as salinity and Iron concentrations. NO_2 and NO_3^- were only significantly correlated with iron levels (Figure 2.4).

2.3.2 Expression of photosynthetic genes

For a comparison of the light dependent phase of photosynthesis a selection of genes were chosen to represent the light harvesting complex (*lhca1*, *lhca2),* photosystems I/II (*psbL*, *psbA*, *psaA*, *psaB*), the cytochrome b6/f complex (*petB*) and the photosynthetic electron transport system (*petF*). Differential expression was compared between polar and tropial sample sites for both SRF and DCM samples. Polar sites showed a general trend of higher expression levels relative to tropical sites for the light harvesting complex genes and photosystems in SRF. *psbL* (p=0.004), *psbA* (p=0.02), (*psaA* p<0.001), *psaB* (p<0.001), *lhc*2 (p<0.001) were all shown to have significantly higher expression levels in polar sites. *lhcaI*, the representative gene for the cytochrome c6-complex, *petB* and the photosynthetic electron transport chain containing *petF* showed no significant difference between polar and tropical communities (p>0.05) (Figure 2.6).

The CBB cycle associated genes were largely expressed to a higher extent in tropical waters (Figure 2.6). *rbcL, sbp and rbcS* were the genes that had marginally higher expression levels in polar SRF sites and this was found not to be significant (p>0.05). *pgk* (p=0.01), *gapdh* (p<0.001), *fbp* (p=0.004), *tka* (p=0.002) , *rpi* (p<0.001) were all found to be have significantly higher expression in tropical sample sites relative to polar sample sites (Figure 2.6). Median expression levels of *fba* and *prk* were also found to be higher in tropical waters although this difference was not significant. The CBB enzymes with the highest expression in tropical

relative to polar environments was *fbp* at 0.88 log2 difference(Figure 2.6). By far the highest median expression levels across all SRF sites were found for the genes encoding *rbcL* and *rbcS* subunits as well as the enzyme *sbp* (Figure 2.6).

The photorespiratory cycle displays a similar relationship between tropical and polar systems as the CBB cycle. There is a general trend of higher expression levels in tropical waters relative to polar waters with *hpr1* ($p<0.001$), *agxt* ($p<0.001$) and *glyA* ($p<0.001$) all having significantly higher expression in tropical waters. Conversly *glyk* had significantly higher expression levels in polar waters ($p=0.02$) (Figure 2.6).

Within in DCM samples there were differing expression patterns although photosystem expression did not largely differ from that of SRF samples with a general trend of significantly higher expression in polar samples. However expression of *lhca1* was not significantly difference whereas *petF* was found to have significantly higher expression in DCM polar samples ($p=0.03$). Unlike SRF water samples there was not an overarching trend of higher expression of CBB genes in tropical samples. Firstly when all Rubisco genes were considered there was a significantly higher expression with polar DCM samples relative to tropical DCM samples ($p=0.03$) and ($p<0.001$) for *rbcL* and *rbcS* respectively. There was no significance difference between polar and tropical DCM sites for genes encoding for *pgk, gapdh* and *tka* unlike within SRF samples. Additionally expression in polar DCM sites for the gene *sbp* was found to be higher although this difference was not significant. Like SRF samples expression of *fbp* and *rpiA* was found to be significantly higher in tropical sites ($p<0.001$ and $p=0.004$ respectively) (Figure 2.6).

A similar expression profile of the photorespiratory system can found between SRF and DCM samples. Like in SRF samples, genes encoding for *hpr1* ($p=0.03$), *agxt* ($p<0.001$) and *glyA* ($p<0.001$) have higher expression levels in tropical DCM samples relative to polar DCM samples. However in DCM there is no significant difference between glyk expression ($P>0.05$) however *pgp* has higher expression in polar DCM samples ($p=0.02$) (Figure 2.6).

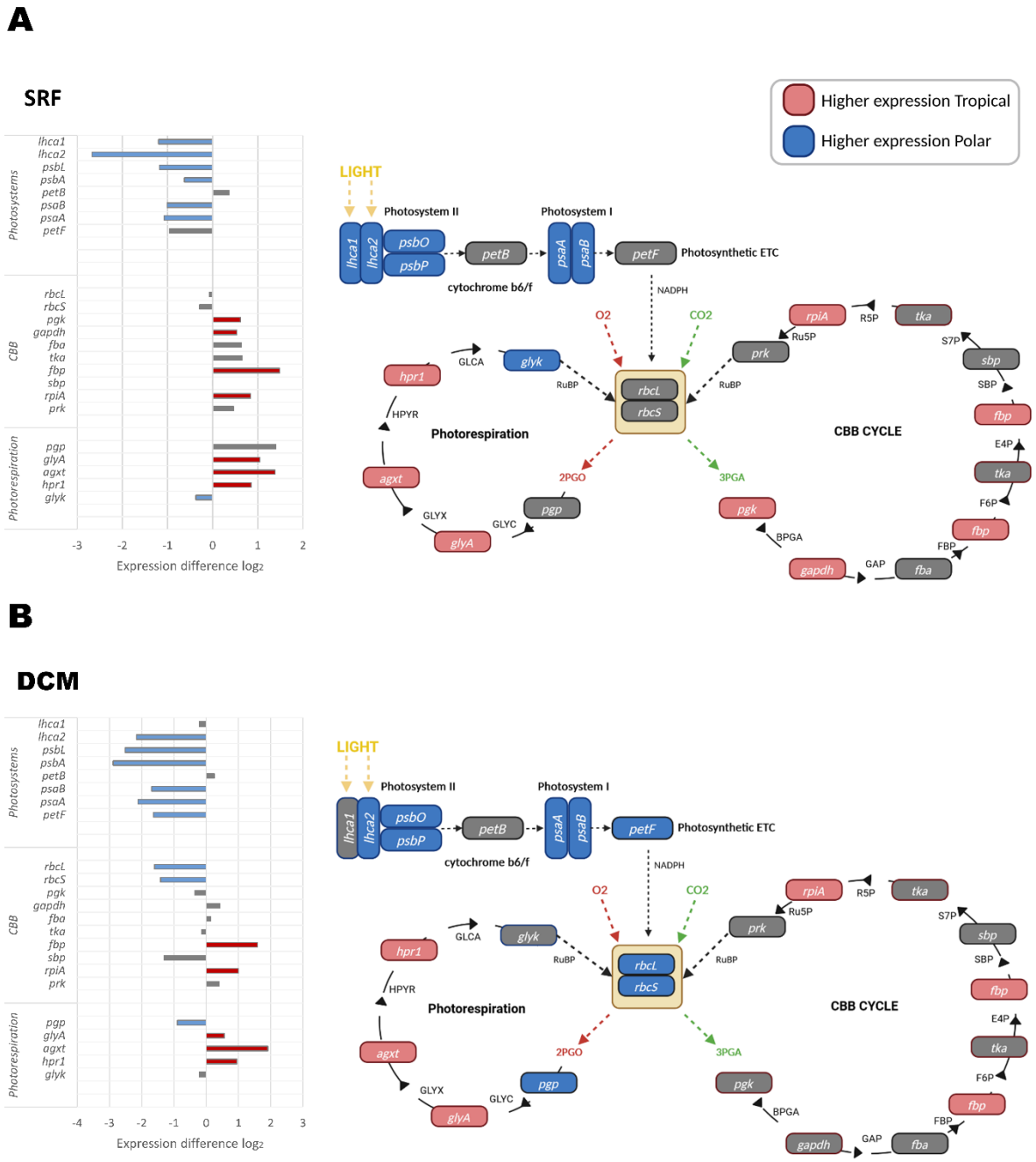


Figure 2.6- A comparative analysis of the relative expression of genes associated with the CBB cycle, Photorespiration and Photosystems. **(A)** Shows the relative expression difference of these photosynthetic genes between polar and tropical SRF sample sites **(B)** Shows the relative expression difference between DCM polar and tropical DCM samples. Difference was calculated as $median\ polar\ expression - median\ tropical$ with significance calculated by a Mann Whitney-U non-parametric test. Blue bars represent a significantly higher expression in polar sites relative to tropical, red bars represent a significantly higher

expression in tropical sites relative to polar. Bars coloured in grey had no significant difference ($p>0.05$).

2.3.3 Comparison of Rubisco form abundance and expression

Rubisco sequences across the marine environments found in the OM_RGC_v2 (assembled by Salazar et al.2019) were divided into the prospective forms. This was done using a combinatorial approach of phylogenetic grouping and diamond annotation based on NCBI-2021 sequence homology. The resulting phylogeny showed all five *rbcL* forms with carboxylation activity were found in this dataset as well as the four Rubisco forms that require an *rbcS*. A maximum likelihood phylogenetic model utilising a Dayhoff substitution matrix demonstrated phylogenetic grouping of forms (Figure 2.7). The lowest number of unique sequences were found within form IB organisms. This may be a result of two factors, initially sequences were grouped by 95% sequence homology within this dataset therefore a lower Rubisco sequence diversity within IB would result in fewer sequences. Secondly size fractionation of water samples would promote for prokaryotic sequences, limiting the number of found form IB sequences, however one would expect a similar scenario for form ID eukaryotic organisms if this was solely the case.

For *rbcL* the green/red divide can be observed in the phylogeny with form II sequences sharing common ancestors sharing a common ancestor with both 'green' and 'red' phylogenetic lineages. This green/red type divide can also be visualised in the *rbcs* phylogeny (Figure 2.7)

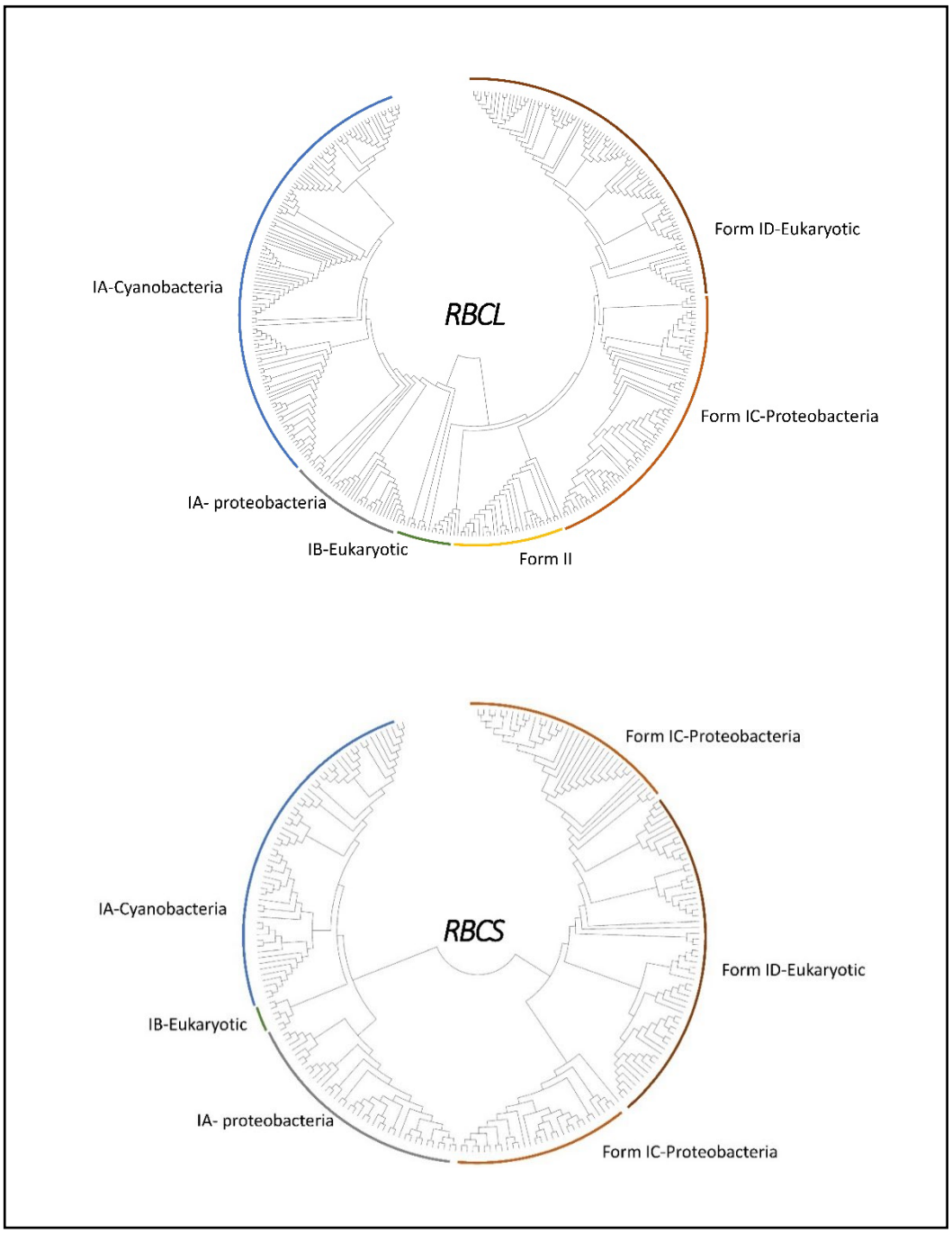


Figure 2.7- Following a tblastn search of the OM-RGC constructed by (Salazar et al. 2019) for Rubisco large and small subunits the resulting sequences were assembled into phylogenetic trees based on their protein sequences. A maximum-likelihood phylogenetic method was used, utilising a Dayhoff substitution matrix. Rubisco forms were annotated using a combinatorial approach of visualising grouping by form and through Diamond+ sequence annotation with the NCBI 2021 database to predict taxonomy. The upper tree represents

the resulting annotation of RbcL sequences. The lower represents the successful annotation of RbcS sequences.

2.3.4 Rubisco form designation and validation

Within form IA, IC and ID Rubisco genes encoding large and small Rubisco subunits are located within the same operon therefore similar expression levels of the two genes is expected. This principle was used to validate Rubisco form designation in (Figure 2.7) expecting a very strong correlation between *rbcL* and *rbcS* expression.

When all SRF and DCM water samples were considered including those in temperate waters we found a strong correlation between expression of Rubisco large and small subunit in the Rubisco forms: IA,IC and ID (Figure 2.8).

Within form IA Rubisco including both Cyanobacterial and Proteobacterial organisms we found a very strong positive correlation between that of Rubisco large and small subunits (r^2=0.972, p<0.001). For form IC correlation analysis there were three samples that presented as distinct outliers for large and small subunit expression out of a total of 89 sample. These three samples were removed from further correlation analysis. Once cleaned form IC Rubisco expression had a strong positive correlation between large and small Rubisco subunits (r=0.84[2], p<0.001). Form ID also had a very strong correlation between that of the small subunits and that of the large subunits relative expression values (r^2=0.90, p<0.001) (Figure 2.8).

However of the four forms assessed, IB Rubisco was found to have the weakest correlation between small and large subunits (r^2=0.38, p=0.000) (Figure 2.8).

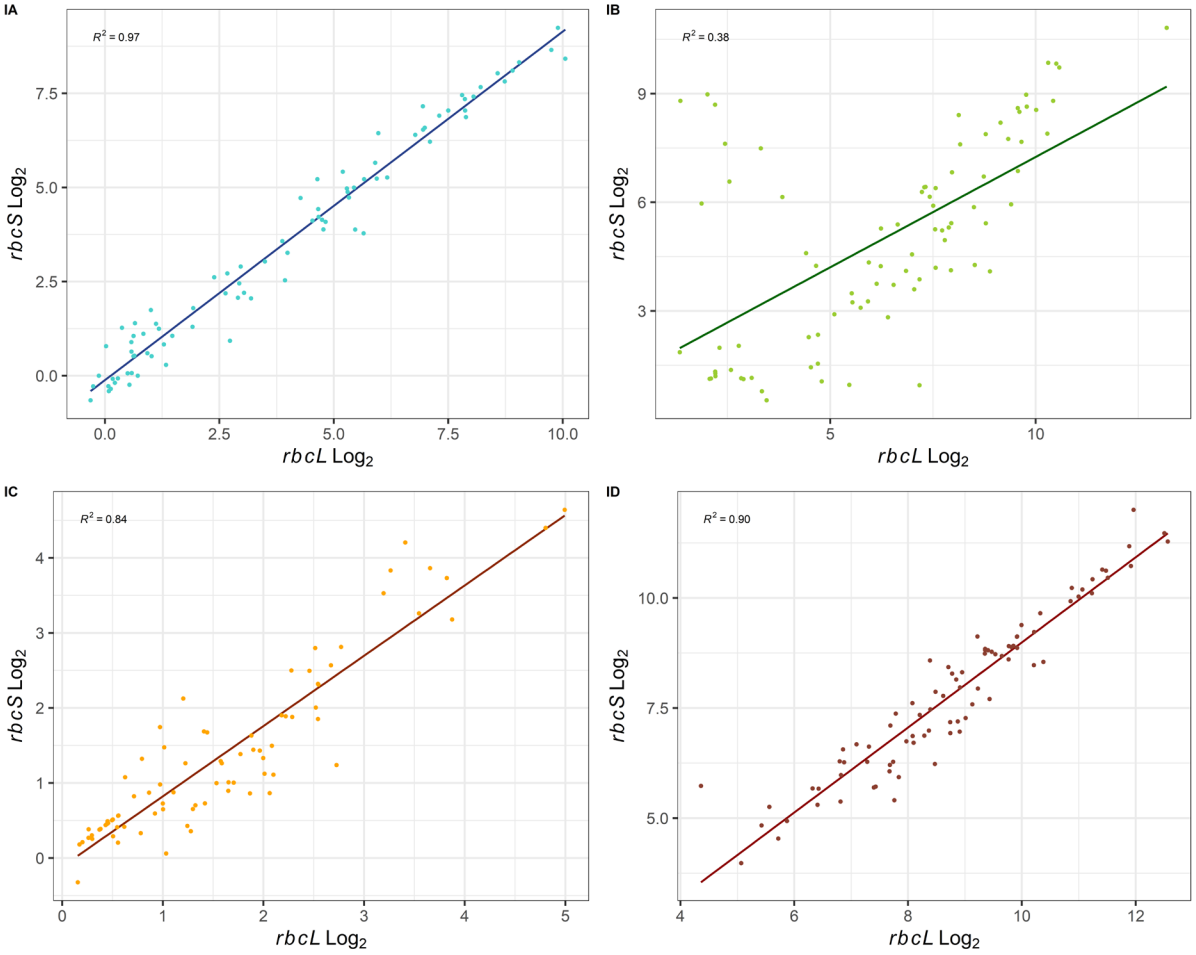


Figure 2.8- The correlation between expression of *rbcL* and *rbcS* within SRF and DCM sample sites (n=89). Each box represents the expression of the individual Rubisco forms. Light blue is the expression of form IA Rubisco derived from cyanobacteria and proteobacteria. Light green represents expression from eukaryotic form IB Rubisco, orange represents form IC expression from proteobacteria and orange represents rubisco subunit expression for form ID Rubisco. For form IC three sample sites were removed as these represented clear visual outliers. The regression line was used to assess linear relationship between *rbcL* and *rbcS* expression with r^2 value shown.

2.3.5 Comparison of *rbcL* and *rbcS* expression between polar and tropical samples

When expression of Rubisco genes are compared for each form between polar and tropical sample sites we see a relatively nuanced response. Firstly form IA *rbcL* associated with cyanobacteria and certain proteobacteria was found to be more highly expressed in tropical SRF samples compared to that of polar SRF samples with a median expression level 5.1 $\log_2$ higher in tropical sites ($p<0.001$) (Figure 2.9). This was the greatest difference in expression levels found in any form. When the relative abundance of form IA *rbcL* within metagenomes of tropical and polar SRF sites was considered, there is an obvious difference between the two environments. Form IA organisms, predominantly in the form of cyanobacteria dominate tropical environments making up 0.15 copies per cell of *rbcL* after normalisation. Conversely form IA *rbcL* only represents $2.7x10^{-4}$ copies per cell in polar SRF sites of which most are derived from proteobacterial origin (Figure 2.9) (Figure 2.10). Opposingly when DCM samples are compared between polar and tropical sites there is no significant difference in form IA *rbcL* expression ($p>0.05$). This is due to the large variation of form IA *rbcL* expression in tropical DCM sites despite form IA *rbcL* being the dominant Rubisco form in tropical DCM sites at 0.13 copies per cell.

tropical DCM sites at 0.13 copies per cell.

Form IB *rbcL* was the only form that had significantly higher expression in polar SRF sites relative to tropical SRF sites ($p=0.017$) (Figure 2.9). However there was no significant difference between expression in polar and tropical DCM samples ($p>0.05$) (Figure 2.9). Form IB *rbcL* gene copies were found to make up $9.7x10^{-3}$ copies per cell in polar SRF samples. Within tropical SRF samples the abundance of form IB *rbcL* was lower only representing $3.4x10^{-3}$ copies per cell. Within DCM polar and tropical samples form IB *rbcL* abundance was very similar at $2.2x10^{-3}$ and $2.4x10^{-3}$ copies per cell respectively (Figure 2.10).

Form IC and Form ID *rbcL* was significantly upregulated in tropical waters ($p<0.001$ and $p=0.044$ respectively) (Figure 2.9). Relative Abundance of gene copy numbers for form IC derived from proteobacteria represented less than $9.9x10^{5}$ of *rbcL* copies per cell in polar waters but represented 0.01 *rbcL* copies per cell in tropical waters. Form ID is by far the most dominant Rubisco form in Polar waters representing 0.05 *rbcL* copies per cell in polar

SRF waters. Looking more closely at form ID phyla. Sequences of the Haptophyta phylum dominate polar waters representing the most abundant form ID form. Form ID *rbcL* also represented 0.03 copies per cell in tropical SRF water sites however in tropical waters Crytophyta was the dominant phyla found here (Figure 2.10). Within DCM samples *rbcL* expression varied much more greatly for form IC and ID Rubisco in tropical sites. This represented a significant difference in expression for form IC *rbcL* with expression being significantly higher in tropical sites ($p=0.015$ (Figure 2.9). For form ID there was no significant difference in expression of *rbcL* genes between tropical and polar DCM samples. Form IC *rbcL* was found to be abundant at less than $9.9x10^{5}$ copies per cell in polar DCM samples but at $7.1x10^{-3}$ copies per cell in tropical DCM samples. Form ID Rubisco was abundant at 0.03 copies per cell in polar DCM samples and 0.01 copies per cell in tropical DCM samples (Figure 2.10).

Form II Rubisco derived from prokaryotes was found to have no significant difference in expression levels between polar SRF and tropical SRF waters and copy numbers represented less than $9.9x10^{5}$ of copies per cell in both polar SRF and tropical SRF sample sites. Form II organisms also had the lowest per cell expression levels of *rbcL* across all Rubisco forms in this study (Figure 2.9) (Figure 2.10). However in tropical DCM samples form II *rbcL* was more abundant at $6.2x10^{-3}$ copies per cell although this did not represent a significant difference in expression between polar and tropical DCM samples.

Differential expression of *rbcS* genes between polar and tropical sites followed a similar trend to that of *rbcl* expression in polar and tropical sites. This is expected due to expression of *rbcL* and *rbcS* being closely correlated (Figure 2.8). For form IA rbcS expression was significantly different between polar and tropical SRF samples ($p<0.001$) but opposingly to *rbcL*, was also significantly different between polar and tropical DCM samples, despite having similarly large variation of expression within tropical DCM samples ($p=0.043$) (Figure 2.11).

Form IB *rbcS* expression differed from that of *rbcL* expression. There was no significance difference in expression between polar and tropical DCM samples, however expression was significantly higher for form IB *rbcS* in tropical SRF samples relative to polar samples SRF ($p=0.047$) (Figure 2.11).

Form IC and form ID *rbcS* expression followed the trends observed in *rbcL* expression exactly. Within form IC *rbcS* expression was significantly higher in tropical SRF and DCM sites relative to polar SRF and DCM sites ($p<0.001$) and $p=0.03$ respectively) (Figure 2.11). For form ID there was no significant difference between *rbcS* expression in DCM sites ($p>0.05$). However, expression was significantly higher in tropical SRF samples relative to polar SRF samples ($p=0.001$) (Figure 2.11).

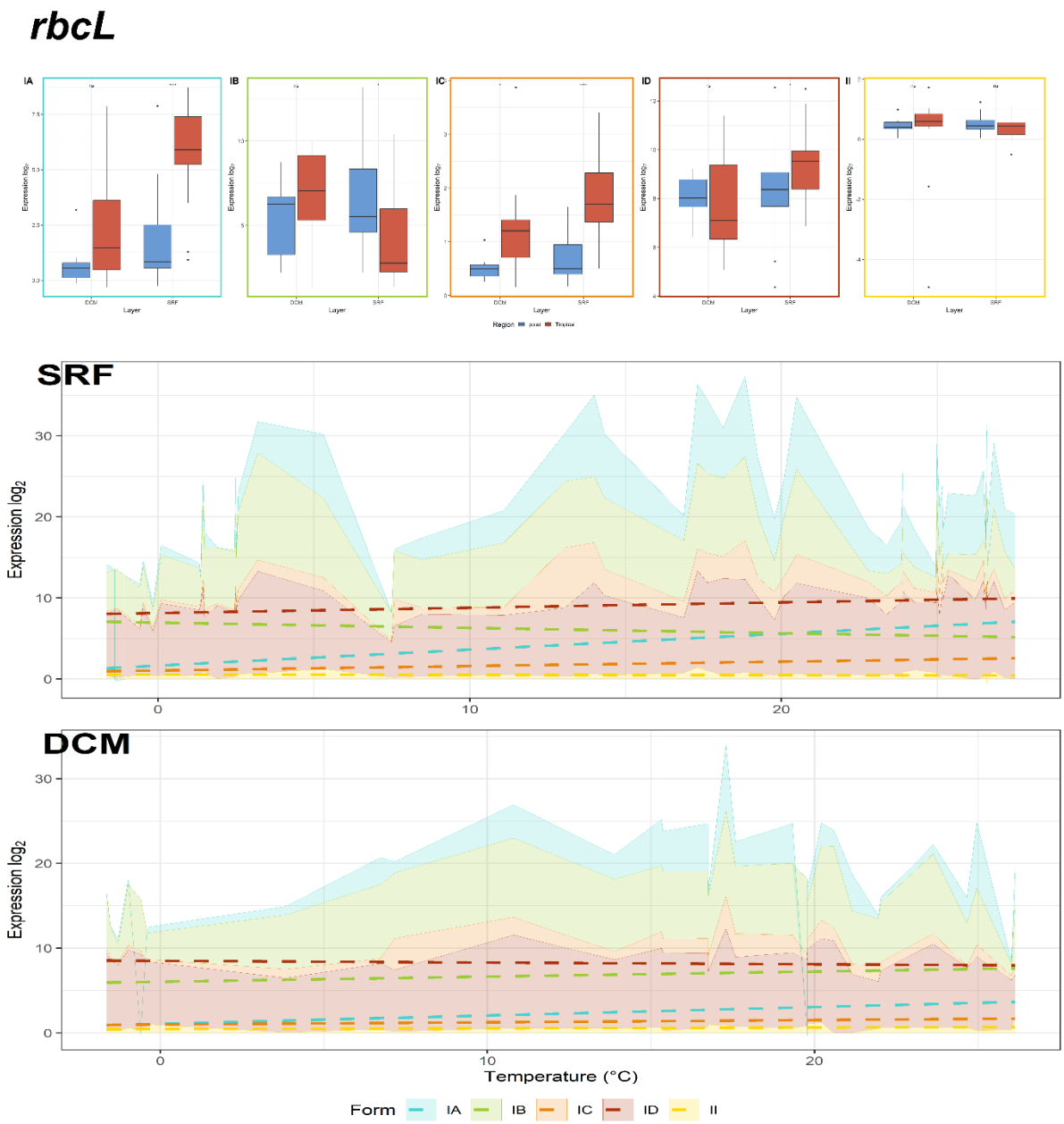


Figure 2.9- The relative expression difference between polar and tropical systems of the *rbcL* forms considered in this study. For this log_2 normalised gene expression values were calculated across all SRF and DCM polar sample sites and tropical surface water sample sites difference was measured by a comparison of Medians through a non-parametric Mann-

Whitney U test for polar and tropical waters. Additionally all SRF sample sites and DCM sample sites were ordered in increasing temperature and cumulative relative expression of *rbcL* is shown. The differing colours indicate the different Rubisco forms. The dotted lines are a linear regression of temperature and expression for each *rbcL* form to highlight the general trend of expression.

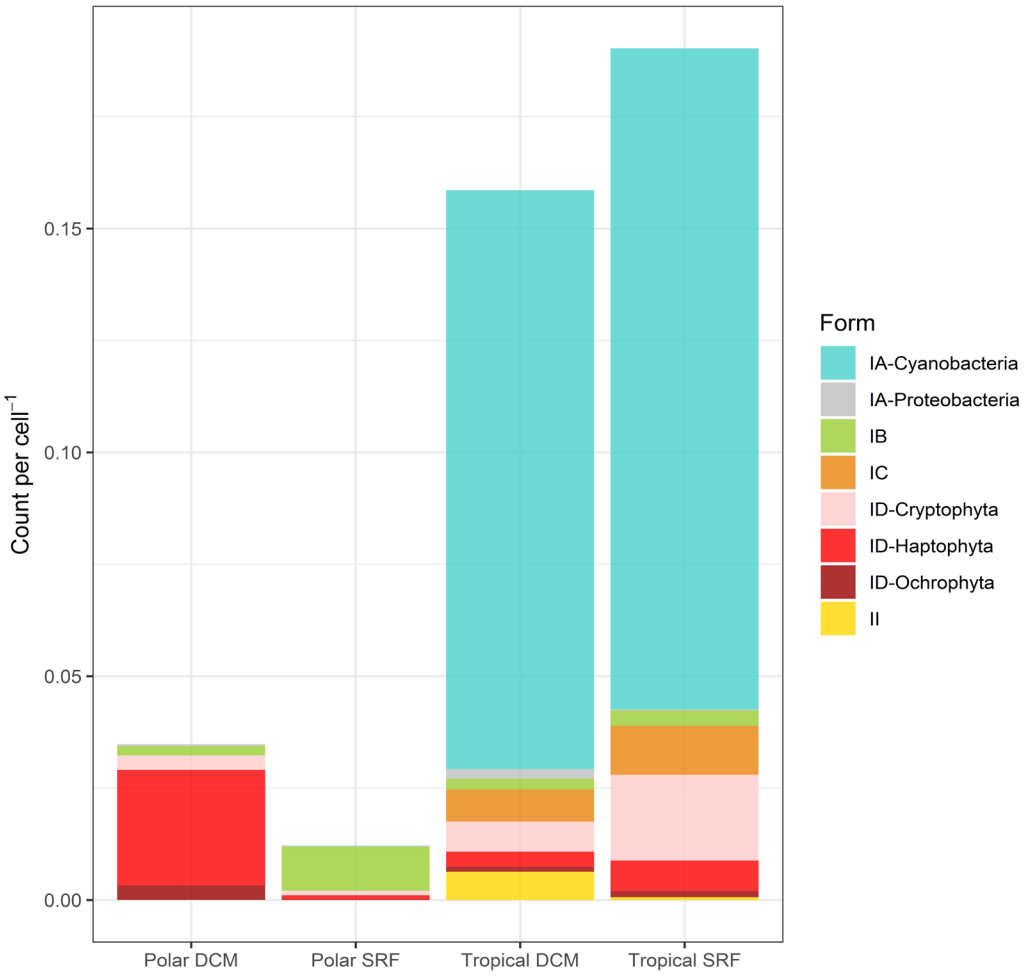


Figure 2.10- The relative abundance of *rbcL* forms for both **(A)** Polar environments and **(B)** Tropical environments within SRF and DCM samples. Relative abundance was calculated by using the total normalised metagenomic copy number for each *rbcL* form across the respective environments. The figure colouration corresponds to each *rbcL* form. Form ID and IA were subdivided into their prospective phyla as they were found to be the dominant form across polar and tropical environments.

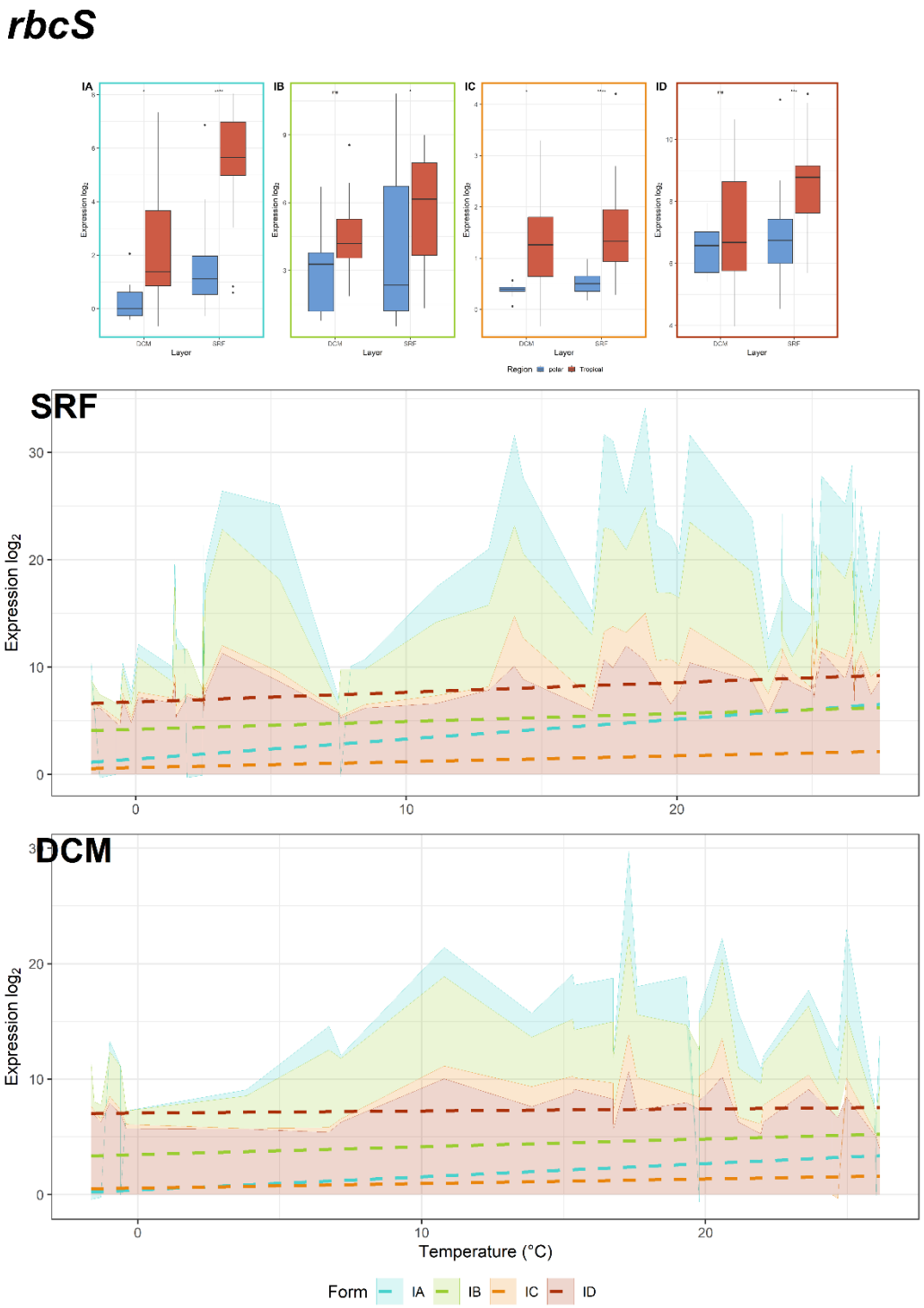


Figure 2.11- The relative expression difference between Polar and Tropical systems of the *rbcS* forms considered in this study. For this log_2 normalised gene expression values were calculated across all SRF and DCM polar sample sites and tropical surface water sample sites difference was measured by a comparison of Medians through a non-parametric Mann-

Whitney U test for polar and tropical waters. Additionally all SRF sample sites and DCM sample sites were ordered in increasing temperature and cumulative relative expression of *rbcS* is shown as a stacked line graph. The differing colours indicate the different Rubisco forms. The dotted lines are a linear regression of temperature and expression for each *rbcL* form to highlight the general trend of expression.

2.3.6 Rubisco co-expression with accessory proteins

For every SRF and DCM sample site including those from temperate and subtropical waters, expression levels were calculated for various accessory genes associated with the Rubisco forms examined in this study. These relative expression levels were used to test for correlation between Rubisco and the accessory protein. In addition the gene *valS* (K01873) encoding for the valyl-tRNA synthetase was used as a negative control. This gene is present across all phyla and constitutively expressed allowing for contrast with the regulated genes examined here. When the average expression of the MG-KEGG genes was calculated across all sites, *valS* was found to be the medially expressed of the 10 PMG genes. The degree of correlation was measured with Pearson's correlation coefficient.

For form IA Rubisco derived from cyanobacteria and proteobacteria correlation was considered for the Rubisco activase *cbbx*, which is also shared across Red lineages. The Rubisco assembly factor *raf2* which transiently binds with Rubisco to form the holoenzyme was also as well as the two genes associated with carboxysomes within alpha cyanobacteria. These are the major carboxysome shell protein *csos2* and the carboxysome associated carbonic anhydrase *csos3* (Figure 2.12).

CbbX derived from cyanobacteria was found to have a weak positive correlation with *rbcS* expression ($r=0.488$, $p<0.001$). This was also the case for *rbcL* and *raf2* ($r=0.437$, $p=0.001$). When considering the carboxysome shell proteins, there was a strong positive correlation between that of the cyanobacterial *rbcL* ($r=0.711$, $p<0.001$) but conversely there was a lightly weaker negative correlation between *rbcL* and the carboxysomal associated carbonic

anhydrase csos3 (r= -0.467, p<0.001). There was no significant correlation between that of *rbcL* and *vals* (r= -0.118, p=0.274) (Figure 2.12).

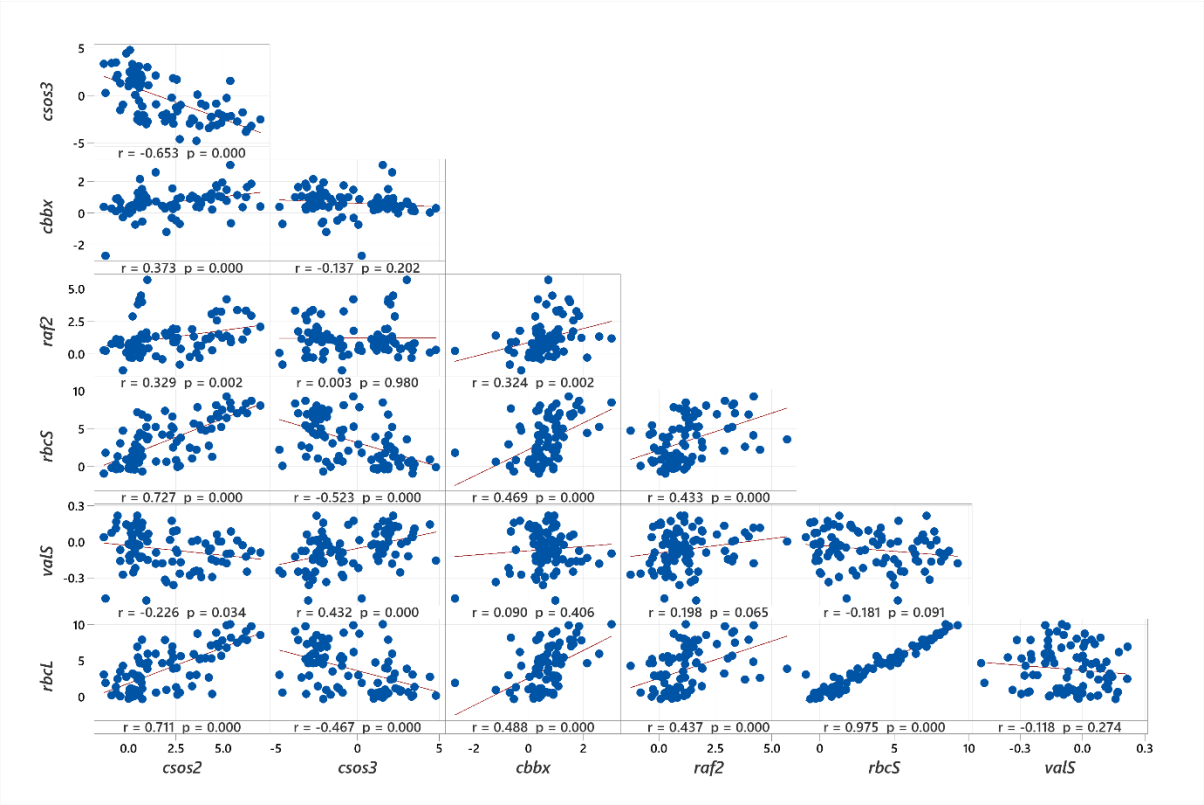


Figure 2.12- A correlation of form IA *rbcL* relative expression (log2) from cyanobacteria and proteobacteria with its associated activase *cbbx*, assembly protein raf2 and the carboxysome genes *csos2* and *csos3*. *valS* was used as a negative control. Pearson's test for correlation was used to assess relationship between relative expression of genes.

For form IB the rubisco chaperone *rbcx* was considered as well as the form IB rubisco activase *rca*. For the chaperone *rbcx* there is a very weak positive correlation between *rbcL* and *rbcx* expression (r=0.308, p=0.003). For the rubisco activase there is a relatively strong correlation between the *rca* gene and *rbcL* (r=0.484, p<0.001). Interestingly the correlation between form IB *rbcS* expression and *rca* is exceedingly high (r=0.934, <0.001). There is no significant correlation between *valS* and any form IB associated gene (p>0.05) (Figure 2.13).

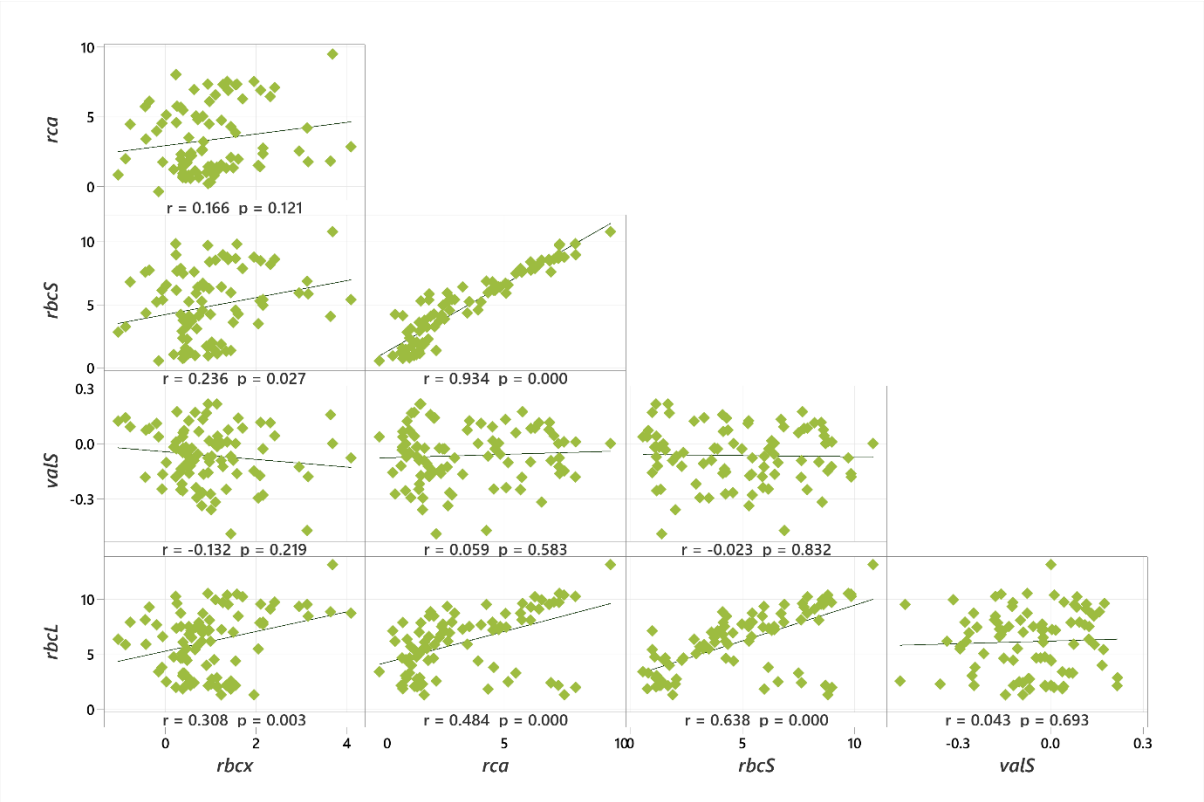


Figure 2.13- Correlation of form IB *rbcL* relative expression (log_2) from Eukaryotes with its associated activase genes *rca and assembly chaperone rbcx*. *valS* was used as a negative control. Pearson's test for correlation was used to assess the relationship between relative expression of genes

For form IC and ID solely the *cbbX* gene derived from proteobacteria and eukaryotes was assessed for correlation with Rubisco large subunit. For form IC this correlation between *rbcL* and *cbbx* was significantly correlated (r=0.561, p<0.001) (Figure 2.14).. For form ID this relationship between expression levels of *rbcL* and *cbbx* exhibited an exceedingly strong positive correlation (r=0.912, p<0.001). For both form form IC and ID there was no correlation between *rbcL* and *valS* (p>0.05) (Figure 2.15).

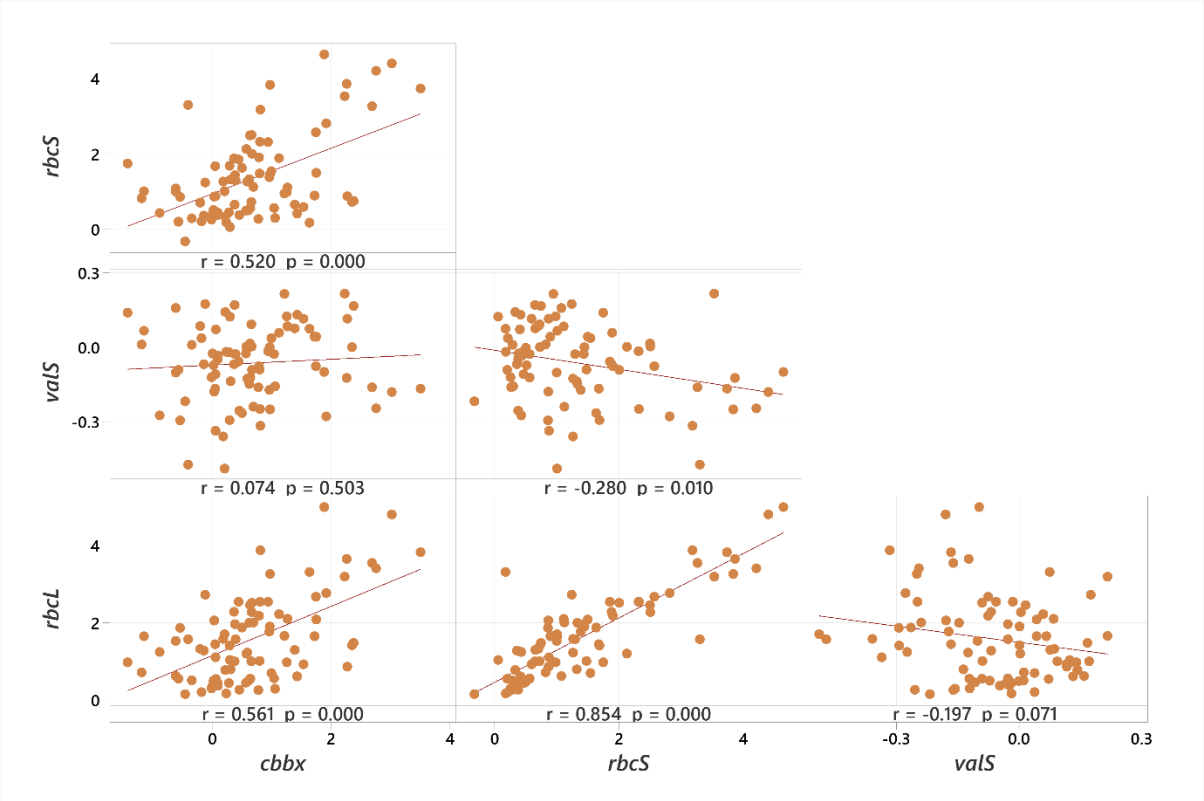


Figure 2.14- Correlation of form IC *rbcL* relative expression (log_2) from Proteobacteria with its associated activase gene *cbbx* and *rbcS*. *valS* was used as a negative control. Pearson's test for correlation was used to assess the relationship between relative expression of genes.

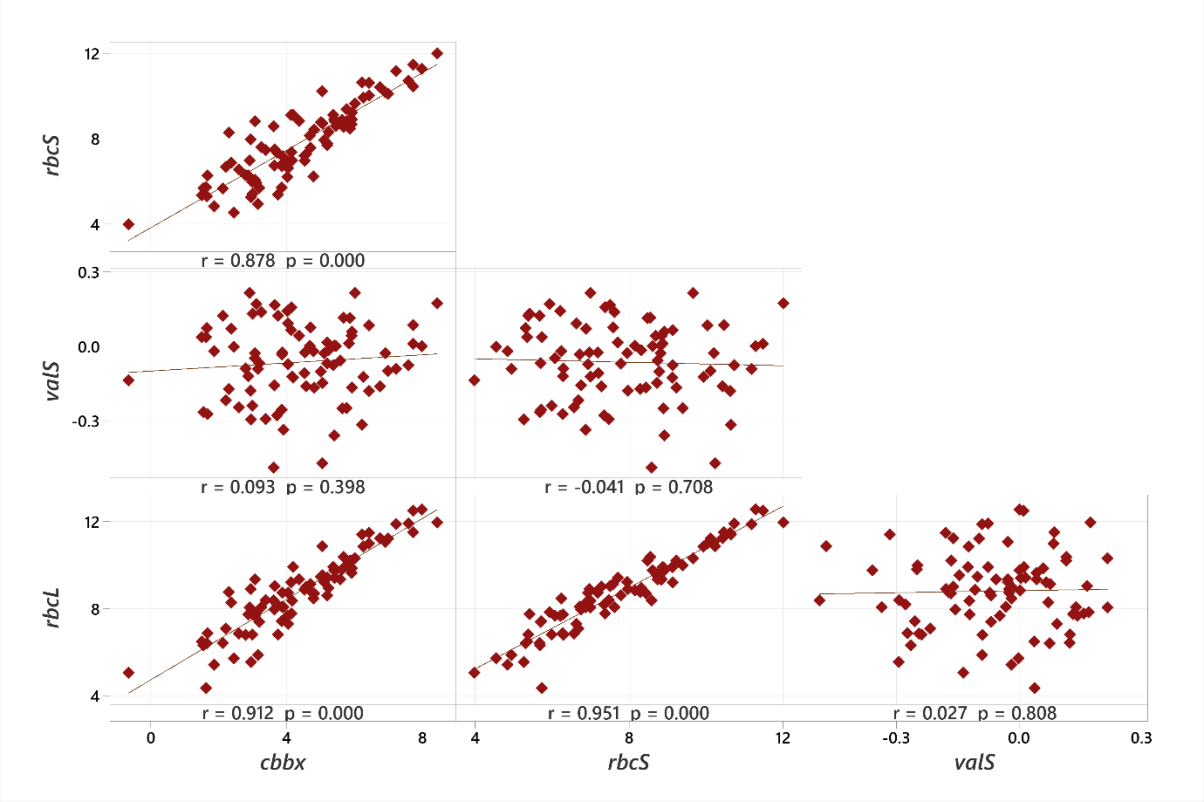


Figure 2.15- Correlation of form ID *rbcL* relative expression (log_2) from Eukaryotes with its associated activase gene *cbbx* and *rbcS*. *valS* was used as a negative control. Pearson's test for correlation was used to assess the relationship between relative expression of genes.

2.3.7 Expression with environmental parameters

To assess gene expression with environmental factors expression was correlated with the environmental parameters of temperature, salinity, oxygen levels, TIC and nitrogen across all SRF (n=57) and DCM sample sites (n=31).

The relative expression levels of *valS* was once again used as a negative control in this study. The *valS* expression levels did not show significant correlation with the environmental parameters of TIC, nitrogen and iron however there was a weak positive correlation with oxygen levels and weak negative correlation with temperature and salinity in both SRF and DCM sample sites (Figure 2.16).

No gene expression in SRF samples and only a few genes in DCM were correlated with nitrogen levels within the system. Interestingly all Rubisco activases but form IC *cbbx* were correlated with nitrogen in DCM samples (Figure 2.16).

Within the photosystems of the light dependant there was a general trend of significantly reduced expression levels with rising water temperature for the genes encoding for the *lhca1*, *lhca2*, the photosystem proteins *psbL*, *psbA*, *psaB* and *psaA* in SRF samples and additionally *petF* in DCM samples. For the *lhca2*, *psbA*, *psbL, psaA* and *psaB* genes there was also a significant positive correlation between expression and oxygen levels in DCM and SRF layers. *petB* had the opposing relationship with the environmental parameters examined here with significant positive correlation with temperature and negative correlations with oxygen in both SRF and DCM samples (Figure 2.16).

For the CBB cycle, for the genes where expression levels correlated with one or more environmental factors there was a ubiquitous negative correlation with oxygen levels in the water system. *pgk*, *gapdh*, *fbp* and *rpi*A expression also had a strong positive correlation with temperature and salinity as well as a weaker negative correlation with TIC levels in SRF sample. In the DCM layer both *rbcL* and *sbp* were negatively correlated with temperature (Figure 2.16).

The photorespiratory pathway in general had a more pronounced positive correlation with temperature than the CBB pathway with the gene *glyA* having an exceedingly strong positive correlation with temperature and an almost equally strong negative relationship with oxygen. This relationship was also the case in SRF and DCM samples for *agxt* and *hpr1* with a strong positive correlation with temperature and a marginally stronger negative correlation with oxygen levels. *pgp* had the lowest expression correlation with environmental parameters however the expression of *pgp* being positively correlated with temperature in SRF samples and negatively correlated in DCM sample. *glyk* expression had the inverse relationship with temperature and oxygen being negatively correlated with temperature and positively correlated with oxygen levels in both SRF and DCM (Figure 2.16).

Within the Rubisco forms form IA, IC and ID *rbcL* had significant correlations with temperature, oxygen, TIC and salinity in both SRF and DCM samples. Additionally, IB *rbcL* was positively correlated with temperature in DCM samples. *rbcS* exhibit the same

expression patterns with environmental parameters as *rbcL* for all but SRF IB *rbcL* genes which were positively correlated with temperature, oxygen, TIC and salinity (Figure 2.16).

The three bicarbonate pumps examined in this study, *sbtA*, *hla3*, *slc4A1* had similar relationships with environmental parameters; being significantly positively correlated with temperature and salinity as well as being negatively correlated with oxygen and TIC levels in both SRF and DCM samples. *sbtA* exhibited the most pronounced correlation with the environmental parameters of temperature and oxygen (Figure 2.16).

Three types of carbonic anhydrase expression were examined in this study. α-ca derived from the gene green pyrenoidal gene *cah3.* As well as a β-ca coded for by the *ptca1* gene. This is known to be associated with pyrenoid function in red algae. The third class being *csos3* associated with bacterial and cyanobacterial carboxysomes. Expression of carbonic anhydrases in this study were shown to have strong correlation with environmental parameters with α-ca, β-ca and *csos3* being significantly correlated with temperature, oxygen, TIC and salinity. α-ca and β-ca expression levels were extremely positively correlated with temperature and extremely negatively correlated with oxygen levels. The inverse was true in *csos3* being strongly negatively correlated with temperature and positively correlated with oxygen levels in both SRF and DCM samples (Figure 2.16).

It is important to note that unpicking correlation of gene expression and the environmental parameters of temperature, oxygen, TIC and salinity is difficult due to the correlated nature of these environmental parameters (Figure 2.4).

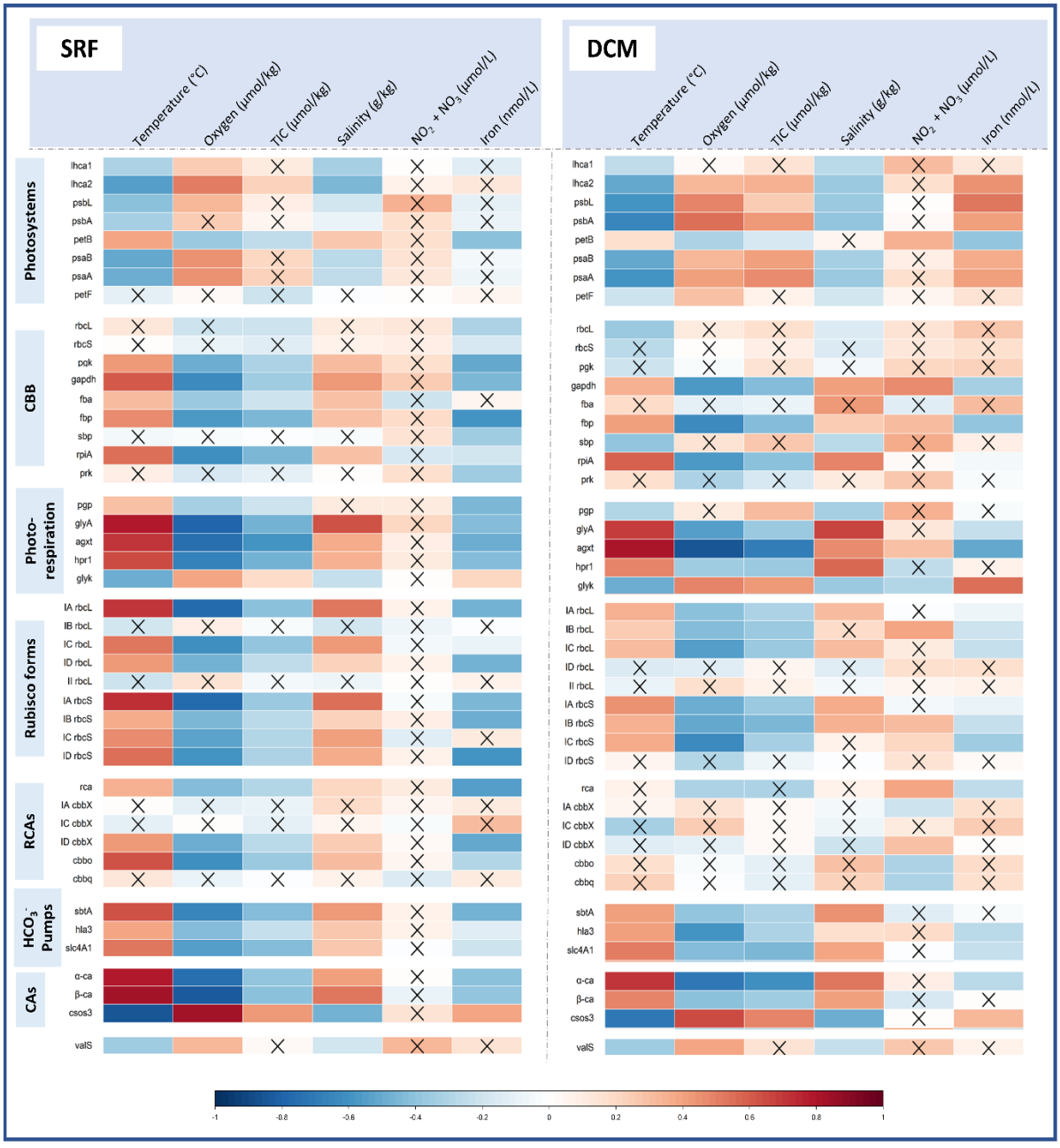


Figure 2.16- The correlation between genes associated with photosynthesis and carbon capture examined in this study with the environmental parameters of temperature, oxygen, total inorganic carbon, salinity and NO_2+NO_3^-. Pearson's correlation coefficient was used to assess correlation strength, squares with 'x's are non-significant correlations ($p>0.05$). Blue colouration represents a negative correlation whereas red represents a positive correlation.

2.4 Discussion

2.4.1 Tropical and polar systems exhibit different expression patterns of metabolic pathways due to environmental conditions

Analysis of surface water samples sites shows a clear distinction in the environmental parameters between tropical and polar environments. Water temperature is the main driver behind this distinction with the polar water examined here not exceeding 8°C and tropical waters being found to be above 17°C at both DCM and SRF layers (Figure 2.1). This temperature distinction has added implications being strongly linked to dissolved gaseous concentrations, salinity and mineral environment within the water systems. All of which play important roles in community dynamics and gene expression within the ecosystem.

Across all sites there was a direct correlation between expression levels of KEGG orthologue genes associated with metabolism and temperature. This is a well-established relationship with metabolic rates rising exponentially with temperature due to increased kinetic energy within biological processes (Gillooly et al., 2001) (Figure 2.5).

Salinity of water systems is also closely correlated with temperature (Figure 2.3). Generally speaking, this is due to increased evaporation of surface water with increasing temperatures resulting in higher salinity. However, there is significant regional and temporal variation based on the influx of fresh water sources (Sigman et al., 2004). This affect is particularly clear in polar waters which had a much higher variation in salinity than tropical waters, likely a result of freshwater influx from ice melts. Within this study the most upregulated gene in tropical systems was found to be the gene encoding for Ectoine Synthase. Ectoine Synthase is involved in ectoine production which acts as a compatible solute to aid against osmotic stress (Widderich et al., 2014) The fact that this gene is so highly expressed in tropical environments suggests that osmotic pressure is a clear stressor for tropical organisms and ectoine production is essential to regulate internal osmotic pressure. Temperature may have a compiling effect on salt stress increasing membrane permeability with higher temperatures (Blicher et al., 2009).

The strongest link between temperature and other environmental factors was between that of dissolved oxygen concentrations and TIC levels (Figure 2.3). Water temperature effects gas solubility with higher solubility in colder waters. Within the water column, oxygen levels

can become more limiting with depth as the rate of respiration begins to out weigh that of photosynthesis due to decreasing light levels. At a point imbalance of respiration to photosynthesis becomes so extreme that the water is defined as anoxic favouring single celled anaerobic organisms (Bertagnolli and Stewart, 2018). However, within SRF waters and the DCM oxygen levels are rarely limited due to the high levels of photosynthesis and an established equilibrium with atmospheric O_2.

The highest environmental variation in tropical samples relative to polar samples was observed in nitrate and nitrite concentrations. Nitrogen levels were particularly high in sample sites located in close proximity to the Galapagos islands (Figure 2.1). These islands are famous for their nutrient rich seas due to strong upwellings carrying nutrients up from depths where nitrogen accumulates (Forryan et al., 2021). Nitrogen has previously been demonstrated to be a strong ecological driver with (Pierella Karlusich et al., 2021) demonstrating that nitrogen fixing diazotroph abundance was closely linked to localised upwellings as well as phytoplankton diversity. These organisms are essential to marine ecosystems playing a critical role in marine nitrogen cycles (Henson et al., 2021). Despite this in this study we failed to find a significant link between gene expression and NO_2 and NO_3^- concentrations in the water system suggesting that it is non-limiting in the sites studied.

Like salinity, TIC and iron concentrations also have links to ice melt, showing higher concentrations and far greater variation in polar waters than tropical waters. Sea ice prevents ocean-air CO_2 exchange but upon melting can deposit encased $CaCO_3$ into the marine system. The resulting dissolution of CO_3^- results in a net increase of TIC in the local water system (Assmy et al., 2013). Additionally bioavailable iron influx to polar systems has been demonstrated as a result of melting ice sheets and glaciers (Bhatia et al., 2013). Within tropical systems iron is often limiting for photosynthesis and this is reflected in an inverse relationship between chlorophyll A and iron concentrations. Influx of iron within tropical systems is typically a result of deposition of aeolian dust carried by wind from Saharan and sub-Saharan Africa. This is reflected by the high iron concentrations found within tropical Atlantic oceans and the iron deplete sites of the tropical Pacific. Cyanobacterial species, namely Trichodesmium sp. have been demonstrated to coordinate

aggregation into multi-cellular ‘ball-like’ structures to increase dust capture in which it can alleviate free iron from oxides contained within the dust (Basu et al., 2019).

2.4.2 Form ID Rubisco organisms dominates carbon fixation in polar environments and cyanobacterial form IA Rubisco in tropical environments

Previous studies have compared Rubisco expression levels within individuals or within the same environmental system across time periods (Young et al., 2015b). In this study we compare Rubisco expression across systems and the globe. For this it is important to first consider the relative abundance of Rubisco forms in polar and tropical waters based on metagenomic copy numbers. There is a clear distinction between polar and tropical waters with form IA cyanobacteria dominating tropical ecosystems and form ID Eukaryotes dominating polar ecosystems (Figure 2.10). This is an established trend with diatoms dominating high latitude ecosystems (Benoiston et al., 2017) and cyanobacteria being the dominant organisms in tropical waters (Capone et al., 1998). Within polar waters cyanobacteria represented <1% of total *rbcL* gene copies found. Cyanobacteria are more closely associated with glacial and ice capped environments in polar environments; characterised by highly changeable temperature and salinity conditions in contrast to the ubiquitously low temperature of the polar oceans (Vincent, 2002) Within the polar marine systems form IA Rubisco derived from proteobacterial organisms represented a higher proportion of abundance in polar waters, being often chemoautotrophic, proteobacteria may be better suited to the prolonged dark winter where sunlight is limiting (Alfreider and Bogensperger, 2018). Additionally form II Rubisco sequences were found to make up only a nominal percentage (<1%) of total abundance in both tropical and polar SRF communities despite being more abundant at tropical DCM sites. Form II Rubisco is often linked to sulphur-oxidative pathways as the energy source, the abundance of form II Rubisco sequences has been shown to be significantly correlated with lower depths in the water column which are defined by oxygen deplete conditions (De Corte et al., 2021, Baltar et al., 2023)

Community dynamics did not significantly differ in Rubisco form abundances between SRF and DCM water layers in both polar and tropical environments. The most significant difference between these layers is the light intensity (Moeller et al., 2019) suggesting that a certain Rubisco form is not better suited to energy flux from high light intensity.

2.4.3 Rubisco expression increases with temperature in marine environments with form IB rbcL being the exception

For forms IA, IC and ID expression of Rubisco large subunits were found to be significantly higher in tropical SRF waters than polar SRF waters with a positive correlation with temperature. However this positive correlation was weaker in DCM samples and for form ID organisms this significant difference in Rubisco expression between polar and tropical sites was lost at the DCM layer. A common theme observed across IA, IC and ID expression in DCM layers is a much more variable level of expression in tropical samples when compared to polar DCM sites (Figure 2.9).

The mechanisms underpinning regulation of Rubisco expression in these Rubisco forms is largely unknown with each form differing in their prospective regulatory elements. Upstream *LysR* family transcriptional regulators are associated with Rubisco regulation in proteobacteria and cyanobacteria. A eukaryotic homologue of y*cf30* is the predominant regulatory element for form ID Rubisco operons. The mechanism of activation in each of these forms is complex, being linked to TIC, light and even rubisco activase expression (Toyoda et al., 2022, Minoda et al., 2010, Böhnke and Perner, 2017). What is clear is that there is a complex cascade of genes across multiple regulatory elements often intertwined with CCM elements and photosystem regulation (Bolay et al., 2022, Minoda et al., 2010).

As a result of this it makes it impossible to categorically predict what environmental factors are influencing Rubisco regulation in the marine environment. However comparing themes of expression in this study with previous studies on Rubisco regulation in single species shows overarching themes. For example the overall higher expression levels of form IA Rubisco in topical waters correlates with a study by (MacKenzie et al., 2005) showing upregulation of *rbcL* within *Synechococcus elongatus* in response to higher temperatures.

For form ID organisms (Young et al., 2015a) demonstrated higher concentrations of Rubisco per gram of biomass in pyschrophillic diatoms from the Western Antarctic Peninsula than mesophilic diatoms (Brown, 1991). The proposed theory behind higher Rubisco concentration in psychrophilic diatoms is to overcome slow kinetic rates which constrain Rubisco at low temperatures (Young et al., 2015a).

This is contradictory to the lack of significance difference in form ID Rubisco expression observed between DCM polar and tropical samples and significantly higher expression in tropical SRF waters. The reason behind this contradictory evidence may be nitrogen limitation. Higher variation in form IA,IC and ID expression at the DCM layer appear to be coordinated with higher variation in nitrogen levels observed across tropical DCM samples (Figure 2.16). Although Rubisco form expression was not linearly correlated with nitrogen concentrations, nitrogen may be limiting past a threshold. This therefore requires a balancing of nitrogen allocations across the cell. In contrast to this Antarctic oceans which are rich in nitrogen due to ocean upwellings allows unrestricted allocation of nitrogen to Rubisco to overcome the bottleneck in these communities (Young et al., 2015a), (Pierella Karlusich et al., 2021).

It is also important to note that inferring a relationship between transcription levels and protein concentrations is imperfect and is an over simplification.

In contrast to the other Rubisco forms Form IB *rbcL* had higher expression levels in polar water than that of tropical waters. This was the sole Rubisco form where this was the case. Upregulation of form IB Rubisco in response to cold conditions has been demonstrated in multiple species of both plants and green algae (Zhang et al., 2002, Ohba et al., 2000, Cavanagh et al., 2023, Devos et al., 1998, Peng et al., 2021). Red algae have frequently been demonstrated to possess Rubisco with specificity levels far higher than that of green algae (Oh et al., 2023). This may be the reason why resource allocation to Rubisco can be moderated in polar environments within form ID Rubisco species, dependant on nitrogen resources. Alternatively form IB organisms require maximum nitrogen allocation to Rubisco in cold environments to overcome bottlenecks due to slow kinetic rates and reduced specificity. This theory is supported by the proportion of total protein Rubisco makes up in form IB aquatic organisms versus form ID organisms with form IB organisms possessing upto 63% Rubisco (Rubisco/ grams of biomass). In comparison diatoms were demonstrated to possess between 1-12% Rubisco (Rubisco/ grams of biomass) (Young et al., 2016).

Form IB organisms have often been shown to possess multiple isoforms of *rbcS* genes which have been demonstrated to confer differential kinetic properties of the holoenzyme. Multiple studies in wheat, Arabidopsis and spinach have shown that differential expressions of the small subunit isoforms, coordinated with temperature, can modify the kinetic

parameters of the Rubisco holoenzyme. Often increasing k_{cat} and decreasing relative specificity at lower temperatures (Yamori et al., 2006), (Cavanagh et al., 2023), (Huner and Macdowall, 1978).

This differential expression of small subunit isoforms may also explain the discrepancies between rbcL and rbcS expression levels in form IB organisms observed in this study. This coupled with the spatial segregation of rbcL (plastid) and rbcS (nuclear) genes in the meant that rbcs and rbcL expression levels were not as tightly coordinated in form IB organisms compared to form IA,IC and ID organisms (Figure 2.8).

Within IA, IC and ID organisms rbcL and rbcS genes are located within operons with the same regulatory promoter. Therefore in this study an almost 1:1 relationship was found between the expression levels of rbcL and rbcS which was to be expected (Figure 2.8).

2.4.4 Rubisco expression correlates strongly with Rubisco activase levels but not with chaperones

Across all forms there was a distinct positive correlation with *rbcL* and the corresponding Rubisco activase (Figure 2.12, 2.13, 2.14, 2.15). This relationship was particularly striking in form ID with an almost one to one ratio of expression of *rbcL* to the *cbbx* gene. This is because within red type organisms *cbbx*, *rbcL* and *rbcS* are located in the same operon so one would expect expression to be comparable. This is also the case for form IC organisms however Rubisco expression did not correlate as strongly with *cbbx* expression. This may due to missed sequences in the annotation of sequences in this study due to a lack of annotated *cbbx*, proteobacterial sequences available on NCBI databases.

Proteobacterial organisms possessing form IA Rubisco can be divided into two sub groups, IA^{QO} and IA^{C} with the latter Rubisco assembling into carboxysomes. Both subforms share the common CBBQO activase complex however this is found further downstream in IA^{C} organisms. This decoupling of Rubisco genes and activase may be why it appears the correlation between form IA *rbcL* and *cbbQ*/*cbbO* expression in proteobacteria is weak.

Within form IB organisms there is also strong positive correlation between the Rubisco activase rca and rbcL but this relationship is far stronger between *rca* and *rbcS*. Both *rbcS*

and *rca* are nuclear encoded and differ in regulation patterns with evidence showing that *rca* (Carmo-Silva and Salvucci, 2013) is redox regulated connected to light intensity.

Across all Rubisco forms there is a weak positive correlation between *rbcL* and their corresponding Rubisco chaperone. This is expected with chaperones only binding transiently, not necessitating tightly linked expression levels. Chaperone genes are also not generally found within the Rubisco operon (Cabello-Yeves et al., 2022)

Within cyanobacteria, form IA assembles into alpha carboxysomes consisting of the Rubisco, the large shell protein *csos2* and *csos3*, a carboxysome associated carbonic anhydrase alongside various smaller linker proteins. Despite being clustered within an operon short intergenic gene spaces may result in intra-operon regulation of genes. (Cai et al., 2008) demonstrated that transcript levels of *cso* operon genes within the proteobacterial *Halobacillus neopolatanus*, varied by magnitudes. Despite being adjacent, *rbcS* and *rbcL* had ten-fold higher transcript levels than *csos2* and *csos3*. This discrepancy of Rubisco levels to *csos2*/*csos3* levels is also represented at the protein level. (Sun et al., 2022)

It appears that *csos3* is also upregulated with temperature and CO_2. This is opposing to form IA Rubisco expression. It is also the opposite of α and β CA levels. The αCA, *ptCA1* studied here have been linked to extracellular conversion of bicarbonates to CO_2 in cyanobacteria (Kupriyanova et al., 2011) whereas βCA, *cah3* has been implicated in pyrenoidal function in green algae (Gee and Niyogi, 2017) and cytosolic localisation in diatoms (Tanaka et al., 2005) involved in the CCM. This trend of upregulation of carbonic anhydrases with decreasing dissolved inorganic carbon levels has been demonstrated as a stress response to growth in low carbon environments. (Clement et al., 2016) Therefore it is intriguing that *csos3* has the inverse relationship with carbon levels and temperature. The significant negative correlation with *rbcL* may indicate malleability in carboxysome architecture with internal CA concentration decrease as Rubisco levels increase.

2.4.5 The light dependant stage of photosynthesis is upregulated in polar systems

Looking more specifically at the light dependant stage of photosynthesis in this study we found an overarching theme of higher expression of genes in both SRF and DCM polar

samples compared to those of tropical (Figure 2.6). This response may be multifactorial based on a number of differences between tropical and polar environments.

Firstly polar environments are highly seasonal with vast changes in sunlight hours over the course of a year. The sampling for the polar waters was conducted between late May and late October. Meaning that large parts of the sampling effort were conducted with 24 hours of daily light. Large algal blooms have been observed during this period in polar waters as a result of increased productivity due to long daylight hours (Tison et al., 2020).

However high-light conditions also come with significant implications for photosynthetic organisms mainly arising from the imbalance of photochemical and enzymatic rates. Photochemical transfer of energy happens rapidly, irrespective of temperature at 10^{-15} µmol s^{-1}, enzymatic rates however lag far behind due to low temperatures (Young and Schmidt, 2020). This necessitates the evolution of biochemical and biophysical mechanisms to alleviate potential damage from this energetic imbalance.

Once such mechanism observed in *Micromonas spp.* shows the upregulation of photosystem II proteins as protein reserves for when photosystems become damaged or in need of repair. This may additionally contribute to higher expression levels within polar regions (Ni et al., 2017).

Another significant factor is iron with levels in polar waters being significantly higher than that of tropical waters. Iron is an essential cofactor involved across many photosystem proteins in the form of haem groups or iron-sulphur clusters. Therefore environmental iron levels are often limiting for the light dependant stage of photosynthesis. It is considered that 30% of marine environments are limited in iron. Within diatoms the c6 complex is often replaced by a plastocyanin, a copper centred homologue. Being dominant in polar waters, this may allow more amplified expression of photosystems (Peers and Price, 2006)

The one exception to higher expression levels in polar waters is found with the gene, *petB* complexes with other pet genes to form the cytochrome b6/f complex. Studies in plants show that while photosystem and light harvesting complex genes are down regulated under heat stress cytochrome b6/f genes, including *petB* are upregulated (Song et al., 2014). This is because cytochrome c6/f is considered to act as a switch between photosystems controlling electron flux and acting as a photoprotective mechanism (Johnson and Berry, 2021).

In DCM water samples we found a strong positive correlation between photosystem expression and iron concentrations in samples. A relationship that was not observed in SRF waters. This suggests within surface waters photosystem expression is regulated to prevent photooxidative damage to cells through excessive energy flux. Within the DCM water layer where light intensity is minimal, photosystem genes are upregulated to a point where iron becomes limiting for the organism.

2.4.6 CBB and photorespiration follow the metabolic trend of upregulation with temperature

After the light dependant stage of photosynthesis comes the light independent stage with carbon dioxide being fixed through the central enzyme of Rubisco. When all forms of Rubisco (including form III lacking carboxylation capacity) were considered, it was found that expression levels did not differ significantly between polar and tropical environments. This was not the case for other CBB associated enzymes with a general upregulation of the pathway in tropical sample sites. This is in accordance with the general trend of upregulation of metabolic processes with temperature. Expression levels within the CBB were by the highest in *rbcL*, *rbcS* and *sbp* by far. Both Rubisco and SBPase have been established as bottlenecks within the CBB cycle due to slow catalytic rates Liang (Liang and Lindblad, 2017), (Hammel et al., 2020). High expression levels within marine systems may be necessary to compensate for these slow turnover rates.

Upon Rubisco fixing O_2 3-PGO is produced which is toxic in high concentrations. Therefore this product must be recycled through the photorespiration pathway. Photorespiration was also generally upregulated in tropical systems relative to polar systems despite lower concentrations of oxygen. This is clearly a proportionate increase to the higher expression levels of the Rubisco forms. The exception to this upregulation in tropical waters was *glyk* which was upregulated in polar waters. Within cold adapted *Chlamydomonas* species *C. nivalis* upregulation of *glyk* was also found to be the case when grown at low temperatures (Peng et al., 2021). Glyk like Rubisco and SBPase may represent a bottleneck in photorespiration and therefore further study here is important.

When considering genes associated with carbon fixation through the CBB cycle, it would also be impossible not to understate the effect of mixotrophy in marine environments. This

adds a layer of confusion when considering carbon cycles in marine systems. The extent of mixotrophy within marine populations contributes to an estimated 12% of marine microeukaryotic gene sequences (Cohen, 2022). This is an emerging area of research which is expanding with the rise of metagenomics. However, it is clear that mixotrophy is widespread amongst microbial marine communities with the majority of organisms balancing both autotrophic and heterotrophic metabolism (Stoecker and Lavrentyev, 2018), (Flynn et al., 2019).

2.4.7 Carbon concentrating mechanisms in marine systems are strongly correlated with environmental parameters

There are two main components to carbon capture mechanisms with marine environments, these are bicarbonate pumps which actively transport inorganic carbon from the environment into cells and secondly there is the carbonic anydrases which convert this bicarbonate to gaseous CO_2 for fixation by Rubisco. In this study we examined, the bicarbonate transporters *sbtA* (commonly found in alpha-cyanobacteria), *hla3* (associated with green algae) and *ptslc4A1* (a carbonate transporter associated with red microalgae). *sbtA*, *hla3* and *ptslc4A1* all showed an inverse relationship with TIC (Figure 2.16). This relationship with TIC is also reflect in the carbonic anhydrases associated with CCM activity in green and red algae (ptca1 and cah3 respectively). Dissolved inorganic carbon levels are higher in polar environments, especially in summer months largely due to the influx of TIC through glacial melt (Lønborg et al., 2020). This coupled with higher productivity during polar summer months may necessitate greater levels of carbon import by bicarbonate transporters. For carbonic anhydrases and bicarbonate pumps it has been demonstrated that upregulation occurs at reduced carbon levels within microalgae (Clement et al., 2016). Although TIC loosely decreases with temperature an increase in temperature also results in a shift of the carbonate equilibria to cause a relative increase in bicarbonate ions in the water system (González-Benítez et al., 2019). The result of which may mask a level of coordination between TIC and bicarbonate pump expression.

2.5 Conclusions and future prospects

This study is the first of its kind giving an insight into photosynthesis regulation on a global scale in marine systems. There is a definitive link between expression of photosynthetic

genes and environmental conditions with all genes coordinated with one or multiple environmental factors.

For the light dependant stage of photosynthesis regulation is inversely linked to temperature with higher expression in polar sample sites with the added limitation of iron at increased depths in the water column. Iron is not significantly linked to expression in surface waters suggesting that there is a redox regulation of photosystems to avoid over excitation, an issue that may be exacerbated at low temperatures.

Interestingly the light independent stages of the CBB cycle and photorespiration appear to have the inverse relationship to the light dependant stages being positively linked to temperature and negatively linked to iron concentration in the majority of cases. This opposing relationship may represent the differences between rate of enzymatic activity and the charge separation rate of photosystems. Charge separation is not affected by temperature and therefore photosystems have to be regulated accordingly for energy dissipation. Alternatively enzymatic rate is largely dependent on temperature and expression is therefore proportional to kinetic rate.

When Rubisco forms were considered, we found nuanced expression patterns emerging between different Rubisco forms as well as a disconnect between *rbcL* and *rbcS* expression in form IB organisms. This is unsurprising as the upstream photosystem apparatus and the CCM architecture differs significantly between organism, generating different Rubisco environments. This coupled with the differing Rubisco accessory proteins and regulatory elements may allow for varying levels of environmental adaption.

A significant limitation of this study is the lack of data associated with light intensity, being intrinsically linked to photosynthesis. The major difference between marine environments and terrestrial environments is the lowered effect of environmental conditions on diurnal cycles with temperature, nutrients, osmotic pressure and gaseous concentrations having little fluctuation. Opposingly, especially within terrestrial plants, temperature, gaseous concentrations and osmotic pressure are fluctuating significantly across the length of a day, making the environment much more changeable, having added implications on photosynthesis. The one exception to this comparison is light intensity, with both marine and terrestrial systems experiencing daily fluctuations.

Climate change and rising water temperatures are a threat to global marine systems, this is particularly true in polar ecosystems where water temperatures are rising at amplified rates. There is a clear correlation between water temperature and the nutrient environment. This has additional links to ecosystem taxonomy and photosynthesis regulation. It is important that we make efforts to curb rising sea temperature in order to prevent added disruption to highly productive but fragile polar ecosystems.

Environmental adaption of Rubisco in the Earth's oceans

3.1 Introduction

In Results 2.3, it was demonstrated that temperature was a significant driver behind expression patterns in the Earth's oceans, being strongly correlated with the regulation of almost all photosynthetic genes. As well as being a significant factor driving expression, environmental temperature also defines the internal temperature of unicellular organisms (Somero, 1995). As a result, enzymes will commonly evolve thermal optimas that reflect the mean temperature of the environment (Feller and Gerday, 2003). With diurnal temperature of oceanic environments varying to a much lesser extent than those of terrestrial environments this phenomenon is particularly prevalent in marine organisms. A comprehensive study demonstrated this shift in thermal optimas within marine environments, highlighting a strong positive correlation between the enzymatic thermal optima and mean annual temperature based on a wide array of enzymes derived from oceanic metagenomes (Marasco et al., 2023).

Contrary to this, previous studies have failed to show a shift in thermal optima within Rubisco species with rate of carboxylation continuing to rise exponentially with temperature (Young et al., 2016). However, there is phylogenetic evidence to suggest evolution of Rubisco to fit the environment (Hermida-Carrera et al., 2017), (Galmés et al., 2014), (Kapralov and Filatov, 2007). Although environmental adaption of Rubisco kinetics may not be as strong as a thermal optima shift, there are subtle differences in the kinetic parameters of specificity and rate between environments. (Capó-Bauçà et al., 2022) provides evidence to suggest these differences may be the result of mutations within the RbcL, highlighted within closely related species of seagrass. Additionally, environmental shifts in the kinetic properties of Rubisco can be observed through the differential expression of RbcS isoforms (Cavanagh et al., 2023). In *Arabidopsis*, a decrease in temperature causes this change in RbcS isoform expression resulting in a subtle change of kinetic parameters better suited to colder conditions with faster, less specific Rubisco being favoured at lower temperatures (Cavanagh et al., 2023).

In this study, we examine the diversity and evolution of Rubisco in seas and oceans by utilising the vast resource of sequences derived from Tara Oceans metagenomes. By

combining metagenomic sequences with their corresponding environmental temperature, we can additionally begin to contrast evolution rates between cold, warm and temperate environments through phylogenetic evolutionary models. In this chapter branch site models are utilised to examine intensification/ relaxation of selection pressure from temperate environments into warm or cold environments within Rubisco forms, additionally episodic and pervasise selection models are utilised to capture selection pressure which are subjected across the Rubisco gene from differing environmental temperatures. In parralel to the codon based phylogeny models the machine learning models of Gaussian process regression and random forest models are used to examine sequential differences between Rubisco proteins derived from warm and cold environments. Guassian process models have previously been utilised in the prediction of various protein characteristics from flourescence, themostability and kinetic properties (Saito et al., 2018, Romero et al., 2013, Iqbal et al., 2023). Thus in this study we apply the Gaussian process model to see if the Rubisco protein sequence is predicative of its environmental temperature which would indicate an underlying environmental adaption**.**

Therefore, our null hypothesis is that we will find no evidence of positive selection in Rubisco and there is no evidence of evolution of Rubisco in relation to the environmental temperature. Additionally, because of this lack of evolution the machine learning models will not be able to determine environmental temperature from protein sequence.

3.2 Methodology

3.2.1 Metagenomic mining

Multiple studies have synthesised gene catalogues from the Tara Oceans genomic information (Delmont et al., 2022), (Delmont et al., 2018) (Vorobev et al., 2020), (Royo-Llonch et al., 2021), (Acinas et al., 2021), (Salazar et al., 2019). These reference catalogues were collated into a searchable tool by (Vernette et al., 2022). Bait from each form of Rubisco (Appendix Table 6.2, 6.3) was used to tBlastn search against the various gene catalogues with an e-value threshold of $1x10^{-10}$.

As well as obtaining the raw protein sequences, it is possible to obtain the DNA sequence and environmental parameters of each sequence found in the reference catalogues. A custom python script was used to collate the above information. As many Rubisco genes were found across multiple sites, the temperature from each site was averaged to give the average temperature assigned to that sequence. Based on the average temperature that each gene was found, a 'temperature category' was assigned to the gene. These temperature categories were defined as Cold (<10 ^{O}C), Temperate (10-20 ^{O}C) and Warm (>20 ^{O}C).

3.2.2 Phylogenetic determination of Rubisco Form and preliminary cleaning

Sequences extracted from the Tara Oceans catalogues were primarily cleaned for duplicate sequences; additionally, partial sequences that were below defined minimum length thresholds were cleaned. The length threshold for RbcL sequences was 350 amino acids and 100 amino acids for RbcS protein sequences.

For form determination of Rubisco sequences a combinatorial approach was conducted, assessing phylogenetic determination after protein sequences were annotated using DIAMOND+ (minimum e-value cutoff $1x10^{-10}$) (Buchfink et al., 2015) implemented on the Galaxy EU servers (Galaxy Version 2.0.15). Using these annotations form III Rubisco sequences from Archael origins were removed. The Rubisco sequences were aligned using MUSCLE alignment default parameters, and a Maximum likelihood phylogenetic tree utilising a Dayhoff substitution matrix was generated. Using the assumption that Rubisco sequences grouped by form and knowing the phylogenetic determination of each sequence, it was then possible to define which phylogenetic clades corresponded to each Rubisco form and therefore each Rubisco sequence could be assigned a form.

3.2.3 Further cleaning of sequences for analysis

Due to the nature of the sequence extraction, there were multiple duplicate genes, as identified by DIAMOND+ annotation that shared the same NCBI indicator but differed marginally in length. This duplicity risked creating a positive bias in subsequent analysis where genes between environments were phylogenetically compared. Therefore, genes

that shared both the same NCBI indicator and were found in the same temperature category were removed, leaving only the longest sequence for that gene. Additionally, a further cleaning stage was conducted upon alignment of each individual Rubisco form using MUSCLE. If a sequence was missing >~15 amino acids at the C/N terminus of the protein, it too was removed from the dataset.

3.2.4 Gaussian Process regression model overview

A gaussian process regression model is a supervised model that aims to predict an observation y from a predictor x with a measure of uncertainty. A gaussian process model is defined as a nonparametric model meaning that it is not limited to a single function unlike linear regression and therefore achieves predictions over a range of possible functions that fit the data (Wang, 2020).

To achieve this estimate one must first define the model priors. This defines the characteristics of the functions used by the model. The prior model function can be divided into three parts. The kernel function, the mean function and the hyperparameters of the kernel function. Firstly, the kernel function is used to represent the relationship between the covariance of x variables (Equation 3.2) (Ebden, 2008). The second component of the prior is the mean function μ (Equation 3.1), this defines the mean of the function values at any given point, without prior knowledge this is defined as 0. Finally, the hyperparameters of variance σ^2 and lengthscales l define the height and smoothness of the kernel function respectively (Equation 3.2)(Ebden, 2008). This allows fine-tuning of the kernel function for the most accurate predictions. Figure 3.1 presents examples of these prior functions. The key principle behind the use of a covariance function is that it ensures x variables that are closely related have a similar y value, otherwise known as the function output value $f(x)$ (Wang et al., 2020).

$$f(x) \sim GP(m(0), k(x, x')) \tag{3.1}$$

$$\mathrm{k}(\mathrm{x}, \mathrm{x}') = \sigma^2\left(1 + \frac{\sqrt{5}|\mathrm{x} - \mathrm{x}'|}{l} + \frac{5|\mathrm{x} - \mathrm{x}'|^2}{3l^2}\right)\exp\left(-\frac{\sqrt{5}|\mathrm{x} - \mathrm{x}'|}{l}\right) \tag{3.2}$$

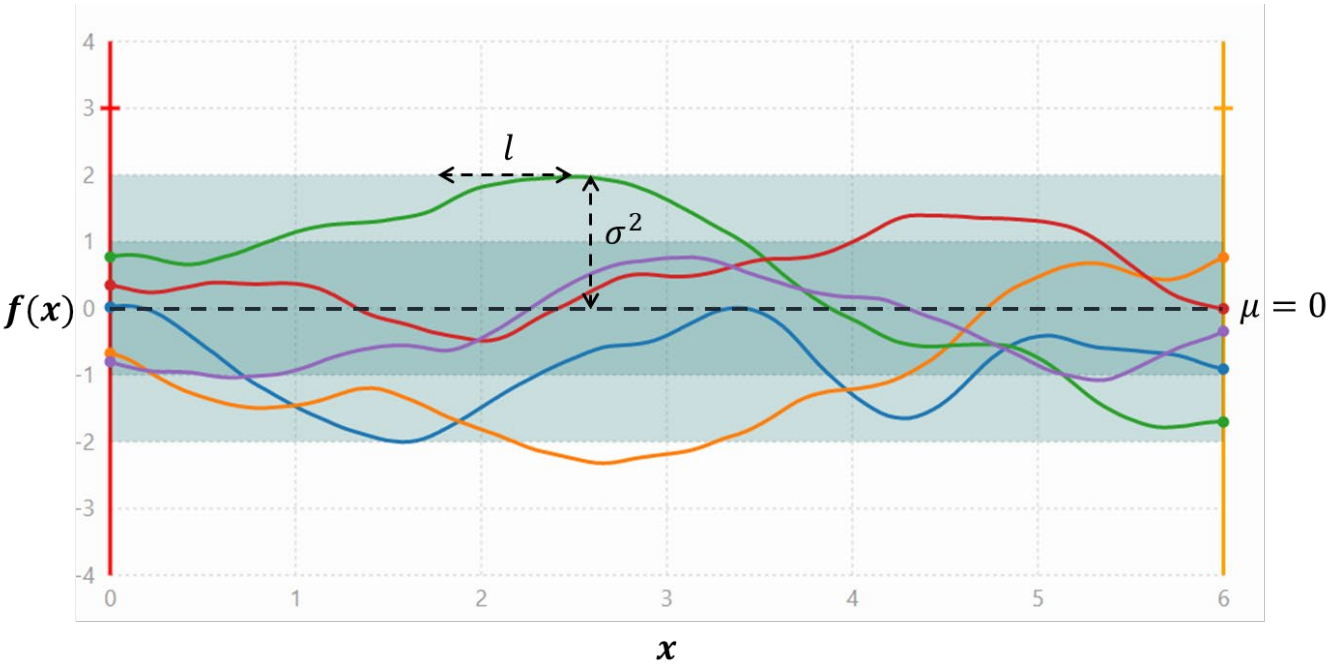


Figure 3.1- A visual representation of the Matern52 kernel function prior to training. The height of the function is defined by the variance (σ^2). μ represents the mean of the function and l is the length scale of the function defining the level of oscillation. st--/int*eractive-gp-visualization github repository*

Training the Gaussian model 3.2.5

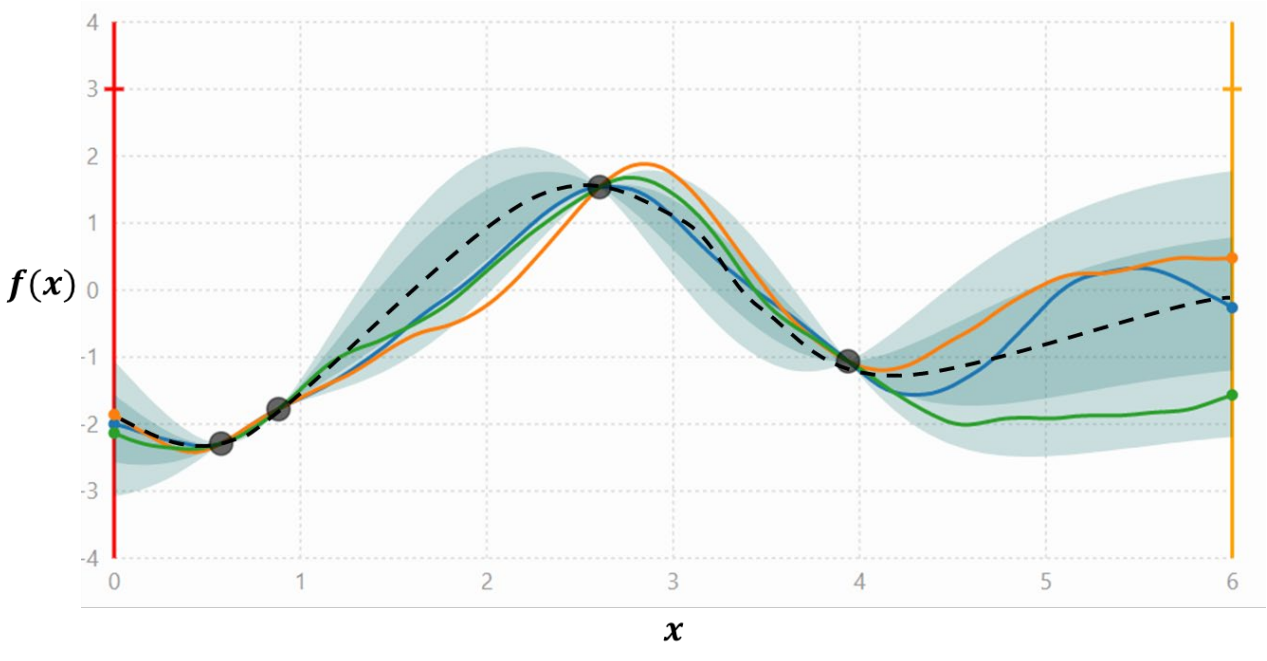


Figure 3.2- *A visual representation of the Matern52 kernel function post training to datapoints. This highlights that only a finite quantity of functions are now applicable to the data. This limits the variance and can be used to predict a mean at unknown x'. The height* of the function is defined by the variance (σ^2). μ represents the mean of the function and l is the length scale of the function defining the level of oscillation. This representation was modified from st-- /interactive-gp-visualization github repository.

The next step is model training of the posterior distribution of functions. For the prior there is an infinite number of functions that can be plotted; however, when training data are incorporated with known x inputs and y outputs that correspond to function outputs $f(x)$ only a selection of functions will fit the data - thus adjusting the mean function (Figure 3.2) (Ebden, 2008).

The model is then optimised through an iterative optimiser that minimises maximum likelihood error in the model by tuning the hyperparameters of the kernel function to get the best fit of the data (Wang et al., 2020).

Finally, the model is used to predict unknown y' from test x_* input. Being Gaussian, this is a measure of probability over the range of normally distributed function values from the posterior giving a mean value for y' and variance. To achieve the calculation of mean and variance the covariance matrix is used. K represents the covariance matrix where the kernel function $\mathrm{k}(\mathrm{x}, \mathrm{x}')$ is used to calculate each element (Equation 3.3) (Ebden, 2008).

$$K = \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{bmatrix} \quad 3.3$$

K_{**} is calculated from the diagonal of the matrix relating to $k(x_*, x_*)$ for test x_* input (Equation 3.4) (Ebden, 2008).

$$K_{**} = k(x_*, x_*) \quad 3.4$$

K_* is a vector of test x_* with trained x values (Equation 3.5) (Ebden, 2008).

$$K_* = [\, k(x_*, x_1), k(x_*, x_2) \ \ldots \ k(x_*, x_n)] \quad 3.5$$

The above components of the covariance matrix are used to calculate the mean prediction for unknown $\overline{y'}$. K^{-1} equates to the inversion of the covariance matrix (Equation 3.6) (Ebden, 2008).

$$\overline{y'} = K_{**} - K^{-1}y \quad 3.6$$

For variance of y', T indicates the transposition of the matrix (Equation 3.7) (Ebden, 2008).

$$var(y') = K_{**} - K_* K^{-1} K_*^T \quad 3.7$$

In this study, both simple and additive kernel architectures were utilised. Simple kernels simply use a single kernel function across all parts of x. Additive kernels use multiple kernel functions over defined parts of x. Upon optimisation, this allows weighting of parts across the length of the protein highlighting lower order interactions (Wang et al., 2020).

3.2.6 One-hot encoding

For the input to the Gaussian Process regression model the protein sequences extracted from the Tara Oceans, outlined above (Methods 3.2.4) assemblies required encoding to a numerical vector. Three main methods of encoding the protein sequence were utilised in this study. Firstly, one-hot encoding. This requires an alignment of all protein sequences found for each of the proteins examined in this study. The alignment was achieved using the MUSCLE algorithm with default parameters. Each protein sequence in the resulting

alignment was then converted to a binary vector where each amino acid is represented within five dimensions of 0s and 1s; for example, A=00001, C=00010 and so on.

3.2.7 VHSE encoding

(Mei et al., 2005) demonstrated that 50 biochemical properties of the 20 amino acids could be condensed to a numerical representation of eight dimensions. Dimensions 1-2 represent hydrophobicity of the amino acids, 3-4 represent the steric properties of an amino acid, and 5-8 represent the electronic properties of the amino acid (Xie et al., 2013).

As described in methods 3.1, the protein sequences were aligned and then a custom python script encoded the aligned protein sequences to a numerical vector where each amino acid was represented by its corresponding VHSE values, for gaps in the alignment the VHSE values were set to 0.00 for all dimensions.

3.2.8 Learnt encoding ESM transformer model

The ESM-2 transformer model developed by (Lin et al., 2022) was used to encode protein sequences into a numerical vector of 1280 elements long. The ESM-2 model is based on a language model that works by 'masking' amino acids along the length of protein and then subsequently predicting the masked item. The model was trained on 50 million protein sequences and encompasses 15 billion parameters in its prediction. (Lin et al., 2022) have demonstrated that these language models have previously been shown to represent biochemical and structural properties of proteins (Lin et al., 2022)and can be used to predict the protein 3D structure comparable to that of AlphaFold2 (Jumper et al., 2021). For this method of protein encoding the esm2_t36_3B_UR50D() model was downloaded from the facebook research/esm github repository. The corresponding python code was used to convert the protein sequences to numerical vectors.

3.2.9 TSNE plots

TSNE plots are a means of visualising underlying trends in highly dimensional data by reducing the number of dimensions to two (Van der Maaten and Hinton, 2008). In this case, binary encoding of the protein alignments were used as the input. Hyperparameters can be manipulated to ensure resolution between datapoints, therefore these hyperparameters

were fixed as perplexity=50, learning_rate=500 and random_state=0. The TSNE analysis was performed using the Scikit Learn package (Version 1.2.2) and visualised using seaborn 0.13.0.

3.2.10 Protein sequence alignments with secondary structural elements

To produce RbcL and RbcS protein alignments with secondary structural elements, Esprit3.0 was used in conjunction with the crystal structures of RbcL from *Arabidopsis thaliana* RCSB: 5IUO (Valegård et al., 2018a) and RbcS from *Chlamydomonas reinhardtii* RCSB: 1GK8 (Taylor et al., 2001). The separate alignments for RbcL and RbcS sequences consisted of the protein sequences from the above crystal structures as well as representative sequences extracted from Tara Oceans metagenomes of form IA, IB, IC and ID origins to highlight sequence diversity across forms. Going forward, the amino acid position index from the RbcS and RbcL alignments with secondary structural elements was used as the reference for all other amino acid positions.

3.2.11 Random forest classifier model to identify residues that differ between 'Warm' and 'Cold' Rubisco sequences

Decision trees define the most efficient way to divide data into defined categories. Decision trees are however not considered robust models as a small change in sampling data can result in a drastically different tree. Therefore, a random forest model randomly shuffles the data into training datasets consisting of a defined proportion of your samples and constructs a decision tree from that data. This process is repeated for a defined number of times. Following the Gaussian Process model, a Random forest model was used to indicate residues of the protein that differed between 'Warm' and 'Cold' environment proteins. Warm proteins were defined as those found above 18 ^{O}C and cold environment proteins were defined as those below 10 ^{O}C. For this, the VHSE encoding of biochemical properties was used. To counteract for noise, VHSE values were averaged across every subsequent five amino acids. The random forest model was built on 10,000 decision trees to minimise the false discovery rate with an additional training split of 90%. The additional hyperparameters

were as follows: measure of impurity='gini', max depth of tree=1, minimum impurity decrease=0.24, bootstrapping=TRUE and random state=69. The model was built using the Scikit Learn package (Version 1.2.2)

3.2.12 Identification of positively selected residues through PAML

Random forest model highlighted regions of proteins that differed in biochemical properties between Warm and Cold environments. In conjunction to this, PAML was used to test for positive selection across the proteins. Firstly, DNA sequences corresponding to protein sequences were extracted from the original scaffolds.

Easycodeml (Gao et al., 2019) built on PAML version 4.9 was used to test for positive selection on residues of form ID rbcL. Firstly, nested site models were used to test for positive selection of codons across all form IA/ID rbcL and rbcS sequences. M0 vs. M3, M1a vs. M2a, and M7 vs. M8 and M8a vs M8 were all used to assess selection. The highest weighting was given to the M8a vs M8 model for positive selection as this is regarded as the most rigorous comparison (Kapralov and Filatov, 2007).

3.2.13 Mixed effect model of variation (MEME) test for positively selected residues

MEME falls into the category of branch-site random effects phylogenetic models. Which means that ω (the ratio of dn/ds substitution events) is allowed to vary across lineages as well as sites. This is comparable to the nested branch site model in PAML; however, it does not require foreground lineages to be defined (Murrell et al., 2012). This is particularly useful for evolution events that are episodic (meaning that they often only occur in a single or a few lineages). As result, the effects of which would not be detected in PAML site based model. For this, rbcS and rbcL dna sequences collated from published metagenomes across multiple studies were first trimmed, removing partial sequences then aligned by codons using MUSCLE MSA in the MEGA11 software package with default parameters. Log ratio test (LRT) >1 and $p<0.05$ indicated residues that were positively selected for. MEME implements conservative detection methods for episodic selection , demonstrating a lower false discovery rate than comparable methods at ~3% false discovery for mediumly conserved gene sets (Murrell et al., 2012).

3.2.14 RELAX for testing relaxation in *rbcS* and *rbcL* genes

RELAX is a model used to test for the strength of selection exerted on defined phylogenetic lineages (Wertheim et al., 2015). Selection strength in the RELAX model is defined as either intensified selection where selection pressure is relatively higher or opposingly, relaxed selection. The latter often resulting in loss of gene functionality (Wertheim et al., 2015). The software works by fitting a model to defined test branches or 'H1' where the variable k is not constrained. This is then compared to the null model of the reference branches 'H0' where k is fixed to 1. LRT is then used to compare the two models with k<1 being defined as relaxed section and k>1 defining intensified selection (Wertheim et al., 2015). In this study, DNA sequences for form IA and ID *rbcL* and *rbcS* genes were aligned by MUSCLE codons. These were inserted into the RELAX program (Wertheim et al., 2015), implemented in Hyphy (Kosakovsky Pond et al., 2020) and supported on the datamonkey.org servers. The hypothesises tested are outlined in Figure 3.3 using incident branches as test branches as outlined. Firstly, for form IA *rbcL* and *rbcS* sequences, the two clades examined (Figure 3.3A and Figure 3.3B), were the divergent lineages pertaining to proteobacteria and cyanobacteria. Following this, each incident branch pertaining to an individual warm sequence or clade was selected as the test branches (Figure 3.3C), then temperate (Figure 3.3D) then cold (Figure 3.3E). Finally, the two basal branches of the cyanobacterial clade and proteobacterial clade were assessed for selection pressure; this was used as the negative control in this analysis. For form ID *rbcL* and *rbcs* genes the same tests were carried out, however the form ID Rubisco sequences did not appear to discrete clades like form IA sequences. Therefore, the two distinctly divergent clades were categorised as clade 1 and clade 2 and these sequences were compared for relaxation of selection (Appendix figure 1).

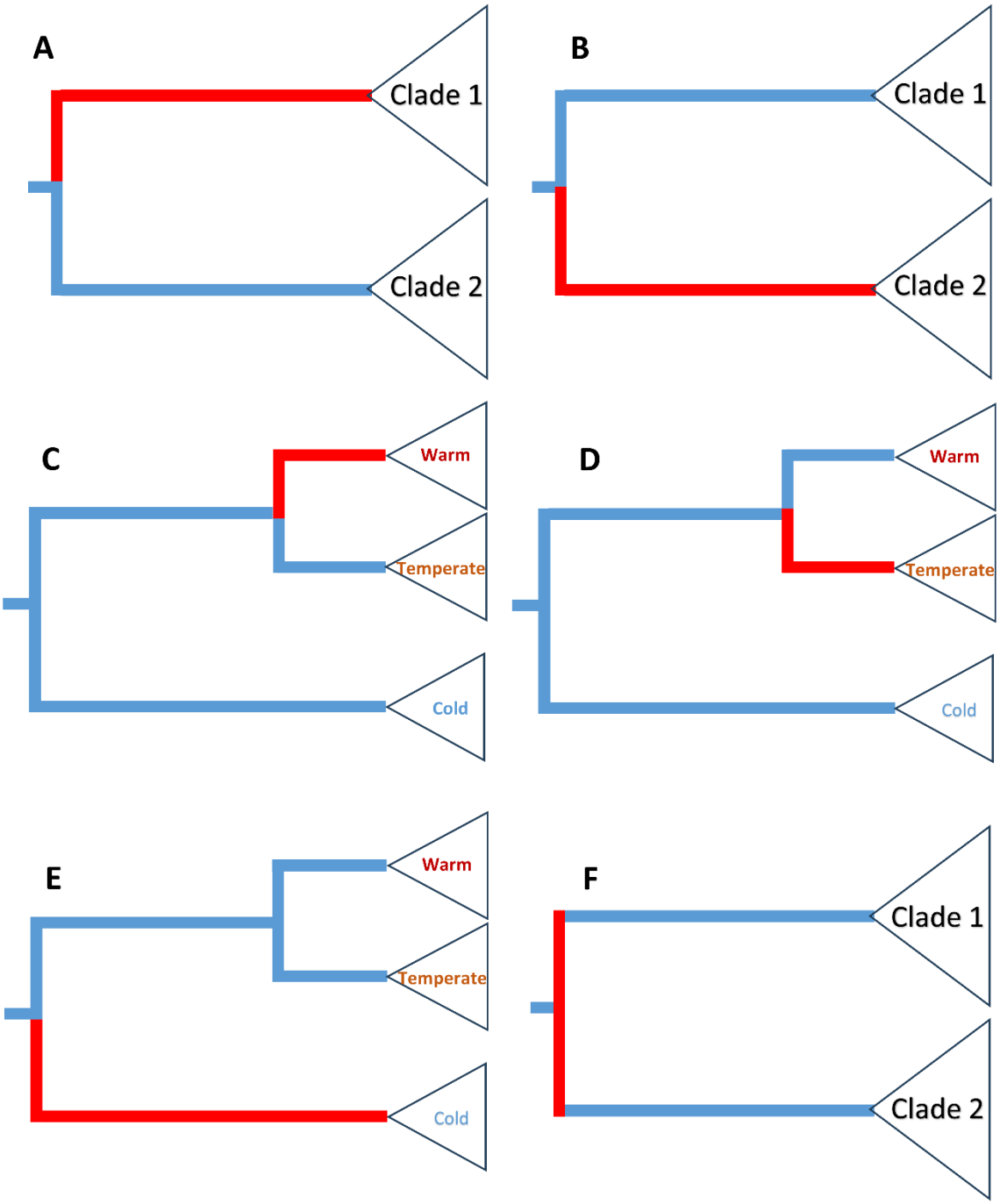


Figure 3.3- A simplified version of tests carried out by RELAX selection pressure on the *rbcL* and *rbcS* genes of form IA and ID organisms. Red lineages represent the test branches whereas the blue indicates the reference branches.

3.2.15 Close contact interactions between RbcS and RbcL subunits

Existing crystal structures for form IA and form ID Rubisco were used to highlight residues on the RbcL and RbcS subunits that were in close contact with each other. For form IA Rubisco, the crystal structure from *Halothiobacillus neapolitanus* (RCSB:7ZBT) (Ni et al., 2022) and *Thalassiosira antarctica* (RCSB:5MZ2) (Valegård et al., 2018b) were used to assess close contact interactions between a single RbcL unit and all interacting RbcS units from the holoenzyme structure. To achieve this the 'contact' function implemented in ChimeraX was utilised; this tool highlighted residues with overlapping Van der Waals (VD) radii of >-0.4 Å which were defined as close contact residues. Positively selected residues were deemed to also be close contact if they within two amino acids down or upstream of residues that matched the above criteria.

3.3 Results

3.3.1 Tara Oceans Rubisco large subunit species

Genes extracted from reference catalogues assembled from multiple studies on Tara Oceans raw genomic reads were collated into a phylogenetic tree. A maximum-likelihood model utilising a Dayhoff substitution matrix was used to achieve this phylogeny (Figure 3.4). It is clear that the individual Rubisco forms form clades relative to the structure of the large subunit (Figure 3.4). Form II Rubisco from proteobacteria are the most divergent Rubisco form representing an outgroup for the other Form I Rubisco types. There are no 'warm' form II Rubisco sequences found across these datasets (Figure 3.4).

There is a clear evolutionary divergence into 'red' and 'green' type Rubisco forms central to this phylogenetic tree; Form IA and IB group together in a discrete clade as do Form IC and ID Rubisco forms (Figure 3.4). Form IA is largely represented by warm and temperate cyanobacteria which represent a divergence from proteobacteria in this clade. This is also the case in the red type lineage with proteobacterial form IC sequences being distinct from the red algal from ID sequences (Figure 3.4). Form IB Rubisco sequences are the most poorly

represented sequences across all the Tara Ocean datasets examined in this study (Figure 3.4).

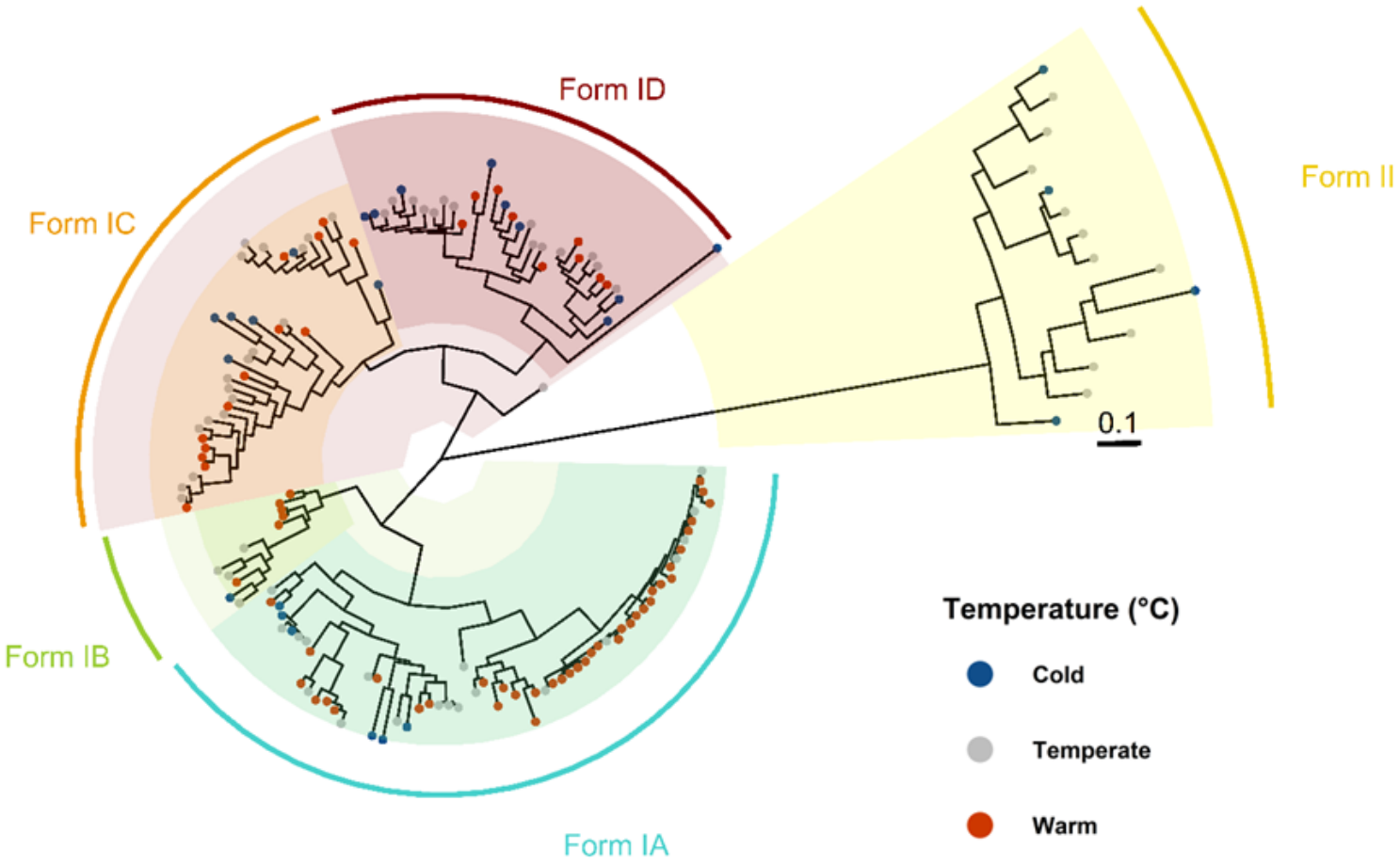


Figure 3.4- RbcL proteins from each individual species extracted from metagenomes (annotated with Diamond) were aligned using MUSCLE. The resulting alignment was used to construct a Maximum-likelihood tree using a Dayhoff substitution matrix for closely related sequences. Each sequence is grouped into its subsequent forms annotated by the colouring. The point at the end of each branch represents the average temperature the sequence was found at. The scale-bar highlights represents the number of substitutions per site.

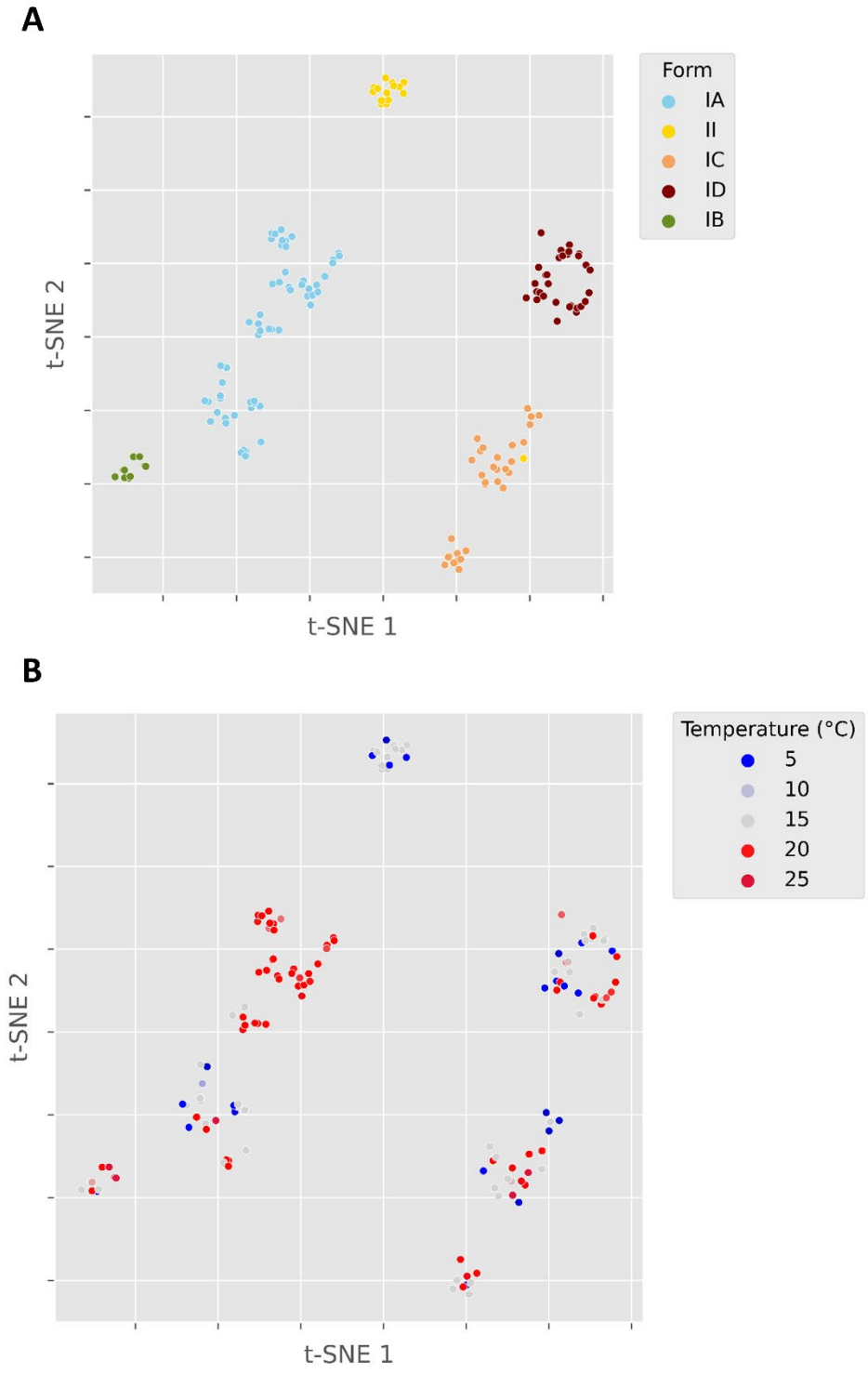


Figure 3.5- TSNE plots of the RbcL metagenome extracted proteins, aligned and binary encoded (perplexity=50, learning_rate=500, random_state=0, SciKit Learn Version 1.2.2): **A**

Colouration of sequences is used to represent the form of the individual RbcL sequences; **B** Colouration represents the average temperature the sequence was found at.

3.3.2 Dimensional reduction of RbcL sequence space

TSNE plots are a means of visually representing highly dimensional data in two dimensions to look for underlying patterns in the data. In this case TSNE plots were used to represent the individual RbcL sequences that were encoded to binary vectors representing individual amino acids (Figure 3.5). From the TSNE plots it is clear to see that the RbcL sequences group based on the form of the RbcL showing clear discrete groups for each; this is comparable to the phylogenetic representation above. When the temperature of the RbcL sequences were considered, there appears to be a subtle degree of grouping within Rubisco forms, however this distinction is imperfect at best (Figure 3.5).

3.3.3 Gaussian process regression model for predicting environmental temperature from RbcL sequence

A Gaussian process regression model RbcL was used to predict average environmental temperature from protein sequence. Simplistic and additive kernel architectures were compared; the 'simple kernel' utilising a single kernel function across the entire length of the protein, and the additive kernel which combines kernel functions for each individual amino acid for aligned RbcL sequences (Figure 3.6).

The model was tested for 'leave-one-out' cross validation. This is where the model is trained on all bar one of the RbcL sequences and then used to predict the temperature for the one sequence that was left out for training. This process is iteratively carried out for all RbcL sequences in the dataset. It clear to see that the additive kernel is far superior with an R^2 value of 0.80 between actual and predicted temperature values. The simple kernel performed poorly R^2 =0.19 (Figure 3.6). This highlights the importance of considering amino acids of each protein sequence individually as opposed to considering the protein sequence as whole.

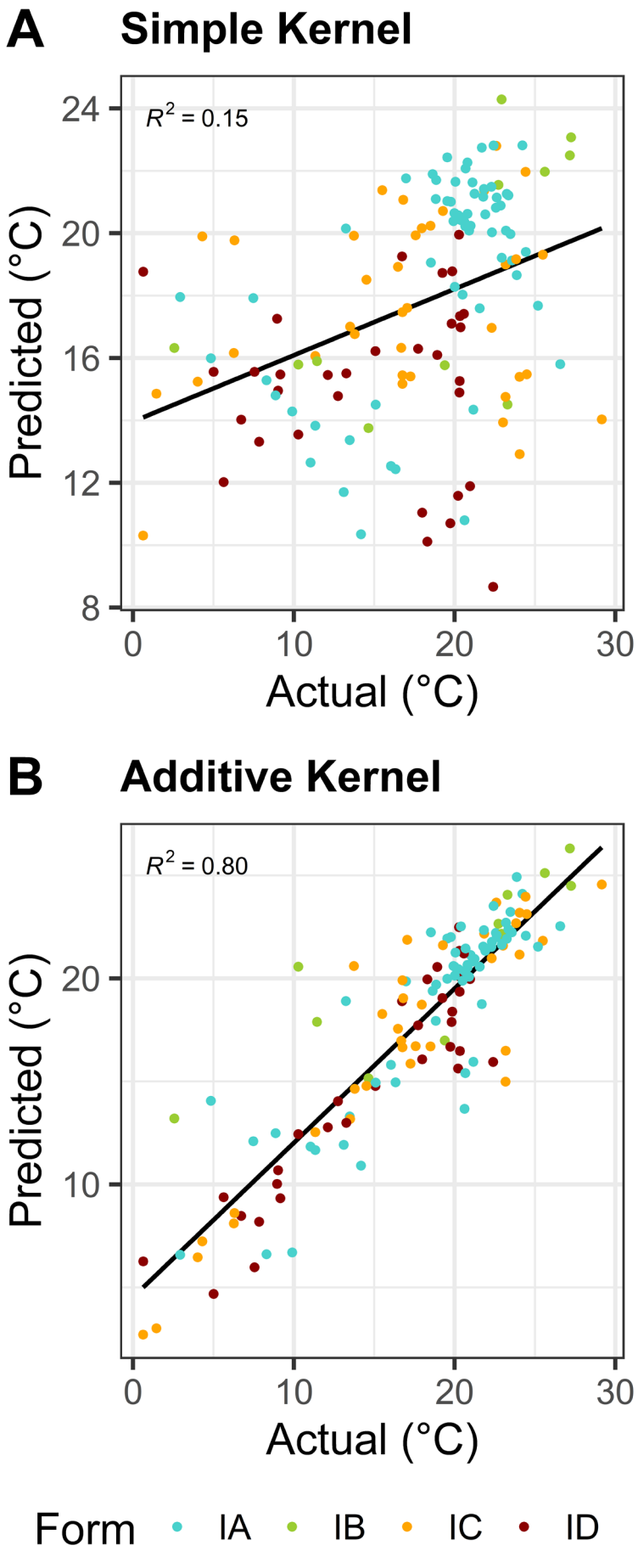

A
Simple Kernel
R^2 = 0.15
Predicted (°C)
Actual (°C)
B
Additive Kernel
R^2 = 0.80
Predicted (°C)
Actual (°C)
Form
IA
IB
IC
ID

Figure 3.6 – A comparison of simple and additive kernel architectures for the prediction of environmental temperature from protein sequence. For both architectures a Matern52 kernel function was used, variance=1 and length scale =1. The colour of the dots describes the Rubisco form. **A)** A simple kernel function uses a single kernel function to explain the relationship between X and Y. **B)** An additive kernel where each amino acid inputted is represented by a different kernel function, this allows weighting for more indicative parts.

For the Gaussian process model sequences must be represented in a numerical vector for regression calculations. How the protein sequences are represented is important as this can confer additional data to aid model prediction. In this study, RbcL sequences were represented in three ways. Firstly, binary representation was a simple encoding of amino acids to a five dimensional vector of 0 and 1s. Secondly, the protein sequences were converted to VHSE encoding which is an eight dimensional representation of amino acids based on their biochemical properties Xie et al. 2013**.** Finally, the ESM model was incorporated into this study which represents each protein sequence as a 1280 character long numerical vector which has been shown to be representative of the protein 3D structure. For the binary and VHSE encoding, a prior alignment was required. Each encoding method was compared for the three encoding methods in a leave one out cross validation as explained above.

Models built on all three protein encoding methods were shown to perform well when predicting environmental temperature from protein sequence. However, it is clear that model performance can be improved with incrementally more complex protein representations (Figure 3.7). The baseline model using one-hot binary encoding had an R^2 value of 0.80. This could be improved by representing the amino acids in a protein by their biochemical descriptors with VHSE encoding (R^2=0.83). Finally, the learnt encoding improved model performance again (R^2 = 0.88) (Figure 3.7).

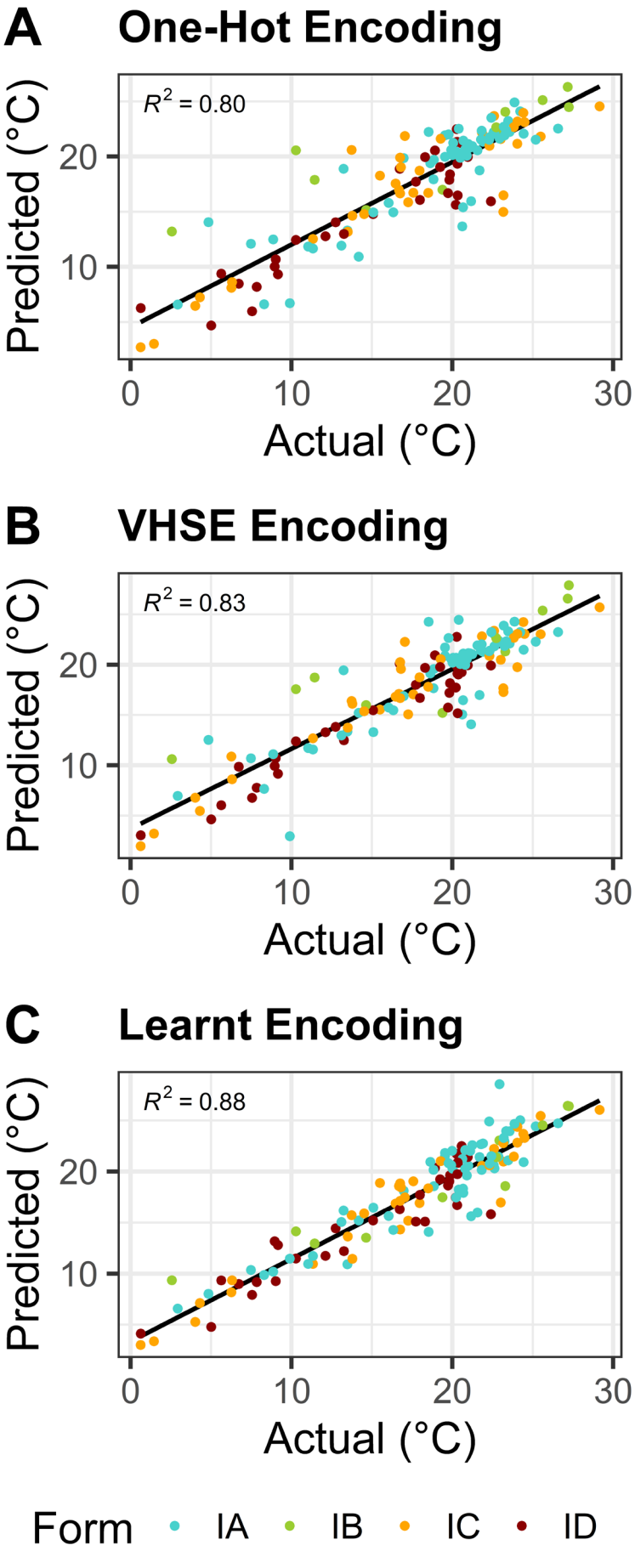
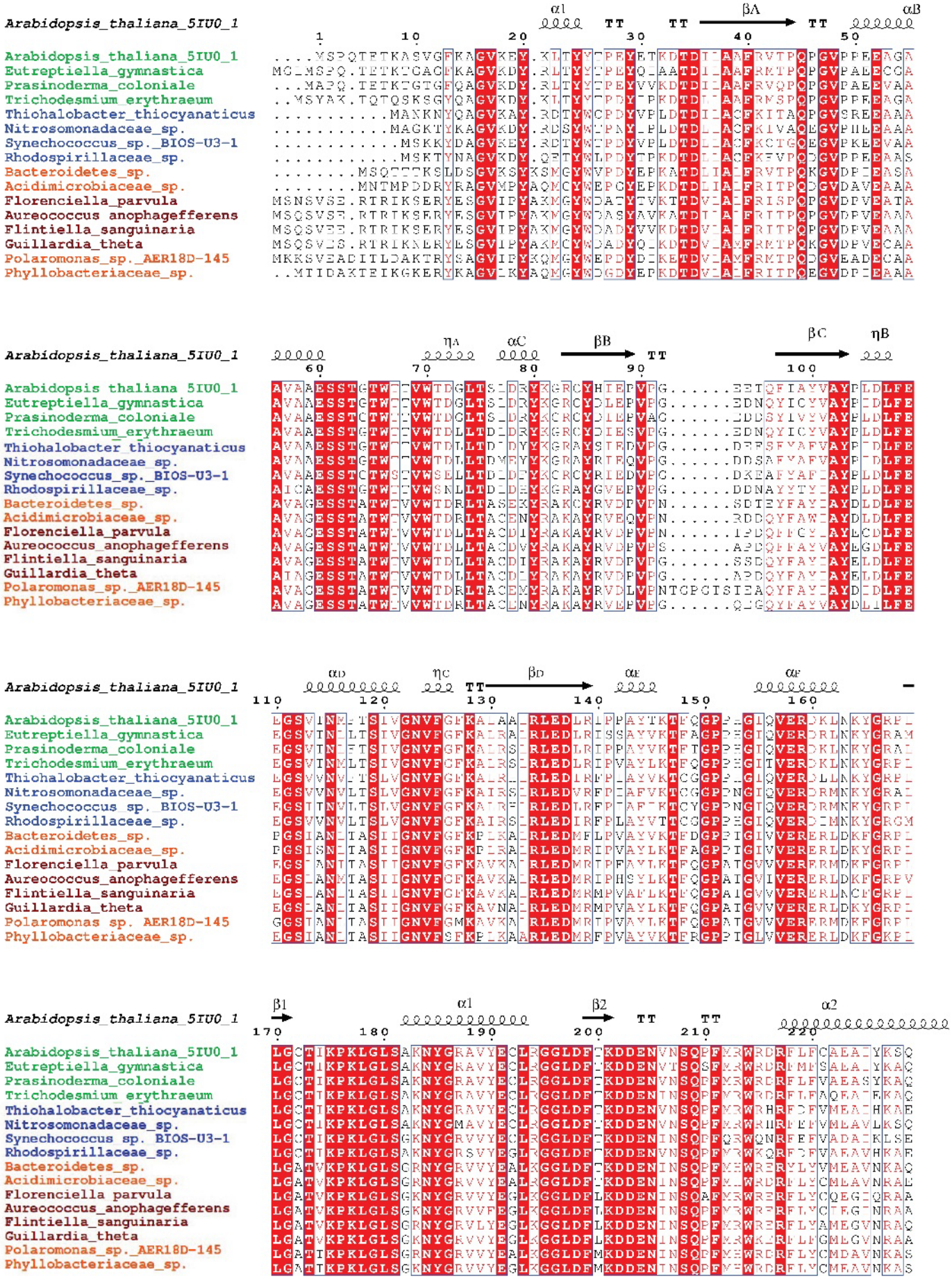

A
One-Hot Encoding
R^2 = 0.80
Predicted (°C)
Actual (°C)
B
VHSE Encoding
R^2 = 0.83
Predicted (°C)
Actual (°C)
C
Learnt Encoding
R^2 = 0.88
Predicted (°C)
Actual (°C)
0
10
20
30
Form
IA
IB
IC
ID

Figure 3.7- A comparison of protein encoding methods for the prediction of environmental temperature from protein sequence. For the different protein encoding methods the kernel function was kept constant using a Matern52 kernel function, variance=1 and length scale =1. The colour of the dots describes the Rubisco form. **A)** Uses a one-hot encoding scheme representing each amino acid as a binary encoding of five dimensions. **B)** VHSE encoding represents the protein by its biochemical properties in eight dimensions. Dimensions 1-2 represent hydrophobicity of the amino acids, 3-4 represent the steric properties of an amino acid, and 5-8 represent the electronic properties of the amino acid Xie et al. 2013**. C)** Learnt encoding is implemented from trained models used to represent protein secondary and tertiary structures Lin et al. 2022. This model represents proteins as numerical vector of 1280 elements in length.

RbcL forms extracted from Tara Oceans metagenomes aligned with known secondary structure from form IB RbcL 3.3.4


Arabidopsis_thaliana_5IU0_1
Arabidopsis_thaliana_5IU0_1
Eutreptiella_gymnastica
Prasinoderma_coloniale
Trichodesmium_erythraeum
Thiohalobacter_thiocyanaticus
Nitrosomonadaceae_sp.
Synechococcus_sp._BIOS-U3-1
Rhodospirillaceae_sp.
Bacteroidetes_sp.
Acidimicrobiaceae_sp.
Florenciella_parvula
Aureococcus_anophagefferens
Flintiella_sanguinaria
Guillardia_theta
Polaromonas_sp._AER18D-145
Phyllobacteriaceae_sp.
Arabidopsis_thaliana_5IU0_1
Eutreptiella_gymnastica
Prasinoderma_coloniale
Trichodesmium_erythraeum
Thiohalobacter_thiocyanaticus
Nitrosomonadaceae_sp.
Synechococcus_sp._BIOS-U3-1
Rhodospirillaceae_sp.
Bacteroidetes_sp.
Acidimicrobiaceae_sp.
Florenciella_parvula
Aureococcus_anophagefferens
Flintiella_sanguinaria
Guillardia_theta
Polaromonas_sp._AER18D-145
Phyllobacteriaceae_sp.
Arabidopsis_thaliana_5IU0_1
Arabidopsis_thaliana_5IU0_1
Eutreptiella_gymnastica
Prasinoderma_coloniale
Trichodesmium_erythraeum
Thiohalobacter_thiocyanaticus
Nitrosomonadaceae_sp.
Synechococcus_sp._BIOS-U3-1
Rhodospirillaceae_sp.
Bacteroidetes_sp.
Acidimicrobiaceae_sp.
Florenciella_parvula
Aureococcus_anophagefferens
Flintiella_sanguinaria
Guillardia_theta
Polaromonas_sp._AER18D-145
Phyllobacteriaceae_sp.

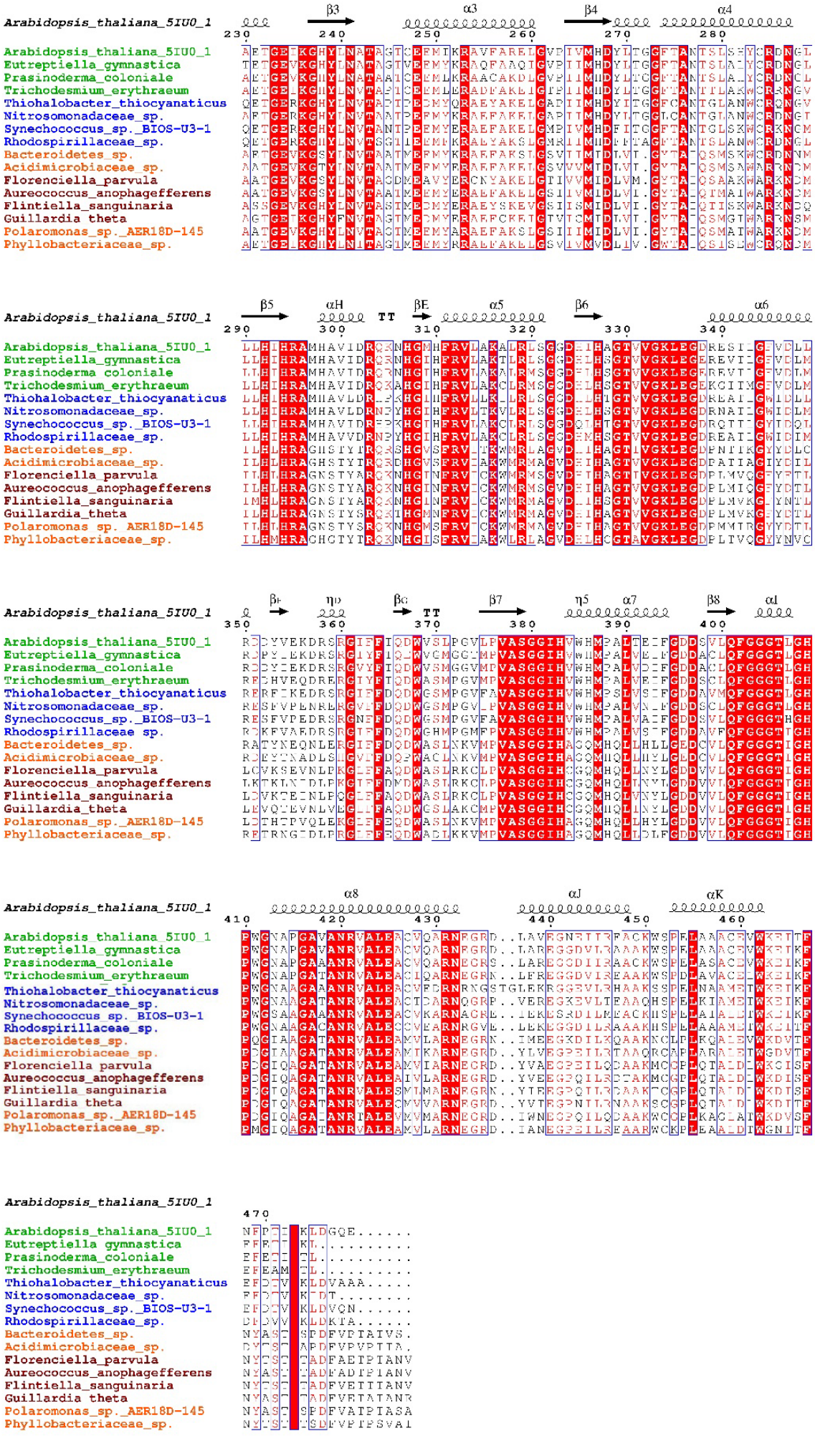

Arabidopsis_thaliana_5IU0_1
Arabidopsis_thaliana_5IU0_1
Eutreptiella_gymnastica
Prasinoderma_coloniale
Trichodesmium_erythraeum
Thiohalobacter_thiocyanaticus
Nitrosomonadaceae_sp.
Synechococcus_sp._BIOS-U3-1
Rhodospirillaceae_sp.
Bacteroidetes_sp.
Acidimicrobiaceae_sp.
Florenciella_parvula
Aureococcus_anophagefferens
Flintiella_sanguinaria
Guillardia_theta
Polaromonas_sp._AER18D-145
Phyllobacteriaceae_sp.

Figure 3.8- An alignment of RbcL protein sequences discovered in this study alongside the published crystal structure of *Arabidopsis thaliana* RCSB: 5IUO (Valegård et al. 2018). Representatives of each Rubisco form were chosen and indicated by their colouration. Consensus is indicated by red colouration of amino acids. Secondary structural units are annotated above relative to 5IUO.

Representative RbcL structures from the Tara Oceans dataset were aligned with the crystal structure for *Arabidopsis thaliana* Valegård et al. 2018. The multisequence alignment with secondary structures overlayed highlights the conserved nature of the RbcL protein across all form 1 types. An extension to the βB and βC loop in few form IC sequences and a short insertion of two amino acids at residue 437 in a form IA protein sequence represent the only structural differences between sequences (Figure 3.8).

3.3.5 Random forest classifier to divide Warm and Cold sequences RbcL

A random forest classifier model was built to highlight residues that differed significantly between 'Warm' and 'Cold' RbcL species. VHSE values were averaged along the length of the protein every five amino acids, breaking the protein into parts represented by its average biochemical properties.

For this, form IA and ID rbcl sequences were used being the most abundant sequences in the seas and oceans (Figure 2.6). Within form IA sequences there were a total of 21 Warm and eight Cold RbcL sequences. 10,000 iterations of the decision tree classifier with 90/10 bootstrapping highlighted that the hydrophobic and steric properties of residues between 89-93 (indexed to Figure 3.8) separated Warm and Cold sequences (Figure 3.8). With no bootstrapping, the residues of 89-93 categorised the form IA RbcL dataset into eight Cold with 39.7% impurity and 18 Warm with 0% impurity.

Amongst the dataset there were 14 Cold and 13 Warm form ID sequences. Residues between 347-351 could be used to divide the dataset based on the hydrophobic properties of the RbcL proteins. Without bootstrapping differences in hydrophobic properties of these residues divide the dataset into 14 Cold with 34.6% impurities and 9 Warm with 0% impurity (Figure 3.9).

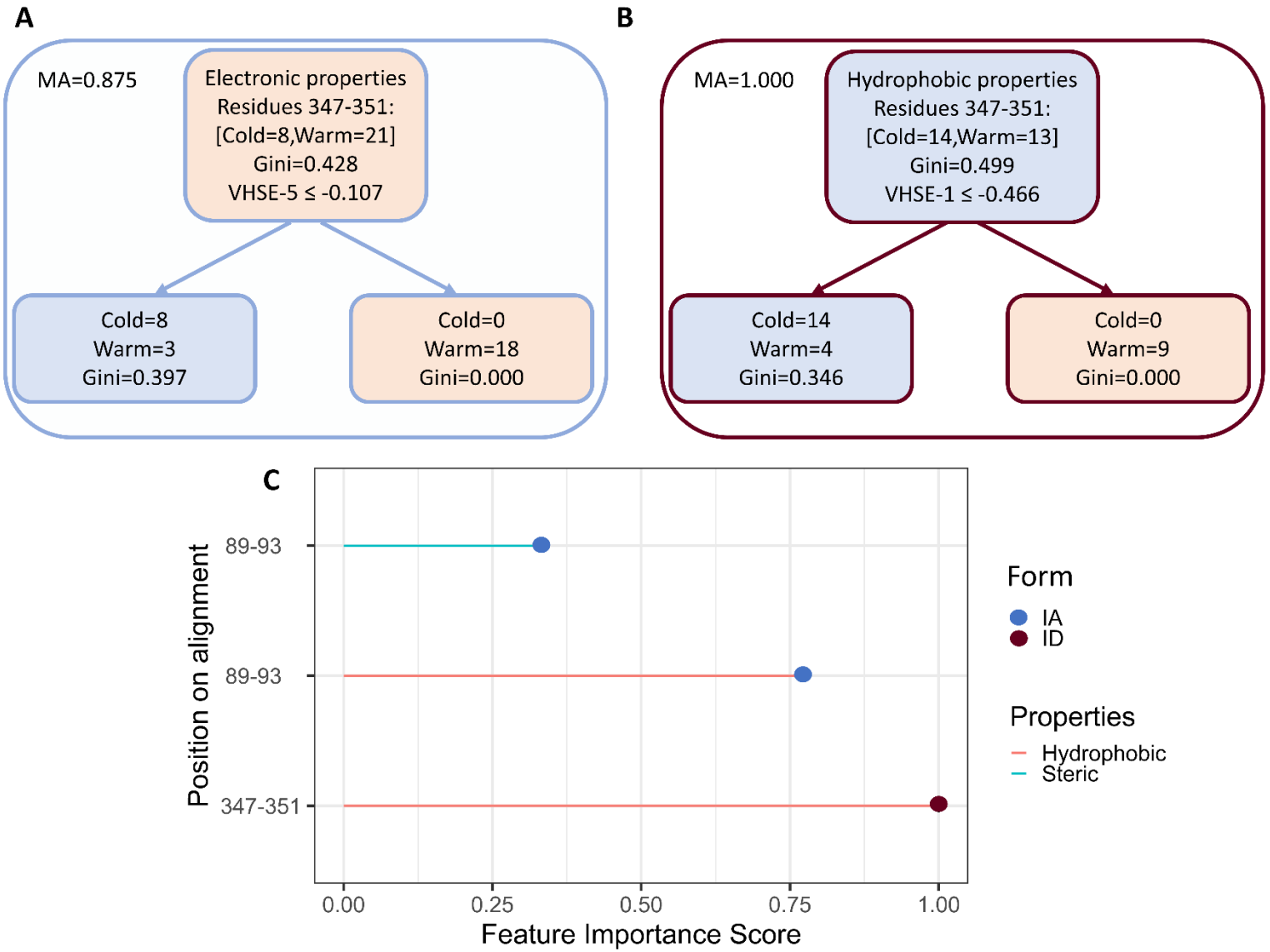

A
MA=0.875
Electronic properties
Residues 347-351:
[Cold=8,Warm=21]
Gini=0.428
VHSE-5 ≤ -0.107
Cold=8
Warm=3
Gini=0.397
Cold=0
Warm=18
Gini=0.000
B
MA=1.000
Hydrophobic properties
Residues 347-351:
[Cold=14,Warm=13]
Gini=0.499
VHSE-1 ≤ -0.466
Cold=14
Warm=4
Gini=0.346
Cold=0
Warm=9
Gini=0.000
C
Position on alignment
89-93
89-93
347-351
0.00
0.25
0.50
0.75
1.00
Feature Importance Score
Form
IA
ID
Properties
Hydrophobic
Steric

Figure 3.9- Form IA and ID RbcL protein sequences were aligned and converted to VHSE encoding, representative of their biochemical properties. These biochemical encodings were averaged every five amino acids along the length of the protein giving average VHSE values for each five amino acid part. These average values for each part were used to divide protein sequences into Cold and Warm categories. Residues listed in the figure are indexed with residues in Figure 3.7, gini is the measure of impurity. **A)** A single decision tree with no bootstrapping, used to divide form IA RbcL sequences into Cold and Warm sequences. The biochemical properties used to divide the sequence are listed as well as the residue location. **B)** A decision tree with no bootstrapping, used to divide form ID RbcL sequences into Cold and Warm sequences. **C)** A random-forest model was implemented to find the most important protein features, indicative of warm and cold environments form IA RbcL and ID RbcL. This involved the construction of 10,000 decision trees with boot-strapping; the most occurring features capable of dividing Warm and Cold sequences are shown. The points of the graph represent the RbcL form, the stick represents the biochemical property that best categorises that form. The Feature importance score represents the proportion of times a certain feature was selected to best categorise the Warm and Cold sequences after bootstrapping over the 10,000 iterations. For the random forest models, the hyperparameters used were: training split=90%, max depth of tree=1, minimum gini impurity decrease=0.24, bootstrapping=TRUE and random state=69. The model was built using the Scikit Learn package (Version 1.2.2).

3.3.6 Test for positively selected sites across the *rbcL* gene

Table 3.1- Form IA *rbcL* test for positively selected sites comparing LRTs of nested models

Model	No. S	Log-likelihood	Model compared	LRT p-value	Positive sites
M3	48	-26593.30			
M0	48	-27453.09	M0 vs. M3	0.00	NA
M2a	48	-27320.88			
M1a	48	-27320.88	M1a vs. M2a	1.00	NA
M8	48	-26544.18			
M7	48	-26544.18	M7 vs.M8	1.00	NA
M8a	48	-26538.39	M8a vs.M8	0.00	NA

Codeml nested models were used to test for positive selection across form IA *rbcL* genes derived from all marine environments found in this study. Principally, the nested models of M0 vs M3 indicated heterogeneity in selection pressures across the rbcL genes (Log_L =--26593.30, -27453.09, $p < 0.0001$, df = 8) (Table 3.1). The LRT models of M7-M8 and M8-M8A were used to assess positive selection and residues across the IA rbcL genes, with M8-8A being the more rigorous test for positive selection. The M8a vs M8 LRT comparison was found to be significant ($p<0.001$), providing evidence for positive selection across the IA *rbcL* genes considered here. Despite this significance, no single loci exceeded the threshold for positive selection (Bayesian Empirical Bayes < 0.95); therefore, the positive selection signal is weak. Additionally the second model used for highlighting positive selection across the *rbcL* gene M7 vs M8 was not significant (Log_L =-26544.18, -26544.18, p =1.00) (Table 3.1).

Alternatively, when episodic positive selection was assessed using MEME selection there was extensive positive selection across the form IA *rbcL* gene with 26 residues significantly positively selected for in *rbcL* phylogenetic lineages (LRT>1, $p<0.05$) (Table 3.2). The positively selected residues were overlayed with close contact residues between RbcL and RbcS determined from the IA crystal structure**.** This highlighted 11 residues on the form IA RbcL, which were positively selected for and in close contact with RbcS protein in the holoenzyme (Appendix Table 1) (Table 3.11) (Figure 3.16).

Table 3.2- Test for Episodic selection amongst form IA *rbcL* residues. Residues with significant selection pressure are shown.

Residue	LRT	LRT p-value	Feature
32	7.78	0.01	α1/βA
42	20.75	0	βA
75	5.28	0.03	ηA/αC
132	10.2	0	βD
138	5.18	0.03	βD
149	36.35	0	αE/αF
156	53.68	0	αF
161	6.42	0.02	αF
172	7.25	0.01	β1/α1
181	6.35	0.02	β1/α1
221	6.06	0.02	α2
226	9.3	0	α2
227	18.69	0	α2
230	10.91	0	α2
235	23.89	0	α2/β3
242	16.99	0	β3/α3
245	6.99	0.01	β3/α3
246	8.78	0.01	β3/α3
267	6.23	0.02	β4
281	8.19	0.01	α4
355	13.63	0	βF/ηD
368	9.29	0	βG/β7
392	12.15	0	α7
439	5.77	0.03	αJ
450	6.3	0.02	αJ/αK
452	6.08	0.02	C-terminal

Table 3.3- Form ID *rbcL* test for positively selected sites comparing LRTs of nested models.

Model	No. S	Log-likelihood	Model compared	LRT p-value	Positive sites
M3	65	-11810.35			NA
M0	65	-12064.47	M0 vs. M3	0.00	
M2a	65	-12058.21			NA
M1a	65	-12058.21	M1a vs. M2a	1.00	
M8	65	-11817.76			NA
M7	65	-11817.76	M7 vs.M8	1.00	
M8a	65	-11817.76	M8a vs.M8	1.00	NA

Codeml nested models were used to test for positive selection across form ID *rbcL* genes derived from all marine environments found in this study. Principally, the nested models of M0 vs M3 indicated heterogeneity in selection pressures across the rbcL genes (Log_L =--11810.35, -12064.47, $p < 0.0001$, df = 8) (Table 3.3). The LRT models of M7-M8 and M8-M8A were used to assess positive selection and residues across the ID *rbcL* genes showing no evidence to suggest positively selected residues as the LRT models were found to not differ significantly (Log_L =-11817.76, -11817.76, p =1.00, df = 2 and Log_L =-11817.76, -11817.76, p =1.00, df = 1 respectively) (Table 3.3).

Again when episodic positive selection was assessed using MEME selection there was widespread positive selection spread across the form ID *rbcL* gene with 25 residues significantly positively selected in phylogenetic lineages (LRT>1, $p<0.05$) (Table 3.4). The positively selected residues were overlayed with close contact residues between RbcL and RbcS determined from the ID crystal structure of *Thalassiosira antarctica* (RCSB:5MZ2) **Valegård et al. 2018.** This highlighted 13 residues on the form ID RbcL, which were positively selected for and in close contact with RbcS protein in the holoenzyme (Appendix 5.1) (Table 3.11) (Figure 3.16).

Table 3.4 – Test for episodic selection amongst form ID *rbcL* residues. Residues with significant selection pressure are shown.

Residue	LRT	LRT p-value	Feature
28	4.72	0.04	α1 /βα gαp
53	4.78	0.04	αβ
99	12.98	0	βC
105	15.52	0	ηβ
153	5.84	0.02	αE/αF
167	12.36	0	αF/β1
183	4.76	0.04	α1
221	6.57	0.02	α2
226	19.75	0	α2
238	7.62	0.01	β3
282	7.02	0.01	α4
291	12.98	0	β5
347	6.47	0.02	α6
350	11.17	0	α6
351	13.71	0	α6/βF
360	6.8	0.01	ηD
372	7.25	0.01	βG/β7
373	6.23	0.02	βG/β7
384	16.03	0	η5
398	14.23	0	α7/β8
431	11.07	0	α8
433	13.46	0	α8/αJ
435	21.39	0	α8/αJ
455	9.16	0	αK
470	9.59	0	C-terminal

3.3.7 Test for relaxation of selection in *rbcL* gene

Table 3.5 -Test for Relaxation or Intensification of Selection Pressure across defined phylogenetic lineages across form IA and ID *rbcL*.

	Test Branches	K	p	LR	Selection
IA	Proteobacteria	0.67	0.00	29.41	Relaxation
	Cyanobacteria	1.68	0.02	23.25	Intensification
	Warm	1.10	0.32	7.05	NS
	Temperate	0.83	0.32	11.46	NS
	Cold	0.57	0.03	19.03	Relaxation
	Basal Branches	4.07	0.09	27.57	NS
ID	Clade 1	0.61	0.42	2.62	NS
	Clade 2	0.72	0.36	1.79	NS
	Warm	1.00	0.64	0.39	NS
	Temperate	0.83	0.39	1.56	NS
	Cold	5.43	0.34	1.72	NS
	Basal Branches	0.74	0.43	1.44	NS

RELAX test was used to assess selection pressures on *rbcL* genes found at different temperatures and across phylogenetic groupings. Each test group was examined for selection pressure five times. Genes were grouped into Warm, Temperate and Cold groups based on the average temperature they were found at. Phylogeny was assessed from Diamond annotation of genes. Selection pressure was assessed for the incident branches, meaning that branches that diverged from reference clade to test clade were measured for selection pressure (Table 3.5).

For forms IA *rbcL* there is a relaxation of selection in genes from proteobacterial species and an intensification of selection in cyanobacterial genes ($K=0.67$, $p<0.001$ and $K=1.68$, $p=0.02$ respectively). Additionally, there was also a relaxation of selection pressure observed in cold form IA *rbcL* genes ($K=0.57$, $p=0.03$) (Table 3.5).

On the other hand there was no significant relaxation or intensification of selection pressure found in form ID *rbcL* sequences for both the phylogenetic groupings and temperature groupings ($p>0.05$ for all test groups). For both form IA and ID sequences, the basal branches used as the negative control in this study were not under significant selection pressure ($p=0.09$, $p=0.43$) respectively (Table 3.5).

3.3.8 Phylogeny of RbcS species extracted from Tara Oceans metagenomes

A maximum-likelihood model utilising a Dayhoff substitution matrix was used to observe the phylogeny of RbcS protein sequences found in Tara Oceans assemblies. RbcS protein sequences group by form.

There is a clear evolutionary divergence into 'red' and 'green' type Rubisco forms central to this phylogenetic tree; Form IA and IB group together in a discrete clade as do Form IC and ID RbcS forms, although form IC RbcS is more closely related to form ID than form IA is to IB.

The form IA clade can be further divided into proteobacteria and cyanobacteria. There are no cyanobacteria that exist at average temperatures less than 10 ^{o}C. No form IB rbcS sequences from cold environments were found either. Within form IC and ID rbcS sequences there is a consistent mix of Cold, Temperate and Warm sequences (Figure 3.10).

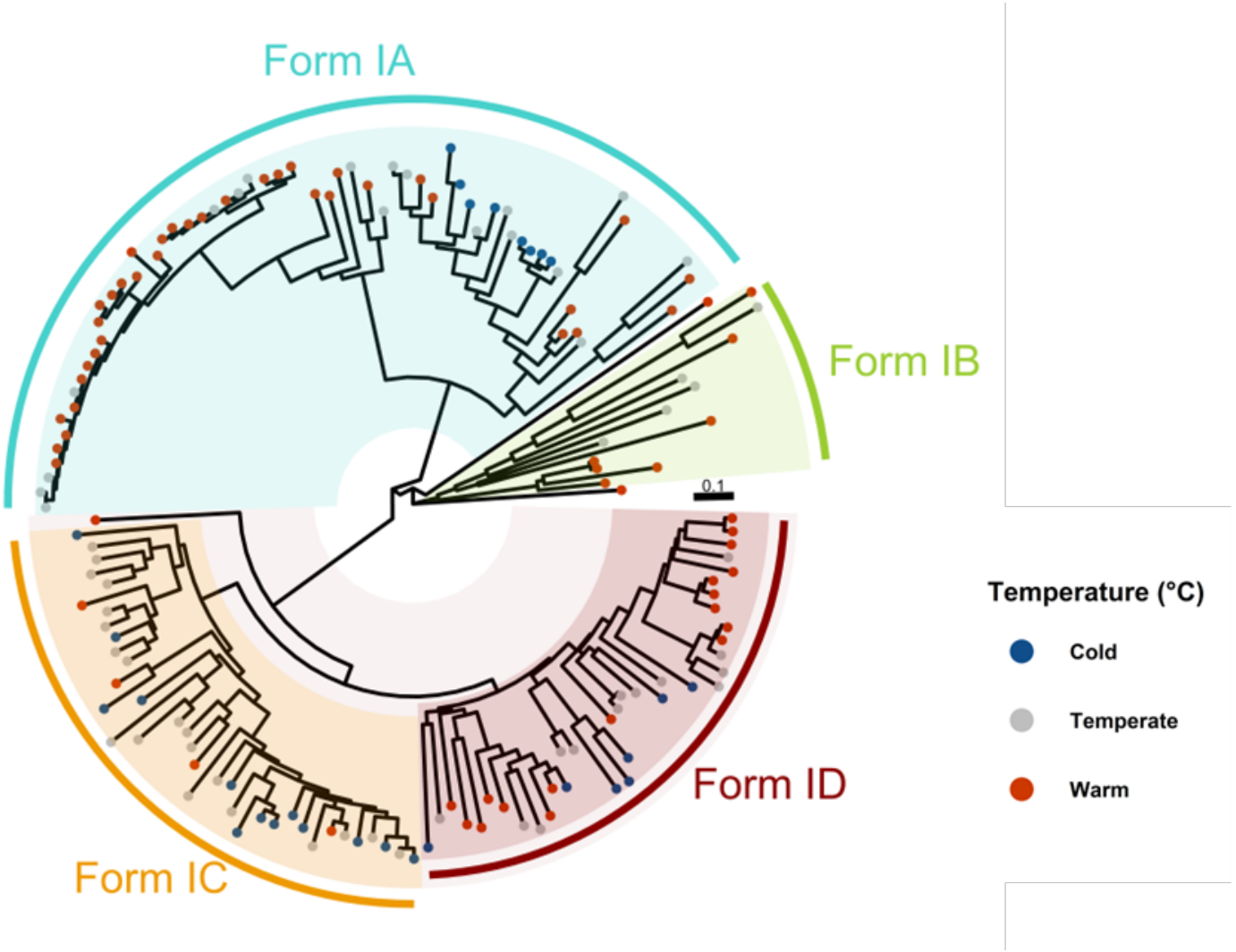


Figure 3.10- RbcS proteins from each individual species extracted from metagenomes (annotated with Diamond) were aligned using MUSCLE. The resulting alignment was used to construct a Maximum-likelihood tree using a Dayhoff substitution matrix for closely related sequences. Each sequence is grouped into its subsequent forms annotated by the colouring. The point at the end of each branch represents the average temperature the sequence was found at. The scale-bar highlights represents the number of substitutions per site.

3.3.9 Dimensional reduction of RbcS sequence space

TSNE, dimensionality reduction plots were used to represent the individual RbcS sequences encoded as binary vectors in 2D. From the TSNE plots it is clear to see that the RbcS sequences group based on form, showing clear discrete groups for each. When the temperature of the RbcS sequences were considered there appears to be a subtle degree of grouping within Rubisco forms; however, this distinction is imperfect at best (Figure 3.11).

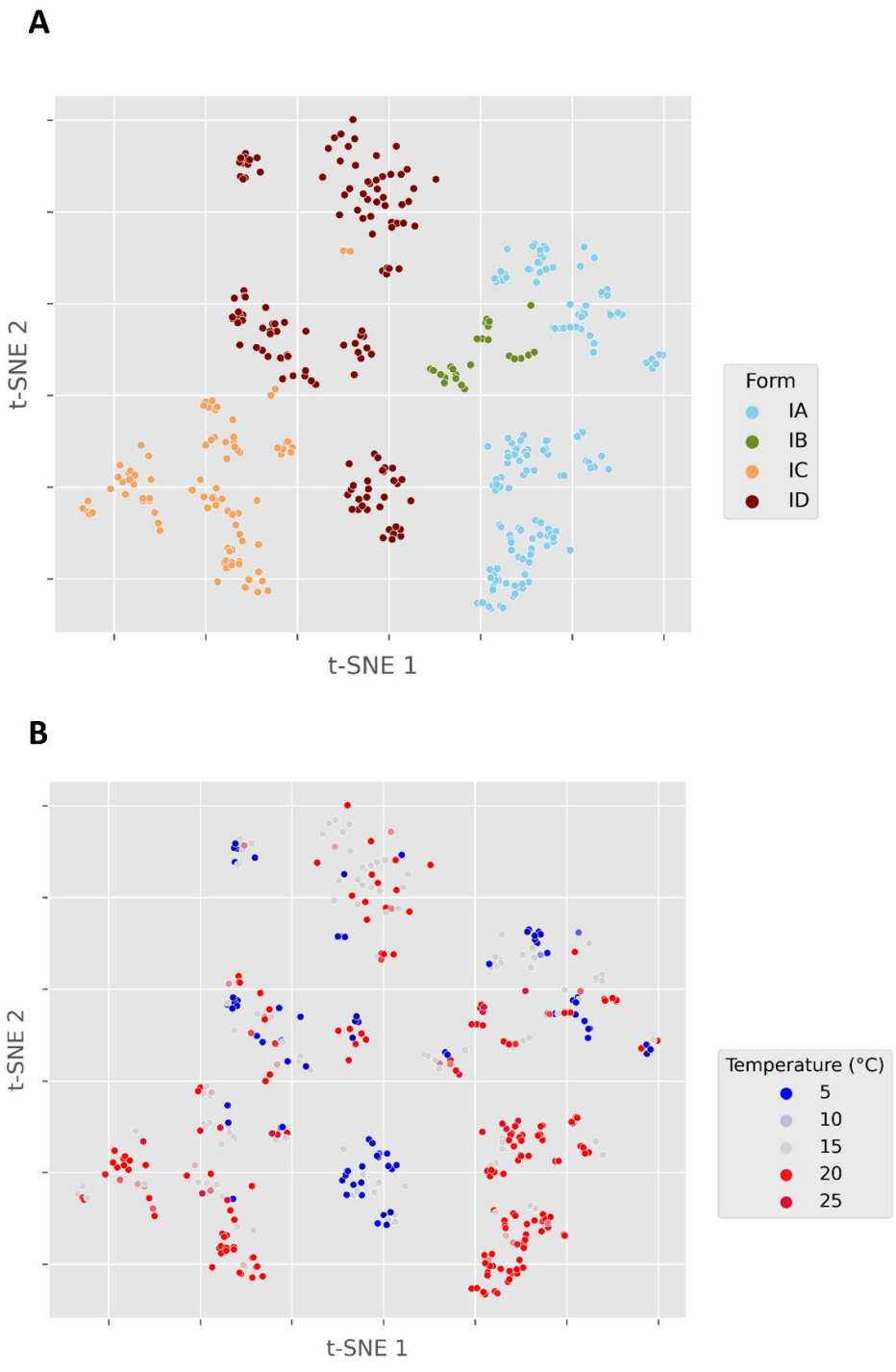


Figure 3.11- TSNE plots of the RbcS metagenome extracted proteins, aligned and binary encoded (perplexity=50, learning_rate=500, random_state=0, SciKit Learn Version 1.2.2) **A** Colouration of sequences is used to represent the form of the individual RbcS sequences. **B** Colouration represents the average temperature the sequence was found at.

Predicting environmental temperature from RbcS sequence through Gaussian Process modelling **3.3.10**

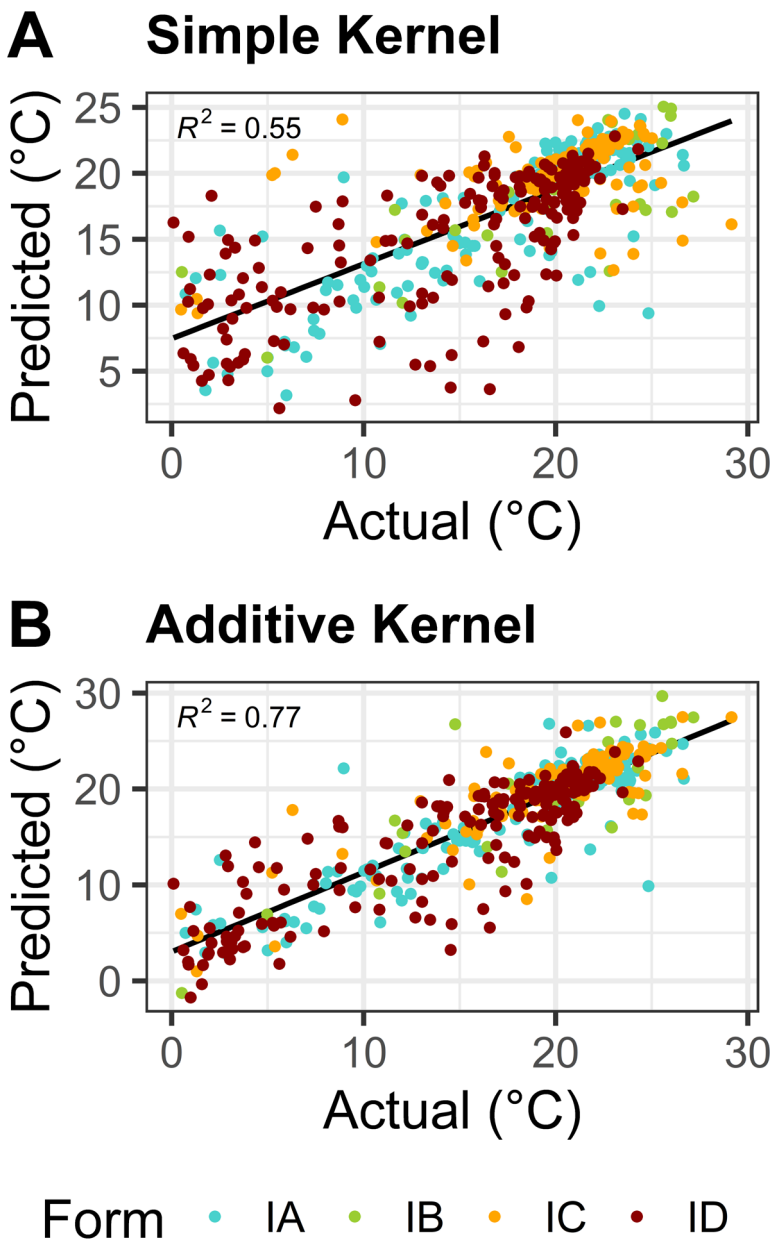


Figure 3.12 – A comparison of simple and additive kernel architectures for the prediction of environmental temperature from RbcS protein sequence. For both architectures a Matern52 kernel function was used, variance=1 and length scale=1. The colour of the dots describes the Rubisco Form. **A)** A simple kernel function, uses a single kernel function to explain the relationship between X and Y. **B)** An additive kernel, where each amino acid inputted is represented by a different kernel function; this allows weighting for more indicative parts.

GPR was used to predict average environmental temperature from RbcS protein sequence. Simplistic and additive kernel architectures were compared using leave one out cross validation. Like for RbcL sequences, the additive kernel is superior at predicting environmental temperature from RbcS sequence with predicted and actual regression calculated at R^2 =0.77. The simple kernel performed poorly with R^2 =0.55 (Figure 3.12).

The protein encoding methods of one-hot, binary and learnt were compared for GPR model performance when RbcS sequences were considered. Models built on all three protein encoding methods were shown to perform well when predicting environmental temperature from protein sequence. Performance increased marginally with encoding complexity. The model built using one-hot binary encoding had an R^2 value of 0.77 when predicted and actual environmental temperatures were compared (Figure 3.13). This relationship was improved by representing the amino acids by their biochemical descriptors with VHSE encoding (R^2=0.78) (Figure 3.13). Finally, the learnt encoding further improved model performance (R^2 = 0.79). Overall, each model performed worse for RbcS than it did for RbcL (Figure 3.13).

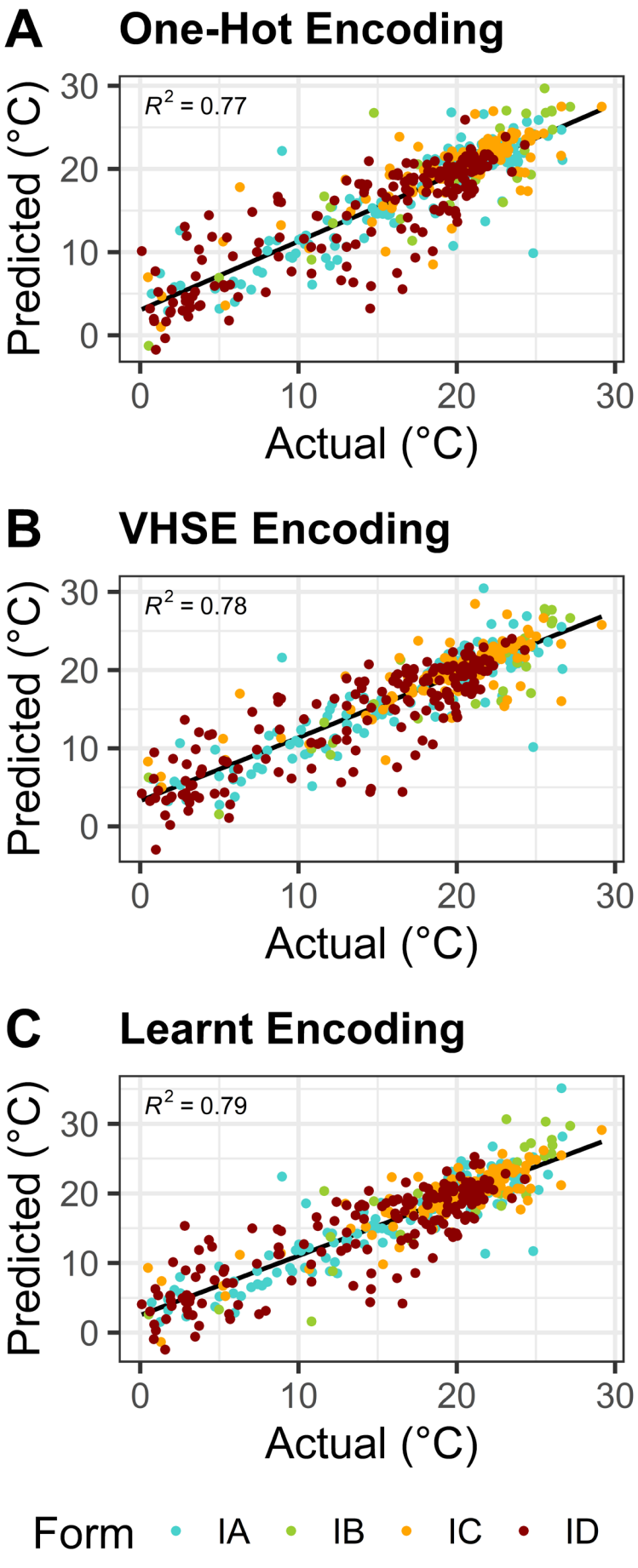

A
One-Hot Encoding
R^2 = 0.77
Predicted (°C)
Actual (°C)
B
VHSE Encoding
R^2 = 0.78
Predicted (°C)
Actual (°C)
C
Learnt Encoding
R^2 = 0.79
Predicted (°C)
Actual (°C)
0
10
20
30
Form
IA
IB
IC
ID

Figure 3.13- A comparison of protein encoding methods for the prediction of environmental temperature from RbcS protein sequence. For the different protein encoding methods the kernel function was kept constant using a Matern52 kernel function, variance=1 and length scale=1. The colour of the dots describes the Rubisco Form. **A)** Uses a one-hot encoding scheme representing each amino acid as a binary encoding of five dimensions. **B)** VHSE encoding represents the protein by its biochemical properties in eight dimensions. Dimensions 1-2 represent hydrophobicity of the amino acids, 3-4 represent the steric properties of an amino acid, and 5-8 represent the electronic properties of the amino acid Xie et al. 2013. **C)** Learnt encoding is implemented from trained models used to represent protein secondary and tertiary structures Lin et al. 2022**.** This model represents proteins as numerical vector of 1280 elements in length.

3.3.11 RbcS forms extracted from Tara Oceans metagenomes aligned with known secondary structure from form IB RbcS

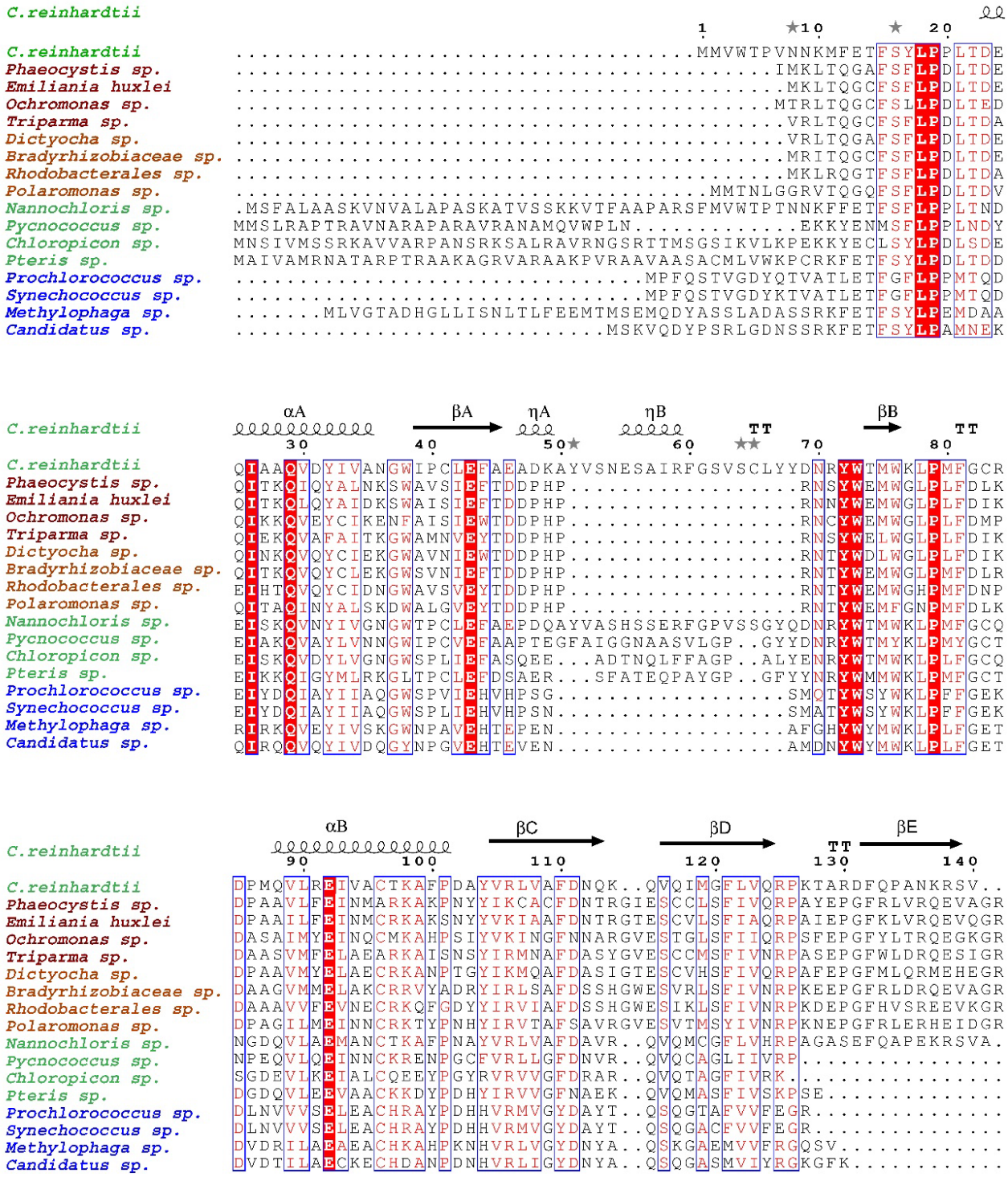

C.reinhardtii
Phaeocystis sp.
Emiliania huxlei
Ochromonas sp.
Triparma sp.
Dictyocha sp.
Bradyrhizobiaceae sp.
Rhodobacterales sp.
Polaromonas sp.
Nannochloris sp.
Pycnococcus sp.
Chloropicon sp.
Pteris sp.
Prochlorococcus sp.
Synechococcus sp.
Methylophaga sp.
Candidatus sp.
αA
βA
ηA
ηB
TT
βB
αB
βC
βD
βE
βF
αC

Figure 3.14- An alignment of RbcS protein sequences discovered in this study alongside the published crystal structure of *Chlamydomonas reinhardtii* RCSB: 1GK8 (Taylor et al. 2001). Representatives of each Rubisco form were chosen and indicated by their colouration. Consensus is indicated by red colouration of amino acids. Secondary structural units are annotated above relative to 5IUO.

Representative RbcS structures from the Tara Oceans dataset were aligned with the crystal structure for *Chlamydmonas reinhardtii* (Taylor et al. 2001). The multisequence alignment with secondary structures overlayed highlights notable differences in the RbcS structures between forms. Firstly, the N-terminal of the RbcS in form IB organisms and the form IA *Methylophaga* sp. is extended by ~30 amino acids. There is also an extension between the βA and βB loops in form IB organisms which is not observed in other forms. Opposingly, in form IC and ID organisms there is an extended C-terminus on the RbcS protein. This extension has additional secondary structures of a βE and βF loop. Unlike the highly conserved RbcL proteins observed in Figure 3.14, there is more variation observed in the secondary structure and residue conservation in the RbcS proteins.

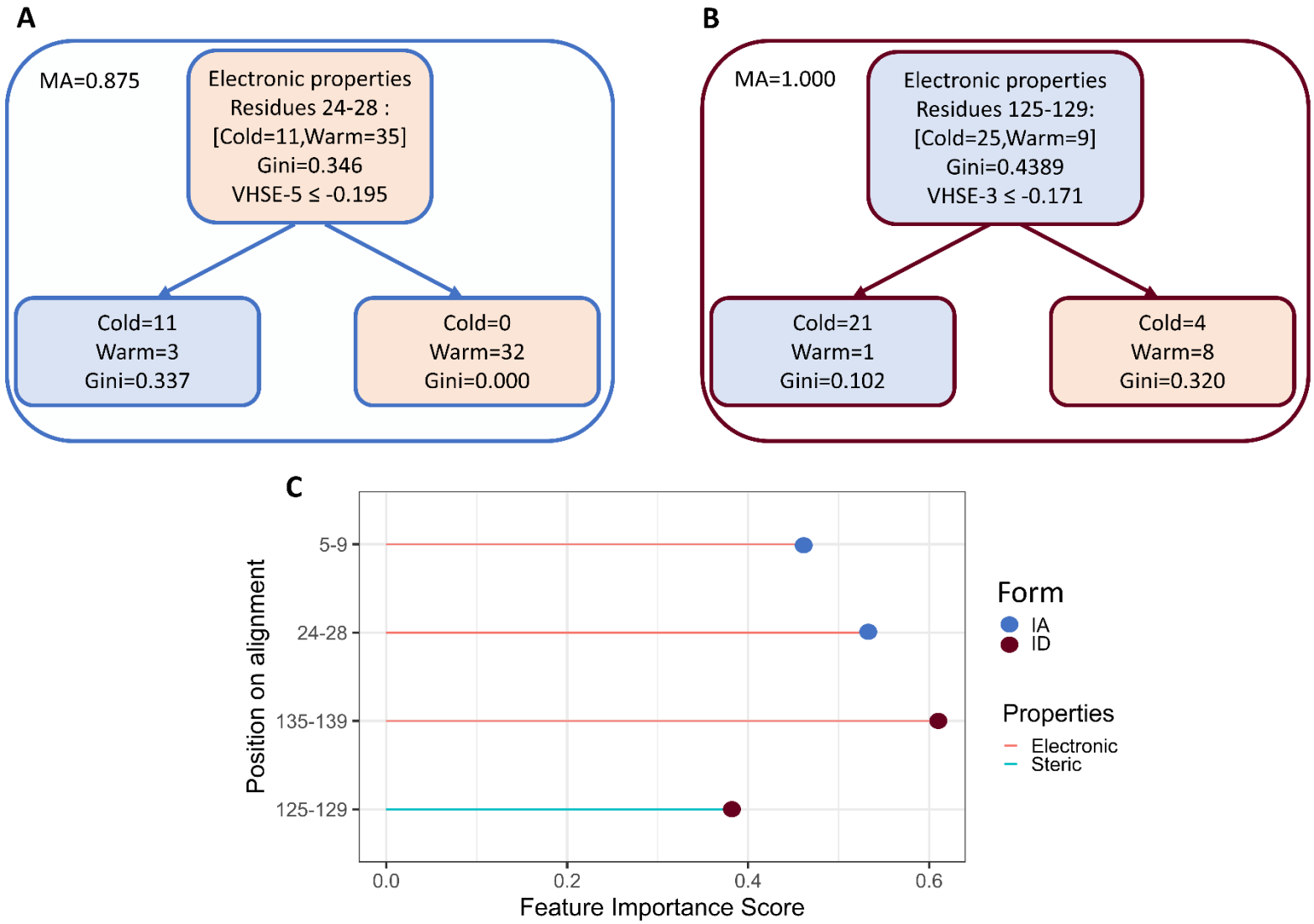

A
MA=0.875
Electronic properties
Residues 24-28 :
[Cold=11,Warm=35]
Gini=0.346
VHSE-5 ≤ -0.195
Cold=11
Warm=3
Gini=0.337
Cold=0
Warm=32
Gini=0.000
B
MA=1.000
Electronic properties
Residues 125-129:
[Cold=25,Warm=9]
Gini=0.4389
VHSE-3 ≤ -0.171
Cold=21
Warm=1
Gini=0.102
Cold=4
Warm=8
Gini=0.320
C
Position on alignment
5-9
24-28
135-139
125-129
0.0
0.2
0.4
0.6
Feature Importance Score
Form
IA
ID
Properties
Electronic
Steric

Figure 3.15- Form IA and ID RbcS protein sequences were aligned and converted to VHSE encoding, representative of their biochemical properties. These biochemical encodings were averaged every five amino acids along the length of the protein giving average VHSE values for each five amino acid part. These average values for each part were used to divide protein sequences into Cold and Warm categories. Residues listed in the figure are indexed with residues in Figure 3.7, gini is the measure of impurity. **A)** A single decision tree with no bootstrapping, used to divide form IA RbcL sequences into cold and warm sequences. The biochemical properties used to divide the sequence is listed as well as the residue location. **B)** A decision tree with no bootstrapping, used to divide form ID RbcS sequences into Cold and Warm sequences. **C)** A random-forest model was implemented to find the most important protein features, indicative of warm and cold environments form IA RbcS and ID RbcS. This involved the construction of 10,000 decision trees with bootstrapping, the most occurring features capable of dividing warm and cold sequences are shown. The points of the graph represent the RbcS form, the stick represents the biochemical property that best categorises that form. The Feature importance score represents the proportion of times a certain feature was selected to best categorise the warm and cold sequences after bootstrapping over the 10,000 iterations. For the random forest models, the hyperparameters used were: training split=90%, max depth of tree=1, minimum gini impurity decrease=0.24, bootstrapping=TRUE and random state=69. The model was built using the Scikit Learn package (Version 1.2.2).

3.3.12 Random forest classifier to divide Warm and Cold Rbcs sequences

A random forest classifier model was built to highlight residues that differed significantly between 'Warm' and 'Cold' RbcS proteins in form IA and ID organisms. VHSE values were averaged along the length of the protein every five amino acids, breaking the protein into parts represented by its average biochemical properties.

Within form IA sequences there were a total of 35 Warm and 11 Cold RbcL sequences. 10,000 iterations of the decision tree classifier with 90/10 bootstrapping highlighted parts made up of residues 5-9 and 24-28 (relative to the Figure 3.14 index) could categorise RbcS proteins into Warm and Cold sequences (Figure 3.15). For residues between both 5-9 and 24-28, the electronic properties of these residues were indicative of Warm and Cold sequences. Residues 24-28 on the RbcS protein had the strongest ability to categorise sequences without bootstrapping. These residues can categorise the form IA RbcS dataset into 11 Cold with 33.7% impurity and 32 Warm with 0.00% impurity based on the electronic properties (Figure 3.15).

Amongst the dataset, there were 25 Cold and nine Warm form ID RbcS sequences. Residues between 125-129 and 135-139 could be used to categorise the genes based on their steric and electronic properties respectively. A single decision tree with no bootstrapping indicated the steric properties of residues 125-129 as the most effective means of categorising Warm and Cold form ID RbcS proteins (Figure 3.15). These properties of the residues categorised the sequences into 21 Cold, with 5.71% impurity, in the Warm category there were eight Warm with 32% impurity (Figure 3.15).

3.3.13 Positive selection within the *rbcS* gene

Form IA *rbcS* gene sites were tested for positive selection using nested models implemented in Codeml. Firstly, the nested models of M0 vs M3 indicated heterogeneity of selection pressures (Log_L =-9686.93, -10221.45, $p < 0.0001$, df = 8). The LRT models of M7-M8 and M8-M8A were used to assess positive selection and residues across the IA *rbcS* genes showing conflicting evidence of positive selection across the genes. The M7-M8 model

comparison lacked significant (p=1.00) meaning that this model of positive selection was ill fitted to the dataset. Opposingly, the more rigorous model comparison M8a vs M8 highlighted was significant, suggesting positive selection along the *rbcS* gene (Log_L =--9696.40, -9679.62, p =0.00, df = 1). Despite this, signals on individuals were not strong enough to highlight positive selection on specific codons along the *rbcS* gene (BEB <0.95).

Table 3.6- Form IA *rbcS* test for positively selected sites comparing LRTs of nested models

Model	No. S	Log-likelihood	Model compared	LRT P-value	Positive sites
M3	57	-9686.93			
M0	57	-10221.45	M0 vs. M3	0.00	NA
M2a	57	-10036.49			
M1a	57	-10036.495936	M1a vs. M2a	1.00	NA
M8	57	-7869.96			
M7	57	-9696.40	M7 vs.M8	0.99	NA
M8a	57	-9679.62	M8a vs.M8	0.00	NA

MEME selection showed there were four residues under episodic positive selection across the form IA *rbcS* gene (LRT>1, p<0.05) (Table 3.7). The positively selected residues were overlayed with close contact residues between RbcL and RbcS determined from the IA crystal structure. This highlighted two residues on the form IA RbcL, which were positively selected for and in close contact with RbcL protein in the holoenzyme (Table 3.11) (Figure 3.16).

Table 3.7- Test for episodic selection amongst form IA *rbcS* residues. Residues with significant selection pressure are shown.

Residue	LRT	p-value	Feature
2	4.53	0.01	n-terminal
49	5.77	0.05	η1
70	5.09	0.07	η2 / β1
126	12.29	0.04	β4

Table 3.8- Form ID *rbcS* test for positively selected sites comparing LRTs of nested models.

Model	No. S	Log-likelihood	Model compared	LRT p-value	Positive sites
M3	58	-9686.93			
M0	58	-10221.45	M0 vs. M3	0.00	NA
M2a	58	-10036.49			
M1a	58	-10036.49	M1a vs. M2a	1.00	NA
M8	58	-9679.62			
M7	58	-9696.40	M7 vs.M8	0.00	NA
M8a	58	-9679.62	M8a vs.M8	1.00	NA

rbcS genes from form ID organisms were examined for positive selection using nested models implemented in Codeml. Firstly, the nested models of M0 vs M3 indicated heterogeneity of selection pressures (Log_L =-11810.35, -12064.47, $p < 0.0001$, df = 8) (Table 3.8). The LRT models of M7-M8 and M8-M8A were used to assess positive selection and residues across the ID *rbcL* genes, showing no evidence to suggest positively selected residues as the LRT models were found not to differ significantly (Log_L =-11817.76, -11817.76, p =1.00, df = 2 and Log_L =-11817.76, -11817.76, p =1.00, df = 1 respectively) (Table 3.8).

Table 3.9- Test for episodic selection amongst form ID *rbcS* residues. Residues with significant selection pressure are shown.

Residue	LRT	LRT p-value	Feature
14	22.36	0	n-terminαl
28	10.68	0	αA
33	5.42	0.03	αA
37	5.47	0.03	αA/ βA gαp
41	12.43	0	βA
70	20.89	0	βA/βB
93	18.96	0	αB
95	23.41	0	αB
96	57.7	0	αB
99	4.89	0.04	αβ
100	7.73	0.01	αβ βC loop
102	11.44	0	αB βC loop
108	19.59	0	βC
114	9.89	0	βC βD loop
115	21.48	0	βC βD loop
117	24.45	0	βC βD loop
118	4.83	0.04	βD
119	33.03	0	βD
120	14.28	0	βD
133	6.3	0.02	βE
135	4.89	0.04	βE
137	16.18	0	βE
139	5.51	0.03	βE
140	11.39	0	βE βF loop
143	28.53	0	βF
145	21.16	0	βF
147	15.1	0	βF
151	16.38	0	αC
152	10.19	0	αC
153	5.06	0.04	αC

MEME selection showed there was extensive positive selection across the form ID rbcS gene with 30 residues under episodic positive selection across lineages (LRT>1, $p<0.05$) (Table 3.9). The residues positively selected were widely spread across the *rbcS* gene; however, 50% of the selected residues were found in the last 35 amino acids of the form ID RbcS protein, clustered around the C-terminal extension only observed in red-type Rubisco (Figure 3.14) (Table 3.9).

3.3.14 Examining relaxation of the *rbcS* gene between Warm, Cold and Temperate environments

RELAX test was used to assess selection pressures on *rbcS* genes found at different temperatures. Genes were grouped into Warm, Temperate and Cold groups based on the average temperature they were discovered. For both *rbcS* forms IA and ID there is a relaxation of selection in genes extracted from cold environments (K=0.42, p<0.001 and K=0.67, p=0.03 respectively) (Table 3.10). For both Temperate and Warm genes in both forms IA and ID there was no significant evidence to suggest an intensification or relaxation of selection pressures (p>0.05). When phylogenetic lineages were considered, within form IA *rbcS* genes there was strong evidence to show an intensification of selection pressures in cyanobacterial lineages (K=5.80, p<0.001) (Table 3.10). Additionally, proteobacterial lineages exhibited a significant relaxation of selection pressure (K=0.65, p=0.01). When both proteobacterial and cyanobacterial branches (basal branches) were examined for selection pressure, there was no significant change in pressures (p=0.76) (Table 3.10).

Alternatively, in ID *rbcS* genes there was no significant selection pressure change linked to phylogenetic groupings examined here. Within incident branches leading to Stramenopile evolution and incident branches leading to Haptophyta and Chlorophyta there was no significance found (p>0.05) (Table 3.10).

Table 3.10 -Test for Relaxation or Intensification of Selection Pressure across defined phylogenetic lineages across form IA and ID *rbcS*.

	Test Branches	K	p	LR	Selection
IA	Proteobacteria	0.65	0.01	12.53	Relaxation
	Cyanobacteria	5.80	0.00	26.52	Intensification
	Warm	1.10	0.73	0.93	NS
	Temperate	1.04	0.68	0.70	NS
	Cold	0.42	0.00	14.54	Relaxation
	Basal branches	4.16	0.76	0.86	NS
ID	Clade 1	5.62	0.68	0.36	NS
	Clade 2	0.79	0.75	0.19	NS
	Warm	11.03	0.33	28.50	NS
	Temperate	1.34	0.25	2.32	NS
	Cold	0.67	0.03	5.24	Relaxation
	Basal branches	3.99	0.63	0.33	NS

Existing crystal structures of form IA and ID Rubisco were used to find interacting residues between RbcL and RbcS units. These residues were compared with residues that were highlighted to be positively selected for in phylogenetic lineages through MEME selection.

Within form IA this highlighted a number of residues across the RbcL protein that are in close contact with RbcS units (<0.4A VWA) (Appendix 1.1). These were correlated with positively selected residues (Table 3.7, Table 3.9). This showed amongst form IA RbcL proteins 38.5% of residues that were positively selected for were in close contact with an RbcS protein subunit. Of the four positively selected residues on the RbcS gene, two of these residues were in close contact with an RbcL protein when the 3D structure was considered.

Form ID RbcL units had a comparative number of positively selected residues to form IA RbcL units (25 and 26 respectively) with 52% of these residues being in close contact with RbcS units. More strikingly, the *rbcS* gene from form ID organisms had 31 positively selected residues across lineages. Of these positively selected residues, 87.1% of them were in close contact with RbcL units in the crystal structure (Table 3.11) (Figure 3.16).

3.3.15 Comparing positively selected residues with close contact interactions between RbcS and RbcL subunits

Table 3.11- Summary of positively selected residues through episodic selection across *rbcs* and *rbcL* genes.

	Protein	No. of Residues under Episodic positive selection	No. of Residues in close contact with opposing subunit	Close Contact (%)
Form IA	RbcL	26	10	38.5
	RbcS	4	2	50.0
Form ID	RbcL	25	13	52.0
	RbcS	31	27	87.1

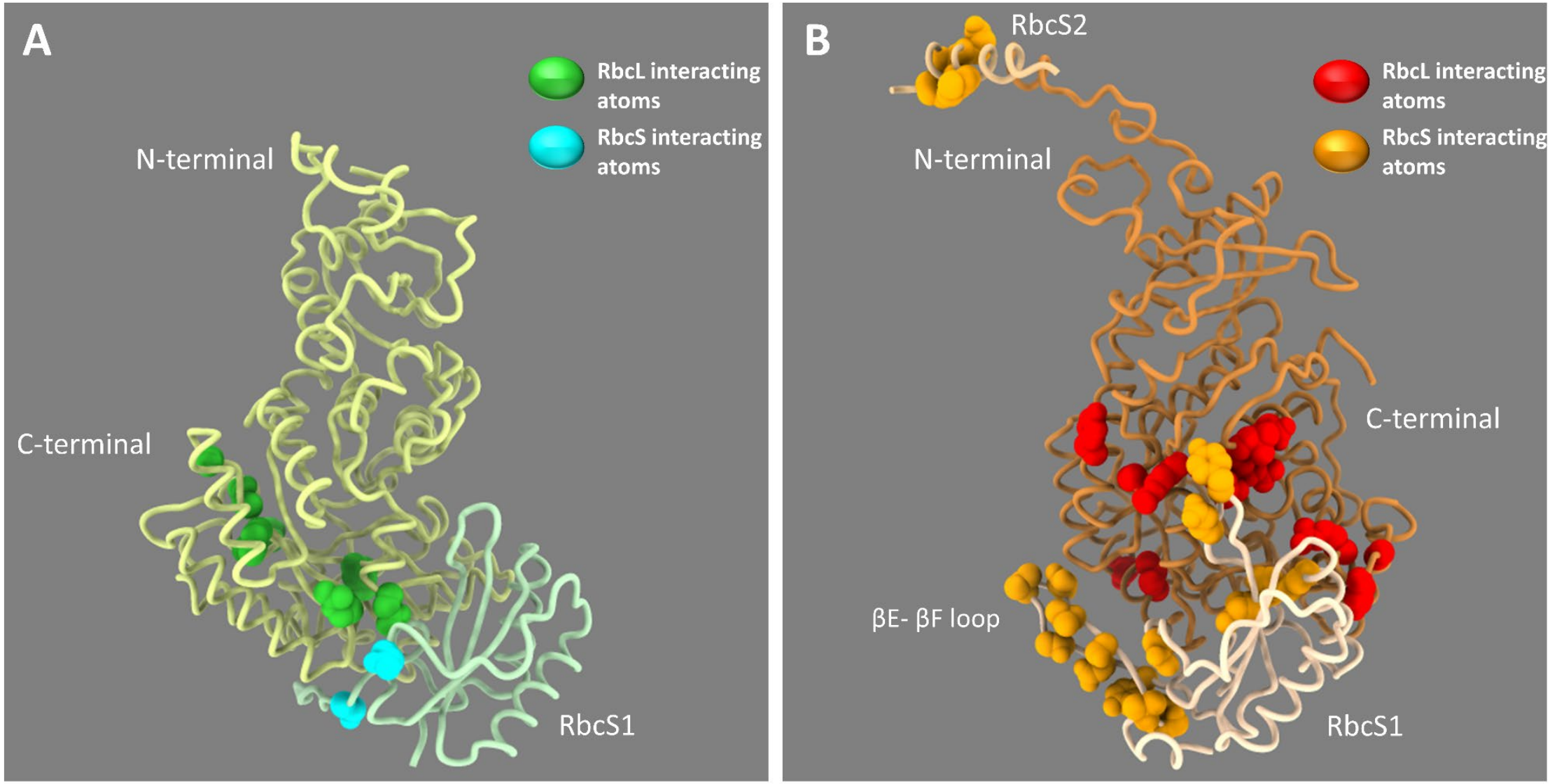

A
RbcL interacting atoms
RbcS interacting atoms
N-terminal
C-terminal
RbcS1
B
RbcS2
RbcL interacting atoms
RbcS interacting atoms
N-terminal
C-terminal
βE- βF loop
RbcS1

Figure 3.16- The loci of the residues that are both positively selected for (Table 3.2, 3.4, 3.7, 3.9) and are in close proximity between the RbcS and RbcL protein. Crystal structures from *Halothiobacillus neapolitanus* (RCSB:7ZBT) Ni et al. 2022 and *Thalassiosira antarctica* (RCSB:5MZ2) Valegård et al. 2018 were used to visualise these interactions in a 3D space. A Highlights the positions of the positively selected interacting residues in form IA RbcS and RbcL proteins IA organisms in this study. One RbcS protein is shown alongside the RbcL unit as RbcS residues that are positively selected and interact with the RbcL protein can be found solely on the RbcS on the N-terminal. **B** Highlights the residues on Form ID Rubisco proteins that were both positively selected for and in close contact interaction with each other. In the case of form ID Rubisco, residues on multiple RbcS subunits were found to positively interact with a single RbcL unit, thus two RbcS units are included in the visualisation. Additionally, the highly selected for ßE-ßF loop is also highlighted.

3.4 Discussion

3.4.1 Gaussian process model highlights predictability of environmental temperature from sequence structure

Gaussian process models have been shown to be powerful tools in predicting common Y value outputs from highly dimensional X inputs. As a result of this, GP models have been utilised as the basis of many directional evolution studies, improving protein fluorescence (Saito et al., 2018), thermostability (Romero et al., 2013) and kinetic properties (Iqbal et al., 2023).

In this study we built a GP model capable of predicting the environmental temperature of the metagenome that an RbcL or RbcS protein sequence was discovered. For both RbcL and RbcS sequences the GP model performed well at predicting environmental temperature from sequence (Figure 3.5). This improved incrementally with protein encoding form. One-hot encoding solely representing amino acids in binary form performed consistently worse across both RbcL and RbcS models. This highlights that phylogeny is not indicative of environment from which the sequence was derived. This is further highlighted by TSNE (Figure 3.3) which were used to interrogate underlying patterns in the sequence data. Based on sequence there is a clear clustering into Rubisco form groups; however, there appears to be a no/ weak clustering of sequences based on environmental temperature.

Alternatively learnt encoding which represents the protein in a numerical vector indicative of structure (Lin et al., 2022) was consistently the superior encoding form. The ESM-2 transformer used for the Rubisco learnt encoding in this study encompasses 15 billion parameters (Lin et al., 2022) which are indicative of protein structure and can be transformed into 3D protein representations. Comparing the ESM-2 encoding (transformed by ESMfold into 3D structures), with Alphafold2 (Jumper et al., 2021) models have shown equal levels of efficacy in predicting the structure of monomeric protein sequences(Lin et al., 2022). Additionally, the ESMfold model performs this task of 3D prediction at speeds 10-fold faster than Alphafold2 due to lack of alignment criteria (Lin et al., 2022). As a result, the superior performance of the learnt encoding Gaussian models used in this study suggests that there are differences in Rubisco subunit 3D structure between temperature environments. The kernel function of the Gaussian process model ensures that ‘similar’ X

inputs produce similar Y outputs. Therefore, we can additionally say that there are structural similarities within similar environments.

3.4.2 Random forest model highlights residues that differ in their biochemical properties in form IA Rubisco

VHSE residues were averaged every five residues along the length of an alignment for both RbcL and RbcS protein sequences. This numerical vector representative of the biochemical properties of each sequence was used to train a random forest model. The random forest model was used to categorise sequences into warm and cold groups based on the average environmental temperature the sequence was discovered.

Form IA Rubisco was successfully categorised into warm and cold sequences highlighting a number of residues that were indicative of their prospective environments. For RbcL the electronic properties of residues between 89-93 were most frequently used to categorise sequences after bootstrapping (Figure 3.8). This equatorial region of the RbcL is outwards facing and has been implicated in carboxysome protein binding and formation with cyanobacteria (Wang et al., 2019).

When RbcS proteins from form IA Rubisco were considered, there were several regions across the protein sequence that were indicative of cold and warm environments (Figure 3.14). With residues between 24-28 being the most important feature in categorising cold and warm sequences. This area is associated with RbcL binding on the RbcS with close contact residues at position 25 on the alignment, which is an area which is also conserved in RbcS binding in form IB organisms (Knight et al., 1990). In both cases the electronic properties of the residues at sites between 24-28 were responsible for categorising sequences. For warm sequences there was a higher propensity of negatively charged residues at sites between 24-28 with a consensus of E-24 and D-27 (as opposed to the uncharged Q-24 and the positively charged K-27 consensus residues in cold species). This is significant as the corresponding contact sites at positions 432 and 433 on the a8 helices of the large subunit are externally negatively charged. This suggest that interacting electrostatic forces between that of the large and small subunit in cold sequences are generally stronger at this site. These increased electrostatic forces are a theme that is

carried across to residues between 117-121 on the alignment with hydrophobic A at 119 in cold and the hydrophilic T-119 consensus residue in warm species. Opposingly, at the interacting site on the a1 helix on the RbcL subunit there is a consensus of V-188 in warm and A-188 in cold species. This insinuates opposing hydrophobic forces in cold sequences suggesting that RbcL and RbcS interactions are weaker.

Additionally, residues of the N-terminal between 5-13 on the alignment were also highlighted as important residues in discriminating hot and cold sequences. This divide appears to be largely due to an insert between residues at positions 5-9. This insert is indicative of form IAq bacteria that lack carboxysomes (Badger and Bek, 2008) and can even be found In IAq operons from proteobacterial species that possess both form IAqo and IAc Rubisco operons (Badger and Bek, 2008). The role of this insert is speculatory and may represent an inhibitory complex to prevent binding of CSO2 carboxysomal protein. Due to its outwards facing nature, it is not involved in interactions with the RbcL and is most likely to be involved in other protein-protein interactions. Interestingly, we found a higher proportion of IAqo sequences in Cold environment sequences highlighting the reduced prevalence of carboxysomal formation in proteobacteria from these environments. This may be due to the increased dissolved TIC in polar regions making advanced carboxysomal structures redundant.

There are biochemical differences between Warm and Cold RbcS sequences, however the extent of environmental adaption is debatable. This is due to the lack of cyanobacterial sequences from cold environments. Differences in biochemistry of Rubisco structures are more likely to highlight differences between photosynthetic machinery in proteobacteria and cyanobacteria with the latter not being found in cold environments.

3.4.3 Positive selection in form IA *rbcL* and *rbcS* gene

When all genes were considered irrespective of phylogeny (PAML analysis) we found no evidence to suggest positive selection at specific residues in either the *rbcL* or *rbcS* genes of form IA organisms (Table 3.1, 3.6). This is unsurprising with previous studies in marine organisms finding similar results (Kapralov and Filatov, 2007), (Goudet et al., 2020). (Kapralov and Filatov, 2007) suggest that the lack of positive selection amongst marine Rubisco genes is due to the increased sequence diversity in these species. In comparison,

there is far less diversity across the Rubisco genes of land plants where positive selection across Rubisco genes is abundant (Bar-On and Milo, 2019). Marine environments, and more specifically the internal environment around Rubisco in marine organisms, are more stable than those of land plants. This is due to buffered diurnal temperature fluctuations and the widespread presence of CCMs in marine systems providing a saturated CO_2 environment (Cabello-Yeves et al., 2022). Based on this environmental stability, one could hypothesis that Rubisco would be allowed to adapt kinetics to better suit the environment. With the diversity of marine environments this may be reflected in the increased sequence diversity across marine organisms. However, this theory is not confirmed by PAML analysis which provides no evidence for such adaptive selection. For this study all form IA Rubisco genes were examined including both cyanobacterial and proteobacterial species. PAML analysis may be improved by dividing the analysis into either cyanobacterial or proteobacterial species; however, this highlights an issue with PAML software, which creates a positive bias for smaller, less diverse datasets (Murrell et al., 2012).

In contrast, when we considered positive selection on phylogenetic lineages we found greater evidence to suggest adaptive positive selection at a number of sites across both the *rbcL* and *rbcS* genes. MEME selection allows for ω to vary from site to site and therefore is more flexible than PAML branch models, which require defined branch site, when multiple factors for evolution need to be considered.

There were 26 sites in total that were positively selected on the *rbcL* gene of form IA species across the phylogenetic groups (Table 3.2). These positively selected sites did not align with sites that were identified to be indicative of Cold and Warm sequences in the random forest model. This misalignment may highlight the complexity of biophysical factors acting as selection pressures on Rubisco genes and we cannot conclude that temperature is the sole driver behind adaptive evolution. Other important factors may play a role in form IA *rbcL* evolution, namely, the presence or lack of CCMs, the formation of carboxysomes, the role of Rubisco accumulation factors in Rubisco assembly and differences in Rubisco activases between proteobacteria and cyanobacteria. All these factors directly or indirectly impact the conservation of particular residues on the *rbcL* for interactions.

One commonality that is found between random forest model analyses and MEME selection analysis is that several highlighted sites can be implicated in close contact interactions with

RbcS units. This shows that interactions between the RbcS and RbcL proteins is essential for function and may pertain to a degree of environmental adaption. For the *rbcS* gene there were four sites that were selected for across phylogenetic lineages of which two are implicated in RbcL binding.

3.4.4 Random forest model highlights biochemically significant areas of form ID Rubisco

Across 10,000 bootstrap iterations of form ID RbcL the random forest model found a single feature capable of categorising RbcL proteins into Warm and Cold groups between residues 347-351 (Figure 3.8). These residues are located within the α6 secondary structural unit of the RbcL protein. This region is interesting as the area of the RbcL protein that these residues are found unusually lacks conservation. However, these residues are directly adjacent to the highly conserved loop 6 of the Rubisco protein which is directly involved in Rubisco catalysis (Parry et al., 1992). Directional evolution studies focusing on the α6 region of the RbcL protein found that changes had a distinctly negative effect, impairing carboxylase activity (Ramage et al., 1998). Based on these results, (Ramage et al., 1998) concluded that the α6 helices can influence movement and position of loop 6 and therefore impacts the catalytic action of the Rubisco enzyme. Interestingly, form IA and IC Rubisco forms have different conserved residues in the α6 helices compared to form IB and ID Rubisco sequences (Matsumura et al., 2012). The mutation Y345F converting the conserved *Rhodobacter sphaeroides* residue to the conserved form IB/ID equivalent residue improved carboxylation rate by 27% (Zhou and Whitney, 2019).

Across form IA *rbcL* genes there is a degree of phylogenetic grouping based on temperature; however, when the form ID *rbcL* phylogeny is considered, cold genes have evolved in multiple lineages and frequently share common ancestors with Warm or Temperate genes. This significantly reduces the chance of there being a common cold/warm adaptive trait across multiple phylogenies. This is further confirmed by PAML analysis which failed to find evidence of positive selection across the entire *rbcL* gene, irrespective of phylogeny.

However, when phylogeny was considered there was extensive evidence for episodic selection in the form ID *rbcL* across lineages. When we compare these highlighted sites with those positively selected in form IA *rbcL,* we find that a similar number of sites are positively selected, and a slightly higher proportion are in close contact with the RbcS subunits in the enzyme structure. Additionally, there are commonalities in the sites that are selected for; namely, residues at 221 and 226 on the alignment which are implicated in RbcL/ RbcS binding, as well as residues at the c-terminal of the rbcL genes. Both form IA in cyanobacteria and red algae share a homologous Rubisco activase, the CbbX protein, which is thought to have been transferred by lateral gene transfer (Zarzycki et al., 2012). The CbbX protein is known to act on the C-terminal of the Rubisco large subunit and is highly species-specific (Mueller-Cajar, 2017). The binding sites of the Cbbx may explain the positive selection observed on the c-terminals of both form IA and ID RbcL. Interestingly, there was also positive selection shared across both form IA and ID on the N-terminal domain between a1 and BA secondary structural units. The residues positively selected for in form IA rbcL and form ID rbcL are directly adjacent to the active site residue at 29 on the alignment (Whitney et al., 2011). This active site on the N-terminal complexes with the C-terminal active sites on the adjacent RbcL unit alignment (Whitney et al., 2011). There is a lack of conservation in neighbouring residues to this active site at position 30 in form IA and ID Rubisco, therefore this region may highlight an area of evolutionary adaption on the large subunit.

Unlike form ID RbcL there were several residues across the RbcS that were indicative of warm and cold environments based on their biochemistry (Figure 3.14). These were largely located towards the C-terminus of the RbcS protein in form ID Rubisco. The most important sites being those between of the BD/BE gap at 125-129 and the BE at residues 135-139. This is a significant area on the red type RbcS as unlike green type RbcS proteins, red type RbcS have an extended BE-BF at the C-terminus (Bracher et al., 2017). This region of the RbcS is involved in extensive interactions with multiple RbcL units in the holoenzyme structure. The beta hairpin loop of the ßE- ßF regions on the RbcS extends into the axial pore of the Rubisco enzyme. This extension is not only thought to aid form ID Rubisco biogenesis, but additionally, interactions are involved in stabilising the enzyme complex owing to high Rubisco specificity found in red algal Rubisco (Oh et al., 2023). Therefore, it particularly

interesting that this area has been highlighted as an area which is divergent between Cold and Warm species.

Not only this, but when we consider the MEME positive selection analysis, a number of the residues between 135-139 of the BE secondary structural element, indicated by the random forest model, are also positively selected for (Table 3.2). PAML analysis failed to highlight any positively selected (Table 3.7) residues across the form ID *rbcS* gene; therefore, it is clear that positive selection, highlighted by the MEME software is a result of adaptive evolution. This highlights the significance of this area as a potential area of evolutionary adaption to cold environments.

When we look more closely and compare the consensus sequences of this structure in Warm and Cold form ID sequences, focusing predominantly on positively selected residues, we find an interesting pattern emerging. Firstly, the consensus residue at 135 in Warm and in Cold is V and D respectively, with reduced conservation in cold sequences. Valine hydrophobic, aspartic acid is negatively charged 137 is Q and T 139 is A and G. Effectively, amongst warm sequences there is predominance of hydrophobic residues which interact with the highly hydrophilic residues on the a5 helices of the RbcL unit. The divergence from this relationship in Cold sequences suggests that the electrostatic forces may be weakened at this point between the RbcL and RbcS.

Based on this weakening of interactions between RbcL and RbcS units in cold species, we can hypothesise that this in turn has an impact on the kinetic properties of Rubisco. Weakened interactions could reduce specificity of the enzyme but allow Rubisco rate to increase when rate is significantly limited due to low temperatures. The one available study comparing red type Rubisco from cold environments demonstrated that psychrophilic diatoms had rates proportionally higher than that of mesophilic diatoms at low temperatures (Young et al., 2015).

Not only do we find positive selection on the form ID RbcS on the BE secondary structural elements, but we also find extensive positive selection across the entire *rbcS* gene in form ID organisms. This is striking, firstly because there were four positively selected residues in form IA *rbcS* genes, compared to the 31 in total across the form ID *rbcS* genes. Secondly of these 31 positively selected residues 87% were in close contact with RbcL units in the

holoenzyme structure. This further highlights the important role that the RbcS protein plays in the biogenesis and function of form ID Rubisco, and highlights that the interactions between RbcL and RbcS are malleable based on either physiological or biophysical factors.

There is further evidence to suggest evolution in response to temperature with a relaxation of selection pressure in form ID *rbcS* genes derived from cold environments. This relaxation of sites associated with RbcL units may drive the loosened epistatic interactions within the Rubisco enzyme, allowing increased kinetic rates at lower temperatures. Additionally, the relaxation and intensification of selection of form IA *rbcS* genes had a strong phylogenetic signal; being linked to proteobacterial and cyanobacterial lineages respectively. In form ID *rbcS* genes, no phylogenetic signals were observed, thus strengthening the argument that the environment is the evolutionary driver.

3.5 Conclusion

In conclusion, this study highlights regions of the Rubisco enzyme which have evolved within form IA and ID Rubisco species. Within form IA Rubisco species, evolution and selection pressure increased with evolution of photosynthesis in bacteria, namely into cyanobacteria. This may be a result of multiple physiological changes such as the ubiquitous presence of carboxysomes across cyanobacteria or the incorporation of Rubisco specific assembly factors and activases. Positive selection in the RbcS unit was limited, but again intensified in cyanobacteria. This is expected as there is evidence to suggest additional RbcS interactions with carboxysomal proteins in its formation within cyanobacteria (Ryan et al., 2019).

However, in form ID Rubisco, there is significant evidence to suggest environmental adaption to temperature especially within the RbcS unit with there being no links to phylogeny. We see a relaxation of selection in Cold species and we propose this has resulted in weakened epistatic interactions between RbcS and RbcL units. We know that the RbcS is responsible in form ID organisms for the observed high specificity of Rubisco. As a result of this, we theorise that the weakened interactions cause a decrease in specificity but an increase in Rubisco rate. This adaption to the environment may be part of the adaptive mechanisms that allow form ID Rubisco to be the dominant Rubisco species in polar environments.

Heterogenous expression of Red Algal Rubisco can Increase Carbon Assimilation and Reduce Water Usage when coupled with Reduced Stomatal Density In Wheat

4.1 Introduction

Future food security is threatened by an ever-changing climate. Desertification is reducing viable farming land (Mirzabaev et al., 2019), whilst increasing temperatures, unpredictable weather (Raza et al., 2019) and increased insect predation (Skendžić et al., 2021)are putting greater stress on the crops that remains. As a result of this efforts to 'future-proof' crop production are being made through engineering drought resistance (Shinwari et al., 2020), insect tolerance (Li et al., 2020) and maximising yields by increasing carbon assimilation (Wu et al., 2023).

The enzyme Rubisco is considered a significant bottleneck in photosynthesis especially within C_3 crops where the enzyme is environmentally constrained through its complex biogenesis (Bracher et al., 2017). As a result of this previous modelling strategies have been implemented to hypothesise the benefits of heterogenetic expression of more efficient Rubisco species in crops (Zhu et al., 2004), (Iqbal et al., 2021), (Wu et al., 2019), (Wu et al., 2023).

The outcomes of these modelling experiments have been conflicting with some studies presenting significant improvements in assimilation (>25%) (Zhu et al., 2004), (Iqbal et al., 2021). On the other hand, studies suggest that altering the maximum carboxylation rate (Vc_{max}) alone is not enough. This is because photosynthesis remains constrained by the supply of ATP from the light reactions and conductance of CO_2 through the mesophyll (Wu et al. 2019). Additionally, Busch (2020) argues that the 'inefficient' process of photorespiration is intrinsically necessary for the plant; acting as an energy dissipation pathway, preventing oxidative damage under high-light conditions. As well as this, the photorespiratory pathway provides the precursors for the synthesis of a number of essential amino acids. Thus significantly reducing photorespiration may have a detrimental effect on the plant.

Despite this conflicting information, a potential application of heterogeneous Rubisco expression is in that of increasing drought tolerance through maximising water use efficiency. Previous studies have demonstrated that it is possible to increase drought resistance in crop plants by significantly reducing the density or size of stomata (Caine et al., 2019) (Xie et al., 2012), (Liu et al., 2015). The effects of which have been shown to not have a detriment on carbon assimilation or yields (Dunn et al., 2019).

In this study we explore the concept of increasing water use efficiency by reducing stomatal density with simultaneous heterologous expression Rubisco from alternative forms. Form IA, IC and ID Rubisco have been shown to have a greater variability of kinetic parameters relative to that of form IB from C_3 plants (Oh et al., 2023). The fastest form I Rubisco can be found in that of form IA organisms of the cyanobacterial *Synechochus sp.* (Lin et al., 2014). Whilst the most specific and efficient Rubisco enzyme previously discovered belongs to a form ID red algal species (Andrews and Whitney, 2003). This poses engineering opportunities that have not been fully explored. As reducing stomatal density will inevitably change the intercellular CO_2 (C_i) environment, we can hypothesise that a foreign Rubisco would be better suited to this decreased C_i environment. This is because aquatic environments can often be limiting in CO_2 and therefore the organisms have adapted accordingly (Bar-On and Milo, 2019).

Through modelling efforts built on the 'big leaf concept' which simulates the carbon assimilation across an entire field as if it was one big leaf (Iqbal et al., 2021), (Rogers et al., 2017), (Bonan, 2019). We explore the effects that reducing stomatal density and heterogenous expression of aquatic Rubisco may have on overall Carbon assimilation across a growing season in wheat. As a result, our null hypothesis is that wheat Rubisco is optimally suited to its environment and carbon assimilation would not be improved by an aquatic Rubisco species. Additionally, we can hypothesise that reducing stomatal density will not influence water usage and carbon assimilation will remain unchanged. Furthermore coupling both changes to stomatal density and Rubisco species will have a detrimental effect on carbon assimilation over an entire growing season. These hypothesises will be examined in this study.

4.2 Methodology

4.2.1 Model overview

This model builds on principles implemented in (Iqbal et al., 2021) who developed an earth system model with a combined sunlit/ shaded model applied for photosynthetic dynamics through a leaf canopy. In short this primarily involves the input of PAR and temperature which are then translated to net radiation absorbed through the canopy using a radiative decay function. From this V_{max} and J_{max} are calculated for the sunlit and shaded portions of the plant and this allows for an overall calculation of net assimilation.

Internal CO_2 concentrations are intrinsic to assimilation rates and this is solved through a Newton-Raphson iterative process balancing outputs of Stomatal conductance and An calculations (Sun et al., 2012). Stomatal conductance requires the environmental inputs of wind, humidity, and soil water content.

Finally calculations of Rn and Gs allow for the estimation of transpiration through the Penman-Monteith equation for transpiration. Transpiration rate has a direct effect on internal leaf temperature and thus another iterative process was used to balance leaf temperature derived from assimilation calculations and transpiration (Figure 4.1).

In this model heterologous Rubisco expression was modelled in a wheat system. The Rubisco kinetic values were taken from a number of previous studies on aquatic Rubisco from form IA, IB, IC and ID organisms (Table 4.1).

For this the Rubisco K_{cat} and $S_{c/o}$ was used, maintaining native expression levels, heat activation values and J_{max} (which is calculated from Vc_{max}).

Table 4.1- The Rubisco kinetic properties used for modelling

Phylogenetic Group	Form	Organism	K_{cat}	$S_{C/O}$	Reference
Angiosperm	IB	*Triticum aestivum*	***2.2***	**100**	Sharwood et al. (2016)
Proteobacteria	IA	*Allochromatium vinosum*	6.7	41	Jordan and Chollet (1985)
	IA	*Rhodobacter capsulatus*	2.5	25.9	Horken and Tabita (1999b)
	IA	*Thiobacillus denitrificans*	1.4	53.4	Hernandez *et al.* (1996)
Cyanobacteria	IA	*Prochlorococcus marinus*	6.6	59.9	Shih *et al.* (2016)
	IB	*Synechococcus elongatus*	9.8	50.3	Shih *et al.* (2016)
	IB	*Synechococcus sp.*	8.6	43.3	Ninomiya *et al.* (2008)
Green algae	IB	*Chlamydomonas reinhardtii*	1.8	64	Zhu and Spreitzer (1994)
Proteobacteria	IC	*Rhodobacter sphaeroides*	3.7	58.4	Gunn *et al.* (2020)
	IC	*Cupriavidus necator*	2.1	74	Lee *et al.* (1991)
	IC	*Bradyrhizobium japonicum*	2.2	74.8	Horken and Tabita (1999a)
	IC	*Xanthobacter flavus*	1.4	44.4	Horken and Tabita (1999a)
Red macroalgae	ID	*Griffithsia monilis*	2.6	167	Whitney *et al.* (2001)
Red microalgae	ID	*Galdieria sulphuraria*	1.2	166	Whitney *et al.* (2001)
	ID	*Galdieria partitia*	1.6	238.1	Uemura *et al.* (1997)
	ID	*Cyanidium caldarium*	1.3	224.6	Uemura *et al.* (1997)
	ID	*Porphyridium purpureum*	1.4	143.5	Uemura *et al.* (1997)
	ID	*Porphyridium cruentum*	1.6	128.8	Read and Tabita (1994)
	ID	*Nannochloropsis* sp.	1	27	Tchernov *et al.* (2008)
	ID	*Olisthodiscus luteus*	0.8	100.5	Read and Tabita (1994)
Diatom	ID	*Cylindrotheca* N1	0.8	105.6	Read and Tabita (1994)
	ID	*Cylindrotheca fusiformis*	2	110.8	Read and Tabita (1994)
	ID	*Thalassiosira weissflogii*	3.2	79	Young *et al.* (2016)
	ID	*Thalassiosira oceania*	2.4	80	Young *et al.* (2016)
	ID	*Chaetoceros calcitrans*	2.6	57	Young *et al.* (2016)
	ID	*Chaetoceros muelleri*	2.4	96	Young *et al.* (2016)
	ID	*Phaeodactylum tricornutum*	3.2	108	Young *et al.* (2016)
	ID	*Fragilariopsis cylindrus*	3.5	77	Young *et al.* (2016)
	ID	*Thalassiosira hyalina*	4.1	99	Valegård *et al.* (2018)
	ID	*Bacterosira bathyomphala*	4.6	87	Valegård *et al.* (2018)
	ID	*Skeletonema marinoi*	4.6	96	Valegård *et al.* (2018)
	ID	*Thalassiosira nordenskioeldii*	4.7	82	Valegård *et al.* (2018)
	ID	*Thalassiosira antarctica*	3.7	90	Valegård *et al.* (2018)
Coccolithophorid	ID	*Pleurochrysis carterae*	3.3	102	Heureux *et al.* (2017)
	ID	*Tisochrysis lutea*	2.2	89	Heureux *et al.* (2017)
	ID	*Pavlova lutheri*	2.5	125	Heureux *et al.* (2017)

4.2.2 Sampling site and measurements

The sample site used for the modelling efforts in this study was the IT-CA2 site (42.3772, 12.0260) located in central Italy. The field in question is situated 200m above sea level and as a result has a relatively low annual temperature of 14^{o}C a year and an annual rainfall of 766 mm. The field is planted on a rotational basis alternating between bare grassland and winter wheat, *Triticum aestivum L*. The years examined were the 2012-2013 growing season with winter wheat being sowed at the start of November 2012 and the crop being harvested in the July of 2013 (Sabbatini et al., 2016). No fertilizer was applied during this growing period and crops were only lightly irrigated in the summer months.

In the field there were two fluxnet towers positioned at opposing ends of the field measuring environmental parameters and CO_2 flux measurements at 30 minute intervals (Sabbatini et al., 2016). Daily averages were taken for each environmental parameter after time periods with solar radiation less than 5 W m^{-2} were removed, as these were deemed as nighttime observations where gas exchange should be minimal (Houborg et al., 2012), (Iqbal et al., 2021).

4.2.3 Intercellular CO_2 and carbon assimilation

Objectively intercellular CO_2 concentrations (C_i) are difficult to measure as it depends on two variables which are challenging to quantify. These are mesophyll conductance levels and intercellular respiration levels as well as abundance and rate of carbonic anhydrase in the chloroplast providing Rubisco with CO_2. Therefore, in earth systems models C_i is simplified to equation (4.1).

$$Ci = Ca - \frac{1.4}{Gb}An - \frac{1.6}{Gs}An \tag{4.1}$$

This highlights the need for a consensus C_i which is complementary to both Gs and Net carbon assimilation. However, to achieve this an initial guess is required for An and Gs calculations. This initial guess is taken as (equation 4.2) which is an estimated ratio for wheat

(Bonan, 2019). For the purpose of this modelling study Chloroplastic CO_2 concentrations are assumed to be equal to C_i concentrations and thus C_i is used throughout.

$$Ci = 0.87Ca \tag{4.2}$$

This estimate allows the calculation of An (equation 4.3) as the minimum of calculated values for the Rubisco limited rate A_c (equation 4.4) the light limited rate A_j (equation 4.5) and the phosphate limited rate A_p (equation 4.6).

$$An = min(Ac, Aj, Ap - Rd) \tag{4.3}$$

$$Ac = \frac{Vcmax(ci - \Gamma *)}{Ci + Kc(1 + \frac{Oi}{Ko})} \tag{4.4}$$

$$Aj = \frac{Jmax(ci - \Gamma *)}{4ci + 8\Gamma *} \tag{4.5}$$

$$Ap = 3Tp \tag{4.6}$$

$$\Gamma * = \frac{Oi}{Sc/o} \tag{4.7}$$

Where Oi (equation 4.7) is the intercellular oxygen concentration, assumed to be a constant 210 mmol O_2 mol^{-1} and $\Gamma *$ is the CO_2 compensation point calculated as per (equation 4.7). Vc_{max} was calculated at 25°C for sunlit and shaded leaves as shown below.

$$Vcmax25 < -Co * Kcat * FLnr * N \quad (4.8)$$

$$N = 1/(SLA * CN))/14.0057 \quad (4.9)$$

$$Vcmax25, sun = Vcmax25 \cdot \{\left[1 - e^{-(Kn+Kb)LAI}\right]\frac{1}{Kn+Kb}\} \quad (4.10)$$

$$Vcmax25, shad = Vcmax25 \cdot \{[1 - e^{-Kn\,LAI}]\frac{1}{kn} - \left[1 - e^{-(Kn+Kb)LAI}\right]\frac{1}{Kn+Kb}\} \quad (4.11)$$

Where Co represents the moles of Rubisco active sites per mole of Rubisco (Houborg et al. 2013), FLnr represents the fraction of total soluble protein that is made up by Rubisco at 25°C, defined as 0.4120 in wheat (Iqbal et al. 2021). K_{cat} is the catalytic rate of wheat µmol m^{-2} s^{-1} and N is the nitrogen content per unit leaf area g N m^{-2} derived from the specific leaf area carbon content (SLA g C m^{-2}) and C:N ratio which are both defined for wheat as 0.07 and 20:1 respectively (equations 4.8, 4.9)

Vc_{max} was adjusted to temperature specific values using equation (4.12) as was Sc/o using specificic heat activation values for the K_{cat} and $S_{c/o}$ of Rubisco respectively. Ta represents the air temperature (°K) and the R constant was 8.314 J K^{-1}. This also allowed the calculation of Q_{10} values (Ito et al. 2015)(equation 4.13).

$$Vcmax = Vcmax25 \cdot e^{\frac{HA}{298.15 \cdot 0.001 \cdot Rgas}1 - \frac{298.15}{Ta}} \quad (4.12)$$

$$Q_{10} = \frac{Vcmax1^{\frac{10}{T2-T1}}}{Vcmax2} \quad (4.13)$$

Stomatal conductance was also calculated using the initial C_i guess (equation 4.2).

$$Gs = g0 + g1bw\,\frac{An}{Cs}hs \quad (4.14)$$

$$Cs = Ca - \frac{An}{Gb} \quad (4.15)$$

For this g0 and g1 represent known minimum and maximum values for stomatal conductance defined as 0.01 and 5.78 in wheat (mol m^{-2} s^{-1}), hs is the external relative humidity and Cs is the surface level CO_2 concentration (µmol mol^{-1}). bw describes soil water content θ. This is calculated as per equation 4.16. θ_w is the soil water content at wilting point, defined as 0.1. θ_c is the soil water content at saturation point. This is soil dependant, for this study the ITA-2 landsite was planted on silty clay loam and therefore θ_c was fixed at 0.477.

$$bw = \frac{\theta - \theta_w}{\theta_w - \theta_c} \tag{4.16}$$

Once stomatal conductance and An were calculated on estimate C_I values an iterative process was used to find a fixed C_i which satisfies both the output of An calculations and Gs.

For this the Newton-Raphson approach was applied as per Sun et al. 2012. This involved an iterative method used to a find a derivative (equation 4.18) which adjusted C_i on each iteration until a convergence was met (i.e the function (equation 4.17) was reduced to <0.001). The function used to represent the relationship between Gs and An was equation (4.17).

$$f(x) = \left[Cs - \frac{A(Ci)}{Gs}\right] -] - [(Ci/Ca)Ca] \tag{4.17}$$

$$f'(x) = \frac{[f(x + f(x)) - f(x)]}{f(x)} \tag{4.18}$$

The adjusted C_i values were then used to calculate final values for An and Gs for sunlit and shaded leaf fractions (equations 4.3, 4.14). mu8

4.2.4 Net radiation absorbed

To calculate net radiation absorbed through a canopy, assumptions were made about the leaf angle based on the growth dynamics of wheat, defined as the Ross indices which

indicates deviation from a spherical growth pattern. This then allowed calculations pertaining to the reflectance and vertical transmission of radiation both upwards and downwards through the canopy.

$$I^{\rightarrow}sha(x) = I^{\rightarrow}d(x) + I^{\rightarrow}bs(x) \quad (4.19)$$

$$I^{\rightarrow}sun(x) = I^{\rightarrow}sha(x) + (1 - \omega\ell)Kb\, I_{sky,b} \quad (4.20)$$

These components allow the calculate the net radiation absorbed by the shaded leaf fraction (equation 4.19) and the sunlit leaf fraction (equation 4.20). The shaded leaf fraction is calculated by the sum off the diffuse ($I^{\rightarrow}d(x)$) and scattered beam radiation $I^{\rightarrow}bs(x)$. Whereas the net radiation absorbed by the sunlit fraction is the sum of the shaded net radiation absorbed, the total downwards direct beam radiation ($I_{sky,b}$), multiplied by the direct beam distinction coefficient (*Kb)* and the scattering beam coefficient ($\omega\ell$) (equation 4.20). More details on how each parameter can be calculated in equations 4.19 and 4.20 can be found in (Iqbal et al., 2021), (Bonan, 2019). When multiplied with reported LAI the above calculations can be used to calculate total absorbed radiation in sunlit and shaded leaf fractions.

4.2.5 Transpiration

Leaf transpiration was predicted using the Penman-Monteith equation for evapotranspiration (Goudriaan and Van Laar, 2012), (Wu et al., 2023) which uses inputs of temperature and net radiation (Rn) for calculations as well as resistance parameters for water movement through the leaf.

$$\lambda T = \frac{sRn + VPD_1 - PaCp/rbh}{s + y(rsw + rbw)rbh} \quad (4.21)$$

$$VPD_1 = SVP_{leaf} - SVP_{dew-point} \quad (4.22)$$

$$SVP = 0.61365e^{\frac{17.502T}{240.97+T}} \quad (4.23)$$

$$dew - point = Ta - ((100 - RH)/5) \quad (4.24)$$

$$PaCp = \left(\frac{101.325}{R_{gas} * Ta}\right) * 1012 \quad (4.21)$$

$$rbh = 100 * \sqrt{\frac{leaf\ width}{u}} \quad (4.22)$$

$$rs = \frac{1}{Gs} x\ p_{air} \quad (4.23)$$

$$p_{air} = ATM \times 100000\ (R_{gas} \times (Ta + 273.15)) \quad (4.24)$$

$$rsw = \frac{rs}{1.6} \quad (4.25)$$

$$rbw = 0.93\ x\ rbh \quad (4.26)$$

Rn is sum of net radiation absorbed by sunlit and shaded leaf partitions (W m^{-2}) calculated from equations (4.19, 4.20). The vapor pressure deficit (VPD) (kPa) (equation 4.22) represents the difference in partial pressure between the leaf and the air. Saturated vapor pressure (SVP) (kPa) (equation 4.23) was used to calculate VPD. y represents the psychometric gas constant of 0.066 kPa. s (constant) is calculated as the difference between SVP at T_{a+1} and T_a[14]. PaCp is dry air density (kg m^{-3}) x specific heat capacity of air J $kg^{-1}K^{-1}$.

rbh, rsw and rbw all represent resistance values. rsw is the stomatal resistance calculated from Gs (equation 4.23), rbw (equation 4.25) is the leaf boundary layer resistance to water vapour and rbh is the leaf boundary layer resistance to heat (equation 4.22). ATM is 1.01 bars.

Once T was calculated, leaf temperature was calculated by rearranging T to assess T_{leaf} (equation 4.27). This was combined with an iterative process which adjusted T_{leaf} sequentially with VPD then Transpiration to a value between 10°C colder and 2°C warmer than air temperature. The resulting value was then used to calculate a final value for An using equation (equation 4.3).

$$T_{leaf} = \frac{\gamma(rtw)Rn\ PaCp\ /\ -VPD_1}{s + y(rtw)} - T_{air} \quad (4.27)$$

$$rtw = \frac{(sRn + VPD_1 - PaCp\ rbh - s\lambda E\ /\ )rbh}{\lambda E \times \gamma} \quad (4.28)$$

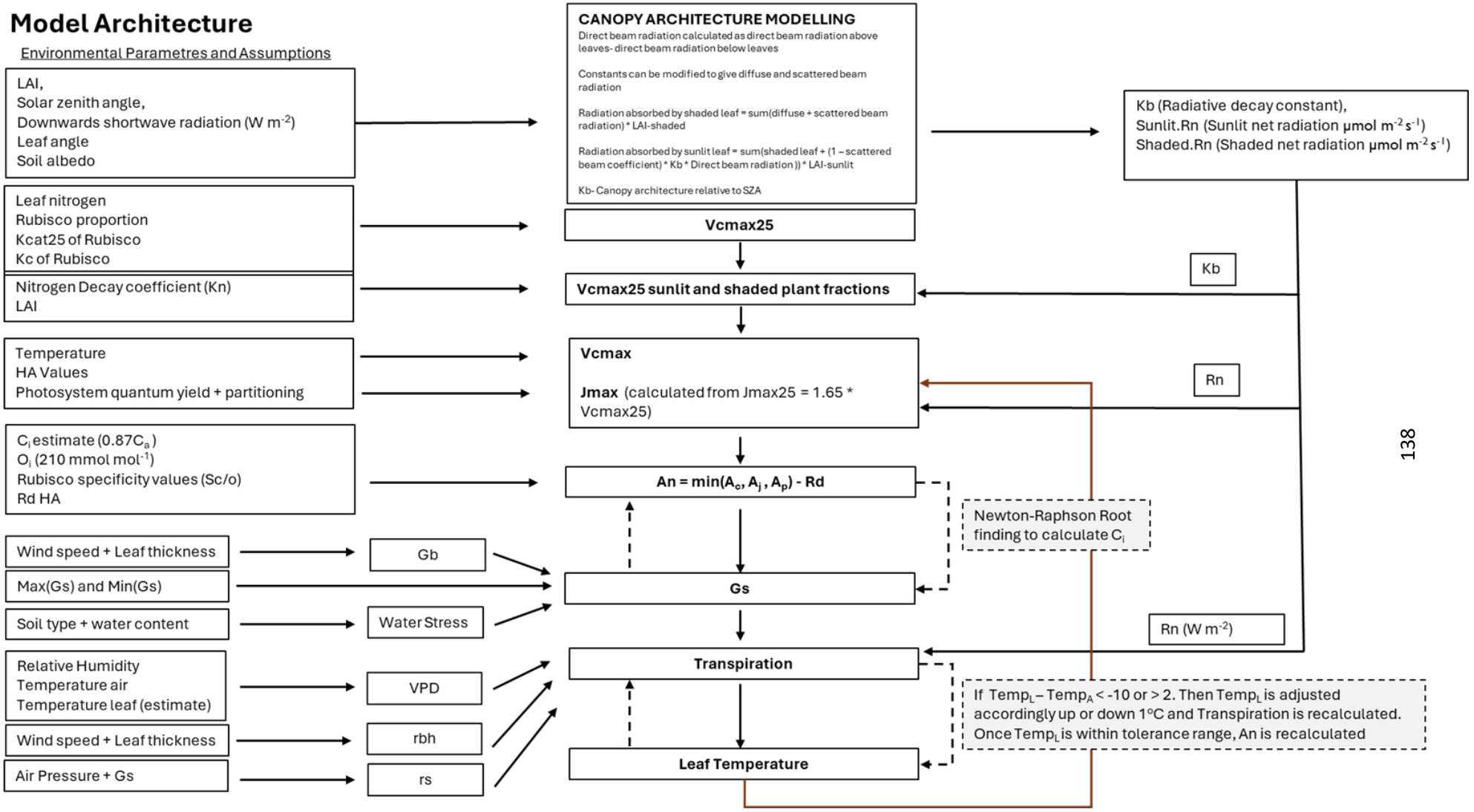


Figure 4.1 – Schematic showing how the field level Carbon assimilation is modelled for wheat across the growing season.

4.2.6 Model validation, statistical analysis and packages

Each daily modelled net assimilation was compared with daily averaged CO_2 flux data from the measurement towers across the entire growing season. To assess the simulations the total mean absolute error and determination coefficient (R^2) were used to compare modelled and observed values using custom functions implemented in Rstudio (V.2023.09.1.494)

Significant difference between total carbon assimilation and transpiration between Rubisco species was assessed using Mann-Whitney U comparison of medians assuming unequal variance (Rstudio 'stats' v.4.3.0). All graphs were generated using ggplot2 (v.3.4.2)

4.3 Results

4.3.1 Comparison of leaf level carbon assimilation in wheat with foreign aquatic Rubiscos

The net carbon assimilation by wheat was modelled at the leaf-level at 25^{o}C and 1800 µmol PAR m^{-2} s^{-1} for predicted C_i concentrations (344.8 µmol mol^{-1}) derived from ambient C_a concentrations. Additionally, the effect on net carbon assimilation was modelled under the heterogenic expression of aquatic Rubiscos assuming equivalent expression levels of foreign Rubisco in a wheat system. At the leaf level, wheat was capable of fixing 27.14 µmol of CO_2 m^{-2} s^{-1} being the 12th most efficient Rubisco modelled in this study (Figure 4.2). Only form ID Rubisco species were higher with Rubisco from *G. monillis* having the highest overall fixation at 30.06 µmol of CO_2 m^{-2} s^{-1} (Figure 4.2). All IB, IC, and IA Rubisco species were less efficient at fixing carbon dioxide under the above conditions with Rubisco from R*hodobacter capsulatus,* form IA, being the least efficient at 14.28 µmol of CO_2 m^{-2} s^{-1} (Figure 4.2).

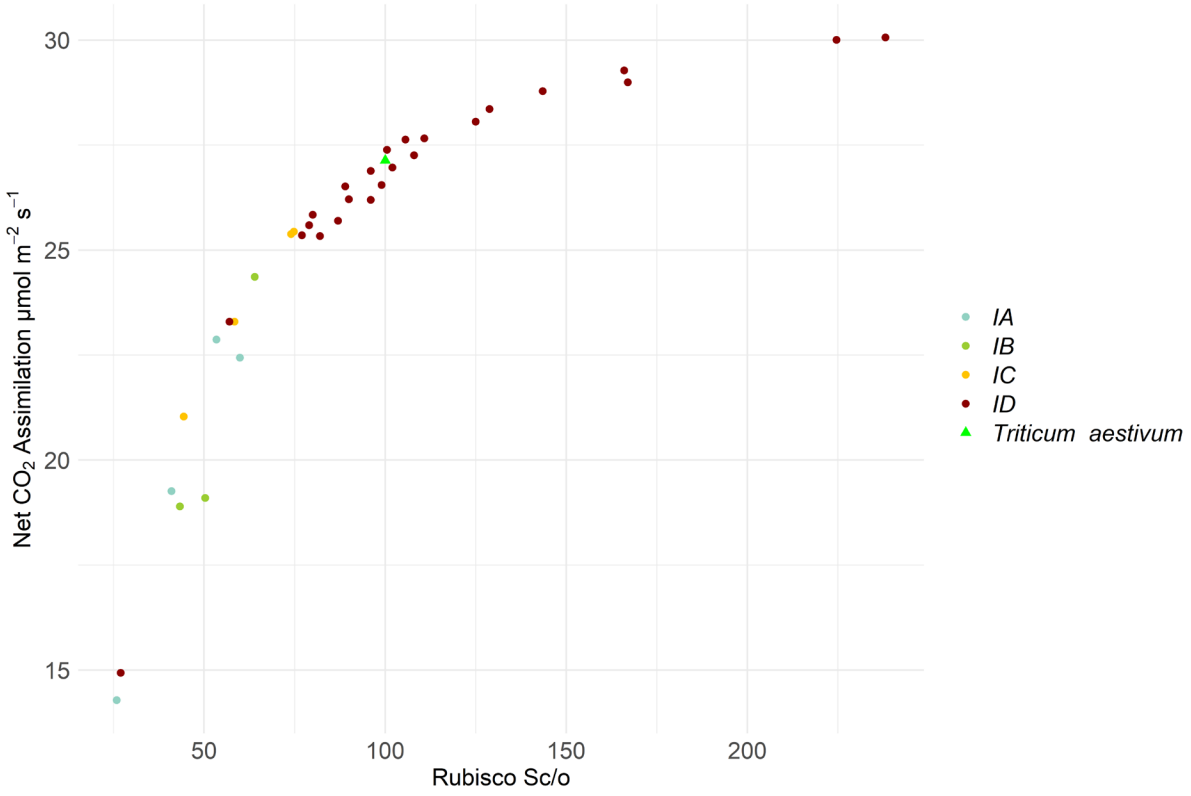


Figure 4.2 – The net assimilation of CO_2 (µmol m^{-2} s^{-1}) was modelled at the leaf level in a wheat system with native Rubisco and heterogenically expressed Rubisco for a number of aquatic organisms of differing Rubisco form (Table 4.1). For this temperature was assumed to be 25°C and net radiation was set at 1800 µmol PAR m^{-2} s^{-1}. C_i was estimated at 0.87Ca giving a concentration of 344.8 µmol mol^{-1}. Sc/o is calculated as $\frac{Vc_{max}}{K_c}\frac{K_o}{Vo_{max}}$.

Being the most efficient Rubisco at ambient CO_2 conditions, the carbon assimilation of *G. monillis* was modelled at a range of C_i environments once again constraining temperature to 25°C under and radiation to 1800 µmol PAR m^{-2} s^{-1} (Figure 4.5). A_c limited rate was superior at low C_i concentrations with heterogenic expression of *G. monillis,* relative to wheat (Figure 4.3). Additionally, the A_j light limited rate was also improved with heterogenic expression of *G. monillis.* increasing the maximum Net assimilation. The relative difference between G. monillis Rubisco and T. aestivum Rubisco decreases with increasing C_i concentrations (Figure 4.3)

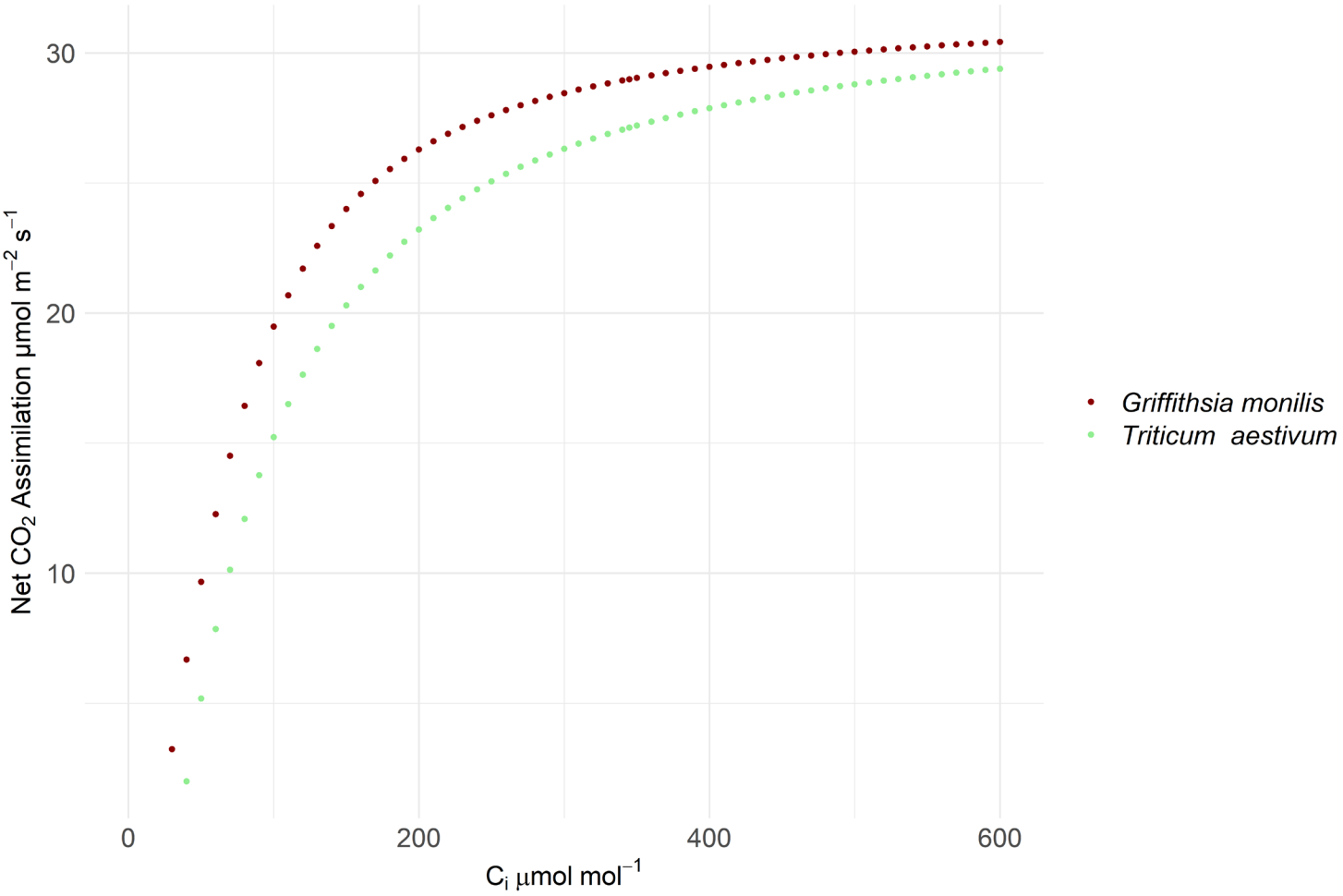


Figure 4.3- A modelled ACI curve for the net carbon assimilation (µmol m^{-2} s^{-1}) in a wheat system with the native wheat Rubisco compared to heterogenic expression of *Griffithsia monills* Rubisco. For this temperature was kept constant at 25°C and net radiation was set at 1800 µmol PAR m^{-2} s^{-1}. C_i µmol mol^{-1} was increased sequentially by 10 µmol mol^{-1}.

4.3.2 Comparison of simulations with real world ecosystem exchange over a growing period in wheat

Using environmental parameters derived from fluxnet data corresponding to Field ITA-2 and LAI measurements, the Net CO_2 assimilation per day was modelled for wheat crop across a growing season of 211 days. This started in November 2012 and ended in July 2013. The

modelled Net daily CO_2 assimilation (µmol m^{-2} d^{-1}) was contrasted with the observed CO_2 uptake measured as the inverse of NEE (µmol m^{-2} d^{-1}) from the fluxnet data tower. This comparison between modelled and observed Net CO_2 assimilation showed an accurate level of performance from the model (r^2= 0.90, MAE =0.98) (Figure 4.4A).

Following the calculation of net assimilation, transpiration was calculated using the Penman-Monteith evapotranspiration equation. For this parameters of Rn (W m^{-2}) and Gs (mol m^{-2} s^{-1}) were required for calculations with Gs being derived from an iterative process, balancing C_i between assimilation rate and Gs. The resulting calculations are shown in Figure 4.4B and Figure 4.4C respectively. This highlights that stomatal conductance and transpiration both closely follow the Net assimilation of the wheat crop.

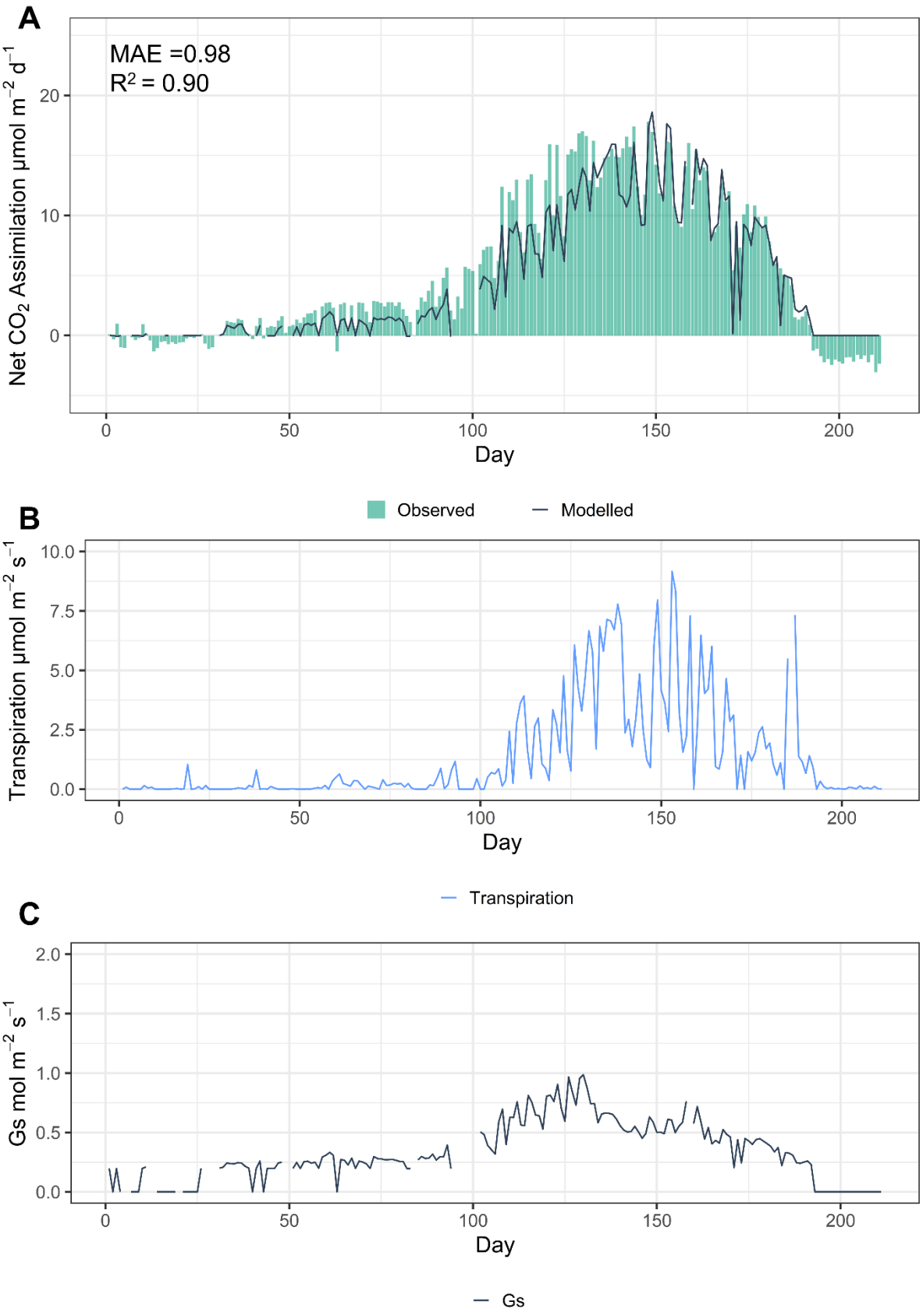

A
MAE =0.98
R2 = 0.90
Net CO2 Assimilation µmol m−2 d−1
20
10
0
0
50
100
150
200
Day
Observed
Modelled
B
Transpiration µmol m−2 s−1
10.0
7.5
5.0
2.5
0.0
0
50
100
150
200
Day
Transpiration
C
Gs mol m−2 s−1
2.0
1.5
1.0
0.5
0.0
0
50
100
150
200
Day
Gs

Figure 4.4- A) The comparison between observed and modelled net CO_2 assimilation values. The time period (days) represents the time across the 2012-2013 growing season for winter wheat at fluxnet site ITA-CA2. Observed values are represented by turquoise bars and modelled values are highlighted by the black line. **B)** is represented as the mean Transpiration (μmol m^{-2} s^{-1}) for each day highlighted by the blue line, calculated for the 2012-2013 winter wheat growing season. **C)** is the average Gs (mol m^{-2} s^{-1}) for each day highlighted by the blue line, calculated for the 2012-2013 winter wheat growing season.

4.3.2 Temperature response of Rubisco

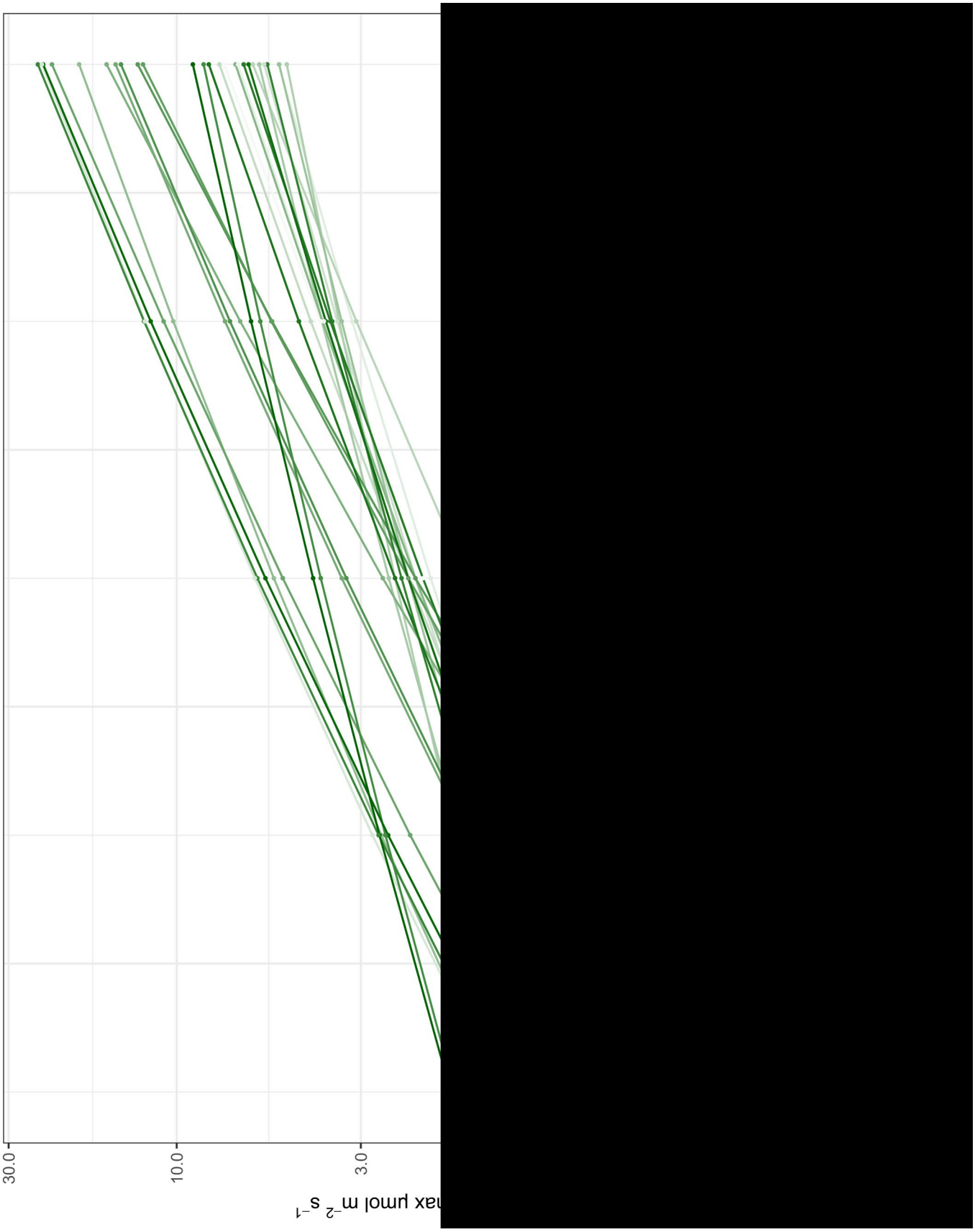


Figure 4.5 – Vc_{max} (µmol m^{-2} s^{-1}) was calculated for Rubisco species reported with corresponding heat activation values (Sharwood et al., 2016), (Hermida-Carrera et al., 2016). Temperatures were fixed at 5,15,25 and 35°C and the x-axis was plotted on a log_{10} scale to linearly represent the exponential relationship between temperature and Vc_{max}. Each line colour represents a different species reported, thus highlighting the diversity in the temperature response across different Rubisco species. For this analysis there was assumed to be no inhibition of the Rubisco enzyme at higher temperatures and the quantity of Rubisco assumed to be 1 mol (Thus Vc_{max} at 25°C is the reported K_{cat}).

Table 4.2- The calculated Q10 values for Rubisco species reported alongside their corresponding Heat Activation values (Sharwood et al., 2016) (Hermida-Carrera et al., 2016)

Species	K_{cat}	Sc/o	Kcat HA	Sc/o HA	Q10
Triticum aestivum	2.2	100	41.2	24.1	1.78
Hordeum vulgare	2.4	99.2	27.9	21.2	1.48
Avena sativa	2.3	99.9	41.5	23.6	1.79
Oryza sativa	2.1	93.1	46.4	24.6	1.91
Solanum lycopersicum	2.3	92.4	34.6	21.8	1.62
Capsicum annuum	1.9	96	39.2	24.1	1.73
Solanum tuberosum	2	95.4	46.2	24.7	1.91
Ipomoea batatas	2.5	98.5	33.4	22.8	1.60
Coffea arabica	2.1	98.7	39	23.4	1.73
Glycine max	1.5	97	55.2	26.5	2.17
Cucurbita maxima	2.2	98.4	48.7	21.1	1.98
Lactua sativa	2.2	94	33.3	21.2	1.59
Brassica oleracea	2.1	96.2	45.7	21.8	1.90
Spinacia oleracea	2.4	97	48	25.2	1.96
Beta vulgaris	2	101	51.2	19.8	2.05
Urochloa panicoides	5.6	78.3	57.3	22.3	2.23
Megathyrsus maximus	5.3	80.3	50.2	22.2	2.02
Panicum deustum	5	84.8	59.5	20.8	2.30
Panicum milliaceum	2.1	79.9	71.6	28.6	2.72
Panicum coloratum	3.4	84.8	58.3	23.5	2.26
Panicum virgatum	3.3	82.6	58.1	23.8	2.26
Panicum milioides	2.2	92.3	68.4	27.6	2.61
Panicum bisulcatum	2.6	87.7	71.2	29.7	2.71
Saccharum officinarum	3.9	82.2	30.2	25.8	1.53
Zea mays	4.1	87.3	31	24.3	1.54
Setaris viridis	5.9	72.7	56.6	25.3	2.21
Cenchrus ciliaris	6	69.9	54.9	20.3	2.16

There is limited information on the temperature response of Rubisco outside of plants. As a result, the idea of using a fixed function, used to represent the relationship between temperature and Vc_{max} was explored. The Vc_{max} was calculated at 5, 15, 25 and 35^{O}C for a number of Rubiscos reported in the literature alongside their respective HA values (Table 4.2) (Sharwood et al. 2016) (Hermida-Carrera et al., 2016). It is also important to note that there was assumed to be no limitation of Rubisco at higher temperatures for these

calculations unlike in further modelling efforts where Rubisco was assumed to begin to be self-inhibited passed 25°C.

From Figure 4.5 as well as the calculated Q10 values (Table 4.2) it is possible to see that there is significant diversity in the temperature response of Rubisco. Q10 values range from 1.53 in *Saccharum officinarum* to >2.6 in multiple *Panicum spp.* As a result of this, using a fixed function to represent the relationship between temperature and Vc_{max} would not be appropriate. Therefore going forward, when modelling aquatic Rubisco in wheat the native HA for wheat was applied.

In previous studies it has been shown to be possible to significantly reduce the stomatal density of wheat to reduce transpiration rates (Dunn et al., 2019) Based on this concept, a number of studies on *Arabidopsis thaliana* were collated to assess the effect of reducing stomatal density on the Gmax of the wheat plant Figure 4.6.

The results from this show that there is a strong linear correlation (R^2= 0.8457) between that of Gmax and Stomatal density. This shows that it is possible to reduce stomatal density by 50% and get a hypothetical reduction in Gmax by 50% also Figure 4.6.

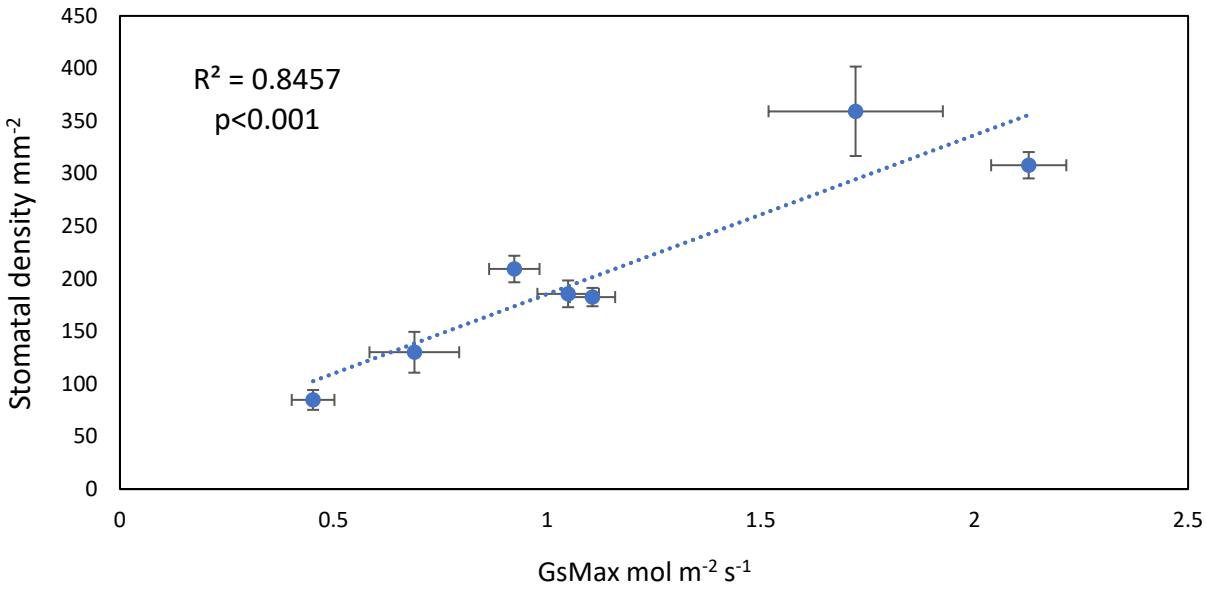


Figure 4.6– The relationship between GsMax and stomatal density was assessed by collating multiple values reported in various studies (Dow et al., 2014), (Lampard et al., 2008), (Hara et al., 2007) on *Arabidopsis thaliana*. A linear function was applied to this relationship through Pearson's rank correlation. The R^2 value represents goodness of fit and p highlights statistical significance of correlation. Errors bars are used to demonstrate standard deviation of reported measurements.

4.3.3 Modelling effects of heterogenous Rubisco expression from aquatic species

The effects of heterogenous expression of aquatic Rubisco in wheat, assuming native expression levels and temperature response were measured over a growing season (Figure 4.7A). For this the carbon assimilation over the growing season was summed giving an overall total for carbon assimilated.

Figure 4.7A and Table 4.3 shows that wheat outperformed the vast majority of aquatic Rubisco species with the exception of a few red algae in terms of carbon assimilated over the growing season. Rubisco derived from *G. monilis* was responsible for the greatest quantity of CO_2 assimilated over the growing season improving expression relative to wheat by 8.76% however this difference was not deemed to be significant (P=0.407). Heterogenous expression of Rubisco, from *O. luteus,* also a form ID Rubisco species, had the greatest negative impact on CO_2 assimilation, reducing total CO_2 assimilated by 63.69% (P<0.001).

CO_2 assimilation over a growing season was also modelled for wheat and heterogenous expression of aquatic Rubisco assuming a 50% reduction in stomatal density and therefore a 50% reduction in Gs_{max}. This showed a reduction but not a proportional reduction in the total Carbon assimilated for wheat going from 850.49 µmol m^{-2} to 734.76 µmol m^{-2}. This was also coupled with a small but significant reduction in C_i concentrations of 37.4 µmol mol^{-1} (P<0.001) within the plant. As well as a 35.59% reduction in the mean transpiration rate (µmol m^{-2} s^{-1}) (Figure 4.5) (Table 4.5).

When the total carbon assimilated in wheat was compared to carbon assimilated through heterogenous expression of aquatic Rubisco with $0.5Gs_{max}$, a different pattern emerged to modelled total assimilation in native Gs_{max}. With native Gs_{max} levels there were only four aquatic Rubiscos that appeared to improve carbon assimilation over a growing season. Opposingly at $0.5Gs_{max}$ there were 13 aquatic Rubiscos that improved assimilation. All these Rubiscos were derived from form ID red algal sources.

Furthermore the relative improvement of carbon assimilation proportional to wheat increased at $0.5Gs_{max}$. For example at native Gs_{max} levels, heterogenous expression of *G. monillis* Rubisco was shown to improve carbon assimilation by 8.75% (Table 4.3). When Gs_{max} was reduced by 50% this percentage increased to 23.58% (Table 4.4) over a growing season. On the other hand heterogenous expression of *T. dentricans* was modelled to have a greater detriment to carbon assimilation with the difference increasing to 69.54% when Gs_{max} was reduced by 50% (Figure 4.7) (Table 4.4).

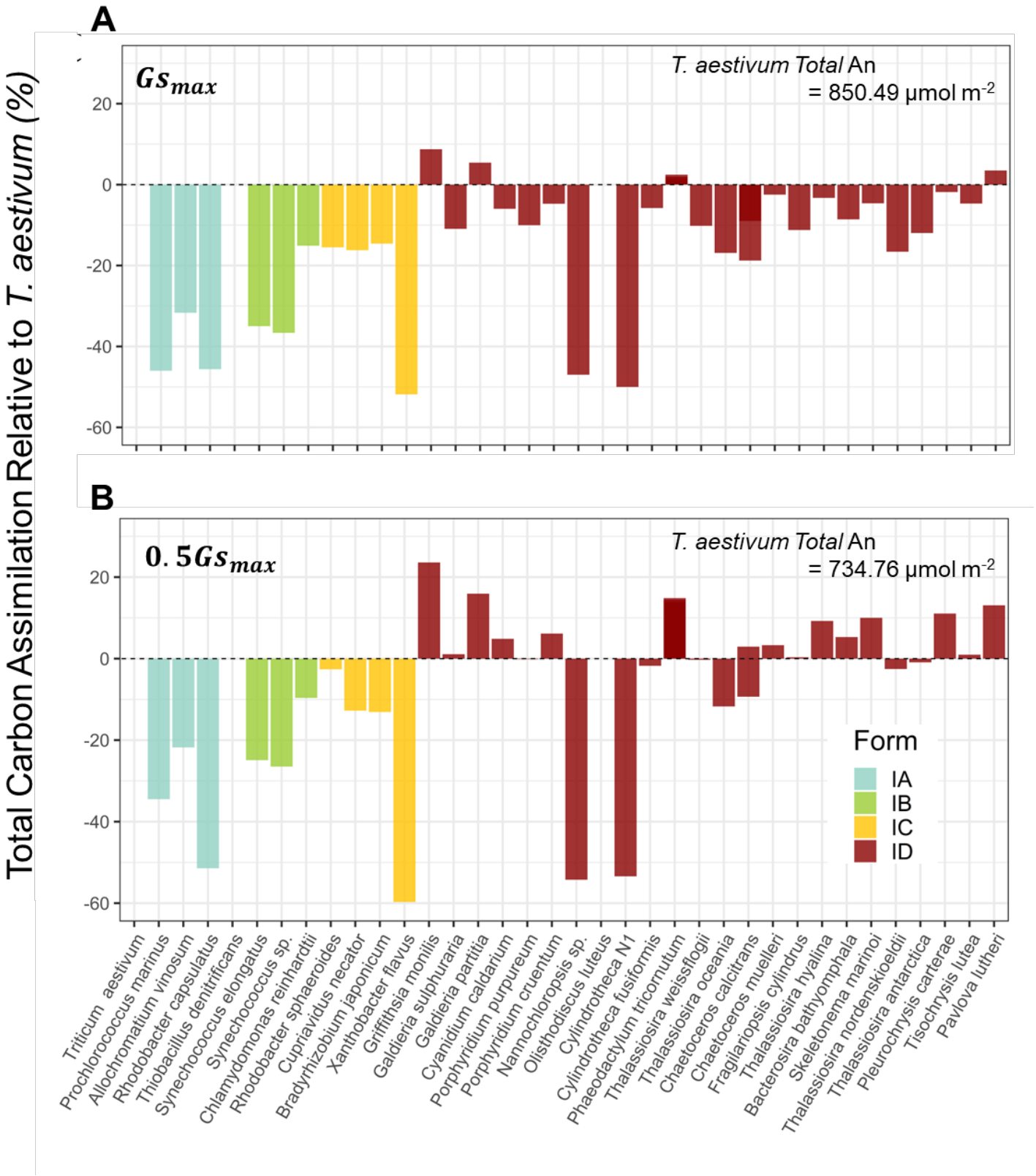


Figure 4.7- A Represents the total carbon assimilation across a growing season relative to wheat simulations with foreign Rubisco derived from the species listed. Relative difference is

represented as a percentage and GsMax was set to native levels. **B** Represents the total carbon assimilation relative to wheat for simulations where Gs_{max} was reduced by 50% to simulate reduced stomatal density as such. Again, wheat Rubisco was exchanged with foreign Rubisco derived from various aquatic sources. The form of which is represented by the colour of the far as described in the figure legend (IA= Turquoise, IB=Green, IC=Orange, ID=Burgandy).

Table 4.3 -Comparison of Total Carbon assimilated by Wheat over a growing season with modelled heterogenous expression of alternative aquatic Rubisco species at normal Gsmax

Species	Form	Total Carbon Assimlation µmol m^{-2}	Relative_Difference relative to *T. aestivum (%)*	P-value
Triticum aestivum	IB	850.487	NA	NA
Prochlorococcus marinus	IA	459.335	-45.992	**0.000**
Allochromatium vinosum	IA	581.144	-31.669	**0.002**
Rhodobacter capsulatus	IA	462.620	-45.605	**0.000**
Thiobacillus denitrificans	IA	330.940	-61.088	**0.000**
Synechococcus elongatus	IB	553.057	-34.972	**0.001**
Synechococcus sp.	IB	538.939	-36.632	**0.001**
Chlamydomonas reinhardtii	IB	722.115	-15.094	0.185
Rhodobacter sphaeroides	IC	718.651	-15.501	0.146
Cupriavidus necator	IC	712.733	-16.197	0.178
Bradyrhizobium japonicum	IC	726.410	-14.589	0.228
Xanthobacter flavus	IC	409.678	-51.830	**0.000**
Griffithsia monilis	ID	924.965	8.757	0.407
Galdieria sulphuraria	ID	757.502	-10.933	0.244
Galdieria partitia	ID	896.551	5.416	0.667
Cyanidium caldarium	ID	799.700	-5.971	0.489
Porphyridium purpureum	ID	765.214	-10.026	0.340
Porphyridium cruentum	ID	810.269	-4.729	0.637
Nannochloropsis sp.	ID	450.889	-46.985	**0.000**
Olisthodiscus luteus	ID	308.823	-63.689	**0.000**
Cylindrotheca N1	ID	425.282	-49.995	**0.000**
Cylindrotheca fusiformis	ID	801.177	-5.798	0.688
Phaeodactylum tricornutum	ID	871.303	2.448	0.379
Thalassiosira weissflogii	ID	763.807	-10.192	0.166
Thalassiosira oceania	ID	706.946	-16.878	0.107
Chaetoceros calcitrans	ID	690.937	-18.760	0.873
Chaetoceros muelleri	ID	829.327	-2.488	0.862
Fragilariopsis cylindrus	ID	754.994	-11.228	0.301
Thalassiosira hyalina	ID	822.483	-3.293	0.665
Bacterosira bathyomphala	ID	777.318	-8.603	0.494
Skeletonema marinoi	ID	811.258	-4.612	0.754
Thalassiosira nordenskioeldii	ID	709.478	-16.580	0.168
Thalassiosira antarctica	ID	748.572	-11.983	0.326
Pleurochrysis carterae	ID	834.924	-1.830	0.968
Tisochrysis lutea	ID	810.722	-4.676	0.711
Pavlova lutheri	ID	880.234	3.498	0.618

P-values were calculated from the median of wheat total carbon assimilation and heterogenic Rubisco assimilation using a Mann-Whitney U test assuming unequal variance.

Table 4.4 -Comparison of Total Carbon assimilated by Wheat over a growing season with modelled heterogenous expression of alternative aquatic Rubisco at half Gs_{max}

Species	Form	Total Carbon Assimilation µmol m^{-2}	Relative_Difference relative to *T. aestivum (%)*	P-value*
Triticum aestivum	IB	734.769	NA	NA
Prochlorococcus marinus	IA	481.320	-34.494	**0.000**
Allochromatium vinosum	IA	574.737	-21.780	**0.003**
Rhodobacter capsulatus	IA	356.857	-51.433	**0.000**
Thiobacillus denitrificans	IA	236.437	-67.822	**0.000**
Synechococcus elongatus	IB	551.684	-24.917	**0.001**
Synechococcus sp.	IB	540.148	-26.487	**0.001**
Chlamydomonas reinhardtii	IB	663.958	-9.637	0.177
Rhodobacter sphaeroides	IC	715.526	-2.619	0.140
Cupriavidus necator	IC	641.008	-12.761	0.167
Bradyrhizobium japonicum	IC	638.533	-13.097	0.214
Xanthobacter flavus	IC	296.119	-59.699	**0.000**
Griffithsia monilis	ID	908.062	23.585	0.397
Galdieria sulphuraria	ID	742.673	1.076	0.252
Galdieria partitia	ID	851.853	15.935	0.652
Cyanidium caldarium	ID	770.526	4.866	0.506
Porphyridium purpureum	ID	733.907	-0.117	0.337
Porphyridium cruentum	ID	779.915	6.144	0.633
Nannochloropsis sp.	ID	336.139	-54.252	**0.000**
Olisthodiscus luteus	ID	223.750	-69.548	**0.000**
Cylindrotheca N1	ID	342.297	-53.414	**0.000**
Cylindrotheca fusiformis	ID	721.796	-1.766	0.669
Phaeodactylum tricornutum	ID	840.750	14.424	0.364
Thalassiosira weissflogii	ID	732.721	-0.279	0.153
Thalassiosira oceania	ID	648.551	-11.734	0.199
Chaetoceros calcitrans	ID	666.081	-9.348	0.870
Chaetoceros muelleri	ID	759.029	3.302	0.865
Fragilariopsis cylindrus	ID	737.207	0.332	0.376
Thalassiosira hyalina	ID	802.658	9.240	0.781
Bacterosira bathyomphala	ID	773.699	5.298	0.479
Skeletonema marinoi	ID	808.308	10.008	0.744
Thalassiosira nordenskioeldii	ID	716.047	-2.548	0.212
Thalassiosira antarctica	ID	728.149	-0.901	0.313
Pleurochrysis carterae	ID	816.089	11.067	0.898
Tisochrysis lutea	ID	741.826	0.960	0.707
Pavlova lutheri	ID	830.830	13.074	0.611

P-values were calculated from the median of wheat total carbon assimilation and heterogenic Rubisco assimilation using a Mann-Whitney U test assuming unequal variance.

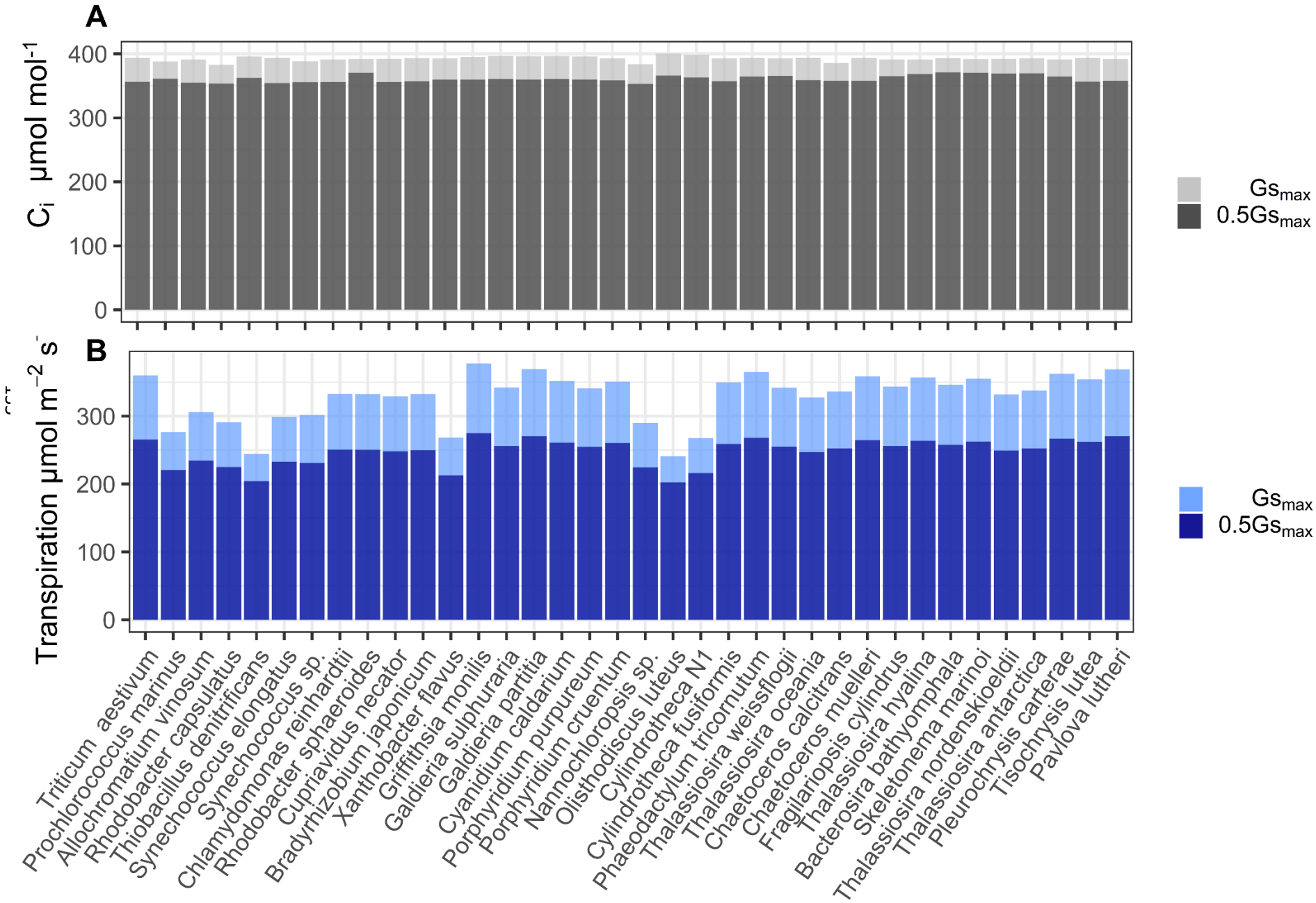

A
B
C_i µmol mol^{-1}
Transpiration µmol m^{-2} s^{-1}
0
100
200
300
400
Gs_{max}
$0.5Gs_{max}$
Triticum aestivum
Prochlorococcus marinus
Allochromatium vinosum
Rhodobacter capsulatus
Thiobacillus denitrificans
Synechococcus elongatus
Synechococcus sp.
Chlamydomonas reinhardtii
Rhodobacter sphaeroides
Cupriavidus necator
Bradyrhizobium japonicum
Xanthobacter flavus
Griffithsia monilis
Galdieria sulphuraria
Galdieria partitia
Cyanidium caldarium
Porphyridium purpureum
Porphyridium cruentum
Nannochloropsis sp.
Olisthodiscus luteus
Cylindrotheca N1
Cylindrotheca fusiformis
Phaeodactylum tricornutum
Thalassiosira weissflogii
Thalassiosira oceania
Chaetoceros calcitrans
Chaetoceros muelleri
Fragilariopsis cylindrus
Thalassiosira hyalina
Bacterosira bathyomphala
Skeletonema marinoi
Thalassiosira nordenskioeldii
Thalassiosira antarctica
Pleurochrysis carterae
Tisochrysis lutea
Pavlova lutheri

Figure 4.8- A is the mean daily C_i (µmol mol^{-1}) modelled in wheat across a growing season. Additionally heterogenic expression of foreign Rubisco was modelled in a wheat system, the resulting mean Ci values are also reported. Light grey bars represent C_i concentrations under native Gs_{max} conditions, dark grey bars represent C_i concentrations when Gs_{max} was reduced by 50%. **B** Is the total of the mean daily Transpiration (µmol m^{-2} s^{-1}) across a growing season for wheat and wheat with heterogenic Rubisco from aquatic algae. The light blue bars represent transpiration rates under native Gs_{max} conditions, dark blue bars represent transpiration rates when Gs_{max} was reduced by 50%.

Species	Form	Mean Transpiration µmol m^{-2} s^{-1} Gsmax	Mean Transpiration µmol m^{-2} s^{-1} 0.5Gsmax	P-value*	Ci µmol mol^{-1} Gsmax	Ci µmol mol^{-1} 0.5Gsmax	P-value*
Triticum aestivum	IB	1.875	1.383	0.058	393.720	356.272	**0.000**
Prochlorococcus marinus	IA	1.439	1.148	0.064	387.865	361.084	**0.000**
Allochromatium vinosum	IA	1.595	1.220	0.097	390.688	354.799	**0.000**
Rhodobacter capsulatus	IA	1.516	1.172	0.499	382.781	353.438	**0.000**
Thiobacillus denitrificans	IA	1.272	1.065	0.184	395.612	362.200	**0.000**
Synechococcus elongatus	IB	1.557	1.212	0.063	393.633	354.333	**0.000**
Synechococcus sp.	IB	1.572	1.202	0.055	388.299	355.596	**0.000**
Chlamydomonas reinhardtii	IB	1.733	1.305	0.099	391.191	356.013	**0.000**
Rhodobacter sphaeroides	IC	1.732	1.304	0.105	391.632	370.332	**0.000**
Cupriavidus necator	IC	1.715	1.293	0.090	391.832	356.168	**0.000**
Bradyrhizobium japonicum	IC	1.732	1.302	0.088	393.334	357.133	**0.000**
Xanthobacter flavus	IC	1.397	1.108	0.244	392.832	359.512	**0.000**
Griffithsia monilis	ID	1.967	1.433	**0.039**	394.560	359.423	**0.000**
Galdieria sulphuraria	ID	1.782	1.334	**0.044**	396.648	360.853	**0.000**
Galdieria partitia	ID	1.923	1.408	**0.036**	396.292	359.851	**0.000**
Cyanidium caldarium	ID	1.832	1.359	**0.039**	396.925	360.894	**0.000**
Porphyridium purpureum	ID	1.777	1.327	0.053	396.060	359.815	**0.000**
Porphyridium cruentum	ID	1.828	1.355	0.053	393.036	358.517	**0.000**
Nannochloropsis sp.	ID	1.511	1.169	0.477	383.781	352.817	**0.000**
Olisthodiscus luteus	ID	1.255	1.055	0.101	400.130	365.994	**0.000**
Cylindrotheca N1	ID	1.393	1.126	0.083	398.257	363.018	**0.000**
Cylindrotheca fusiformis	ID	1.821	1.350	0.054	393.013	357.222	**0.000**
Phaeodactylum tricornutum	ID	1.782	1.329	0.087	393.542	365.293	**0.000**
Thalassiosira weissflogii	ID	1.705	1.286	0.085	393.089	359.069	**0.000**
Thalassiosira oceania	ID	1.751	1.315	0.108	393.709	358.021	**0.000**
Chaetoceros calcitrans	ID	1.867	1.378	0.059	385.699	357.931	**0.000**
Chaetoceros muelleri	ID	1.887	1.388	0.056	393.510	364.962	**0.000**
Fragilariopsis cylindrus	ID	1.789	1.334	0.088	390.967	364.935	**0.000**
Thalassiosira hyalina	ID	1.859	1.373	0.058	390.963	368.528	**0.000**
Bacterosira bathyomphala	ID	1.805	1.342	0.060	393.322	370.755	**0.000**
Skeletonema marinoi	ID	1.849	1.368	0.059	391.597	370.316	**0.000**
Thalassiosira nordenskioeldii	ID	1.728	1.298	0.060	391.796	369.020	**0.000**
Thalassiosira antarctica	ID	1.759	1.314	0.057	393.286	369.588	**0.000**
Pleurochrysis carterae	ID	1.888	1.390	0.057	391.327	364.545	**0.000**
Tisochrysis lutea	ID	1.844	1.365	0.060	393.466	356.472	**0.000**
Pavlova lutheri	ID	1.922	1.408	0.053	391.765	357.957	**0.000**

*P-values were calculated from the median of Transpiration or C_i values under Gsmax and 0.5 Gsmax conditions. A Mann-Whitney U test assuming unequal variance was used to assess this difference.

4.4 Discussion

4.4.1 Model evaluation

This study builds on the models developed by (Iqbal et al., 2021) and previous implementations of Earth systems models used to simulate photosynthesis (Chen and Blankenship, 2011) (Houborg et al., 2012). Like previous ESM studies it continues to demonstrate a high level of performance in predicting carbon assimilation over growing seasons.

Modelling efforts to predict the impact of the heterogenous expression of aquatic Rubisco on wheat photosynthesis has been limited. (Zhu et al., 2004) demonstrated that the heterogenous expression of *G. monillis* may result in a 25% increase in carbon assimilated. However this study was a snapshot of a single day in a growing season and the temperature was assumed to be an optimal 25^{O}C (Zhu et al., 2004). *G. monillis* was also examined in this study as it is widely regarded as the most efficient Rubisco found to date due to its high specificity to K_{cat} ratio). However the potential improvement in carbon assimilation over a growing season was calculated to be closer to 8% (Figure 4.8).

A significant limitation of this study is highlighted by the predefined fixed Carbon: Nitrogen ratios used as well as the fixed Rubisco:TSP content. We previously demonstrated (Figure 2.8) that Rubisco large and small subunit expression has a highly nuanced relationship with the environment and temperature, varying significantly across parameters. (Wu et al., 2019) has made efforts to correct for this over-simplification of nitrogen budgets. In (Wu et al., 2019) they successfully combine photosystems models with APSIMv.7.8 yield models. The APSIM component allows for dynamic estimates of specific leaf nitrogen levels based on developmental and environmental conditions (Wu et al., 2019), (Wu et al., 2023).

Another significant limitation stems from the lack of HA values for Rubisco measured from aquatic organisms. Comparing the Vc_{max} of Rubisco with accompanying HA values demonstrated a greater variability in Q_{10} values than was previously considered for Rubisco (Sage, 2002). As the rate of Rubisco has an exponential relationship with temperature the specific effect of temperature on carbon assimilation is amplified.

In this study a significant modification made to the model developed by (Iqbal et al., 2021) was the fixing of J_{max} to native values from wheat. ESM models (Kattge and Knorr, 2007) (Medlyn et al., 2002) estimate J_{max} to be 1.67 x Vc_{max}. In the model developed by Iqbal et al. 2021 Vc_{max} changed based on the heterogenous expression of Rubisco and J_{max} increased/ decreased proportionally. In this study J_{max} was fixed for the native value of wheat and Vc_{max} depended on Rubisco kinetic properties. As a result of the fixation of Jmax to native conditions, percentage increases in carbon assimilation in this model are significantly lower than those demonstrated by the heterogenous expression of plant Rubisco in (Iqbal et al., 2021).

4.4.2 Heterogenous expression

Rubisco is historically considered the key bottle neck in photosynthesis. However recent modelling efforts by (Busch, 2020), (Wu et al., 2019), (Wu et al., 2023) have demonstrated that heterogenous expression of Rubisco may not be as productive as previously envisioned (Zhu et al., 2004).

What appears to be imperative is a more holistic response to improving photosynthesis, focusing on multiple aspects of photosynthesis simultaneously to improve overall production. (Busch, 2020) demonstrate that under 'optimal conditions' Rubisco is not the limiting factor and that photorespiratory pathways can be a necessary energy sink for the plant. Additionally, (Wu et al., 2023) demonstrate that a combinatorial improvement of Rubisco, mesophyll conductance and the electron transport chain, are necessary to improve overall Carbon assimilation and yields.

In this study we demonstrate that the previously calculated improvements to carbon assimilation achieved through heterogenous expression of Rubisco are overexaggerated (Zhu et al., 2004, Iqbal et al., 2021) (Figure 4.3, Figure 4.5). Despite this it is still possible to achieve up to 8% increase in carbon assimilation in wheat through heterogenic expression of *G. monillis* which when considered over an entire field could result in significant improvements to yields.

In this study we also demonstrated that many aquatic Rubiscos, if expressed in wheat would result in a detriment to growth, still assuming native expression levels. This is due to the fact that many aquatic Rubiscos, particularly those derived from bacterial sources are often very

fast but lack specificity (Davidi et al., 2020). Within these microbial organisms the kinetic properties are considered advantageous due to presence of highly efficient CCMs allowing the accumulation of CO_2 to high concentrations around the Rubisco enzyme (Sun et al., 2022).

Previous studies have shown that it is possible to express form IA Rubisco in tobacco (Chen et al., 2022) and carbon assimilation levels can be comparable to that of the wild type despite expression only being at 40% of native levels (Chen et al., 2022). However, to achieve this plants must be supplemented with a 1% CO_2 environment, artificially increasing C_i levels.

However, when hypothetically expressed within wheat as wheat examined here, under ambient CO_2 conditions, the lack of specificity becomes a detriment.

The few Rubisco species that were modelled to improve carbon assimilation in wheat were all derived from red algae (containing Form ID Rubisco). Form ID Rubisco has been demonstrated to frequently break the canonical Rubisco trade-off theory, being commonly found to have far higher specificity than would be expected from its K_{cat} (Oh et al., 2023), (Flamholz et al., 2019), (Young et al., 2016). As a result of this higher specificity Carbon assimilation can be improved in crops that lack a CCM such as wheat.

4.4.3 Form ID allows maintenance of carbon assimilation whilst reducing transpiration

As outline above the efficacy of heterogenous expression of Rubisco to improve Carbon assimilation in crops is overstated. However a potential application of heterogenous Rubisco expression may be in improving water usage within crops.

The concept of significantly reducing stomatal density in wheat has been demonstrated by (Dunn et al., 2019), (Caine et al., 2019) (Xie et al., 2012), (Liu et al., 2015). Through manipulating expression of epidermal patterning factors it is possible to achieve stomatal density reductions by >50%. The result of this being proportional reductions in Gs_{max} and significant increases in stomatal resistance (Bertolino et al., 2019).

A similar concept was applied in this modelling study, reducing the Gs_{max} of wheat by 50% as if one had reduced stomatal density by such. As a result of this change, modelled mean transpiration levels in wheat dropped >30%. However, a similarly proportional decline in C_i

levels observed or carbon assimilation over the growing season was not found when wheat Gs_{max} and wheat $0.5Gs_{max}$ were compared. These results are concurrent with the *in vivo* experiments in wheat Dunn et al. 2019.

When the reduced stomatal density was compared with heterogenous rubisco simulations an interesting pattern emerged. Under Gs_{max} conditions the greatest increase in carbon assimilation was achieved through heterogenous expression of *G. monillis* totalling a net increase of 8%. However, at 50% Gs_{max} this difference in assimilation between wheat and *G. monillis* increased to 24%. This is a theme that is observed across a number of red algal Rubiscos with total assimilation increasing, relative to wheat when Gs_{max} was reduced by 50%. This discrepancy in relative assimilation must be attributed to the higher specificity levels in red algae. This improvement is reflected in the ACI curves comparing native wheat Rubisco and heterogenic expression of *G. monillis* Rubisco. *G. monillis* being more specific and faster than wheat Rubisco resulted in an improvement in both the A_c and A_j limited rate of photosynthesis. (Wu et al., 2023) demonstrated that an improvement in the Rubisco rate alone only improves the A_c limited rate of the enzyme and as result has negligible effect on carbon assimilation at typical C_i levels. Improving A_j has a more pronounced effect on assimilation as this difference is observed at C_i concentrations >300 µmol mol^{-1} (Wu et al., 2023). As the mean C_i of wheat over the growing season at Gsmax and 0.5 Gsmax was found to be 393.72 and 356.27 (µmol mol^{-1}) respectively, this would explain why the transgenic expression of faster, less specific Rubisco from namely cyanobacteria had a detrimental effect on carbon assimilation.

Considering the ACI curve of wheat and *G. monillis* further, at increasing C_i concentrations over 300 µmol mol^{-1}, the discrepancy between *G. monillis* and wheat assimilation begins to narrow. This explains why the relative difference between wheat and *G. monillis* at 0.5Gsmax is 24% and at native Gsmax the difference is 8%. This is due to the fact that a more specific, faster Rubisco improves carbon assimilation proportionally higher at reduced C_i concentrations (observed as the reduced C_i concentrations at 0.5 Gsmax).

It is important to note that carbon assimilation does not necessarily equate to increased yields despite the aforementioned study by (Dunn et al., 2019) showing that reducing stomatal density not significantly reduce yields. Additionally, transpiration is a necessary process for the plant, maintaining leaf temperatures up to 13°C below ambient conditions

(Deva et al., 2020). This study focused on winter wheat being harvested in early July (Sattorini et al., 2016) meaning that temperatures rarely exceeded 25^{o}C. However crops grown through summer months may require irrigation and a high rate of transpiration to maintain homeostasis within the plant.

4.5 Conclusion and future prospects

This study builds on previous ESM models (Iqbal et al. 2021), (Chen and Blankenship, 2011) (Houborg et al., 2012) which have been demonstrated to be incredibly useful resources in modelling photosynthesis over extended periods. It also highlights the fact the heterogenous expression of Rubisco can improve carbon assimilation over a growing season however previous gains in carbon assimilation have been overstated.

Despite this, heterogenous expression of *G. monillis* Rubisco, when coupled with reduced stomatal density has significantly improved water use efficiency as well as amplified improvements in carbon assimilation, relative to wheat. This is due to the high specificity of *G.monillis* Rubisco allowing it to perform more effectively under reduced C_i levels as a result of reduced Gsmax.

Previous studies have demonstrated the potential improvements that red algal Rubisco could make on improving crop yields however heterogenous expression of red Rubisco in green systems is still a significant stumbling block in this process. Previous efforts have failed to successfully express form ID Rubisco in plants (Whitney et al., 2001), (Lin et al., 2018). Recent efforts by (Zhou et al., 2023) have shown that it is possible to use a form IC Rubisco chassis with modified form ID Rubisco parts for successful expression in Tobacco. However, the illusive red algal chaperones, facilitating complete heterogenous Rubisco expression, are still illuding researchers to this day. Even with the discovery of the red algal chaperones, achieving native expression levels of a foreign Rubisco is a further challenge, but one that would be worthwhile based on the simulations conducted in this study.

A future world that is hotter and more arid in many parts of the world, possess an existential risk to human food security. This study demonstrates that the hypothetical coupling of reduced stomatal density and heterogenous expression of red type Rubisco could help

alleviate this challenge simultaneously increasing water use efficiency and carbon assimilation.

5.1 General Discussion

Rubisco is the enzyme central to photosynthesis fixing carbon to the five carbon RuBP for the synthesis of more complex sugars. Rubisco is the dominant carbon fixing enzyme on the planet being responsible for 250 billion tonnes per year globally (Bracher et al., 2017) of which 50% of that is taken up by marine organisms (Irion et al. 2021). Despite this importance, much of the focus on Rubisco has been within land plants and the need for its optimisation within economically important crops. This is because Rubisco is considered to be an inefficient enzyme with a slow catalytic rate and promiscuity, frequently incorrectly binding O_2. Despite the prevalence of Rubisco in marine environments and improved Rubisco kinetics that we observed in aquatic systems (Bar-On and Milo, 2019) these environments still remain underexploited for their diversity. As a result of this the diversity, ecological importance and potential application for improving crop yields of marine Rubisco is the focus of this study.

Marine systems are vulnerable to climate change experiencing inflated temperature rises, especially within polar regions (Rantanen et al., 2022) and this temperature rise has knock-on effects shifting the marine biochemical environment. In this study we explored the effects that temperature and other environmental parameters have on the expression of photosynthetic genes within marine systems. The most notable factor driving expression patterns was environmental temperature with almost all photosynthetic gene expression being significantly correlated with temperature. In opposition to the consensus, photosynthetic genes corresponding to the photosystem architecture were negatively correlated with temperature differing significantly in expression levels between tropical and polar environments. From this negative correlation we hypothesised that this was a necessary mechanism in balancing flux of energy from photosystems to the slow catalytic rates of enzymes at low temperature in order to reduce reactive oxygen species build up. A more simple explanation may be the fact that many of the metagenomic samples were taken during austral summer. This is a time of maximal productivity in polar waters and as a result this explains the higher expression of photosystem genes in this period.

When Rubisco abundance and expression was assessed. It was shown that form ID Rubisco sequences were the dominant form in polar environments and alpha-cyanobacterial form IA Rubisco dominated tropical environments. This corresponds with a number of previous

studies assessing taxonomic abundance of these regions (Cabello-Yeves et al., 2022) When Rubisco and Rubiscosome genes were considered a highly nuanced expression pattern emerged. This deviates from the established trend in form IB land plants where Rubisco expression decreases with rising temperature (Ohba et al., 2000), (Cavanagh et al., 2023), (Devos et al., 1998), (Peng et al., 2021). The driving hypothesis between this negative correlation found in land plants is that Rubisco concentration must increase under cold temperatures to negate the effects of slowed kinetic action. Despite this form IA, IC and ID were all shown to be positively correlated with temperature in the earth's seas and oceans. A possible explanation for this positive correlation with temperature is that there is environmental adaption of the Rubisco enzyme within these species to overcome slow kinetic constraints in cold climates. This was explored in chapter 3.

Rubisco sequences from form IA and ID organisms were the focus due to the high abundance in polar and tropical environments. A Gaussian process model built on sequences extracted in this study highlighted that it was possible to predict environmental temperature from sequence structure. This was particularly true when the ESM-2 transformer was applied to sequences as a numerical representation of 3D structure (Lin et al., 2022). This model points to underlying patterns in the sequence data that were indicative of environmental adaption.

Despite this, previous experiments have shown evidence of environmental evolution within form IB land plants, but have failed to demonstrate environmental evolution in form IA and IB microorganisms (Kapralov and Filatov, 2007), (Goudet et al., 2020). PAML models (Yang, 2007) have been the basis for this assessment of selection however the architecture of the PAML site-models requires a consensus of positive selection across all lineages and thus creates a bias for smaller datasets (Murrell et al., 2012). Additionally (Bouvier et al., 2021) argues that evolution in Rubisco is under represented by the phylogenetic bias observed across Rubisco studies. Therefore PAML may be inappropriate when analysing evolution across clades from within a single Rubisco form examined here. As a result MEME selection (Murrell et al., 2012) was used in conjunction with RELAX selection (Wertheim et al., 2015) programmes implemented. This allows for variation across lineages.

Within form IA organisms RELAX (Wertheim et al., 2015) showed increased selection pressure on cyanobacterial lineages relative to proteobacterial lineages. This intensification

of selection may be explained by the ubiquitous presence of carboxysomes in alpha-cyanobacteria (Cabello-Yeves et al., 2022) and the lack of in many proteobacterial species possessing CbbQO form IA operons. This discrepancy in carboxysome presence was also detected by the random forest model highlighting residues of the form IA N-terminus which are only found in proteobacteria lacking carboxysomes (Badger and Bek, 2008). We can posit that the ubiquity of carboxysomes in cyanobacteria, results in overall increase in epistatic interactions between Rubisco and carboxysome shell proteins and therefore necessitates a greater level of selection pressure applied to the form IA *rbcL* genes.

Unlike within form IA organisms, there was no phylogenetic link to the intensification or relaxation of selection pressure in form ID Rubisco species. However there was widespread positive selection on the *rbcS* gene at loci that were in close contact with the RbcL protein. The majority of these positively residues were located on the ßE-ßF loop at the carboxy-terminal of the RbcS protein which were also highlighted as indicative of 'warm' and 'cold' sequences by the random forest model. This is significant as this extension of the RbcS protein is not found within form IB or IA RbcS proteins. There has been a significant focus in the literature around form ID Rubisco as many species break the canonical trade-off of rate and specificity with specificity far higher than would be expected (Oh et al., 2023). It is known that the specificity of Rubisco is derived in the large part from interactions with the small subunit (observed as differences in specificity between form II and form I Rubisco). It is also true that the most significant structural difference between red and green type RbcS proteins is visualised as the extended ßE-ßF loop of red Rubisco which has a highly invasive interation with the axial pore of Rubisco (Joshi et al., 2015). Therefore it would not be grandiose to hypothesise that the high specificity of form ID Rubisco is derived from the interactions of the ßE-ßF loop with the large subunit (which does not differ significantly between red and green Rubisco species). This is a hypothesis that requires further examination but to date investigation remains challenging due to our inability to successfully express form ID Rubisco in a heterologous system. This is due to the absence of knowledge on the hypothetical red-type Rubisco chaperone which appears to be required for Rubisco assembly within green systems (Oh et al., 2023). Additionally due to the contiguous nature of Rubisco genes located within the chloroplast of red organisms, nuclear transformation of existing red systems such as the well-established *Phaeodactylum tricornutum* is redundant.

Recent efforts by (Zhou et al., 2023) showed that it was possible to express form IC Rubisco within Tobacco and modify loop 6 with elements from *G. monillis*, improving the specificity of the heterologously expressed enzyme (Zhou et al., 2023). A similar approach, modifying the readily expressed form IC chassis with *G. monillis* ßE-ßF loops from the small subunit may be a viable option.

Most importantly efforts to heterlogously express form ID Rubisco in green systems are with the assumption that it will significantly improve crop yields. Zhu et al. 2004 predicted that heterologous expression of *G. monillis* could improve carbon assimilation by in excess of 30% in certain crop species. In this study we modelled the effects of expressing a number of algal and bacterial from aquatic environments including *G. monillis*. Opposingly to previous findings the maximal gain in carbon assimilation was found to be closer to 8% across an entire growing season in winter wheat. This discrepancy in improvements is due to the fact that in Zhu et al. 2004 assumed optimal light and temperature. Under suboptimal conditions (cold and low light) improvements in net carbon assimilation are not solely limited by the rate of Rubisco. Greater improvements in carbon assimilation may be observed in summer crops.

In this study we demonstrated that greater improvements in relative carbon assimilation can be observed when we simultaneously reduce Gsmax through reducing stomatal density. This lowers the internal CO_2 environment in which a more specific Rubisco can perform more efficiently. *G. monillis* Rubisco has the highest specificity to rate ratio currently observed. As a result the greatest modelled improvements in wheat crop with reduced stomatal density were observed through the heterologous expression of *G. monillis*. This modelling effort shows that it is possible to simultaneously improve water use efficiency and improve carbon assimilation through the heterologous expression of form ID Rubisco.

Conclusions and prospects 5.2

This study examines Rubisco from marine systems and highlights its diversity in expression patterns and evolution across forms and phylogenetic lineages from different environments. Most importantly this study demonstrates an impetus for further research into Rubisco from marine environments. Rubisco derived from form ID organisms showed there was the

potential to increase plant productivity and reduce water usage in economically important crops.

Future work must continue to kinetically categorise marine Rubisco with a focus on form ID organisms due to their heightened specificity. Alongside this research we must continue to find the illusive red-type chaperones to allow heterologous expression in green systems. In a future world where climate change continues to increase global temperatures and the aridity of large swathes of crop land, this study highlights an engineering opportunity to maintain high food productivity under increasing environmental pressures.

6. References

ACINAS, S. G., SÁNCHEZ, P., SALAZAR, G., CORNEJO-CASTILLO, F. M., SEBASTIÁN, M., LOGARES, R., ROYO-LLONCH, M., PAOLI, L., SUNAGAWA, S., HINGAMP, P., OGATA, H., LIMA-MENDEZ, G., ROUX, S., GONZÁLEZ, J. M., ARRIETA, J. M., ALAM, I. S., KAMAU, A., BOWLER, C., RAES, J., PESANT, S., BORK, P., AGUSTÍ, S., GOJOBORI, T., VAQUÉ, D., SULLIVAN, M. B., PEDRÓS-ALIÓ, C., MASSANA, R., DUARTE, C. M. & GASOL, J. M. 2021. Deep ocean metagenomes provide insight into the metabolic architecture of bathypelagic microbial communities. Communications Biology, 4, 604.

ADENAN, N. S., YUSOFF, F. M. & SHARIFF, M. 2013. Effect of salinity and temperature on the growth of diatoms and green algae. Journal of Fisheries and Aquatic Science, 8, 397.

ADLER, L., DÍAZ-RAMOS, A., MAO, Y., PUKACZ, K. R., FEI, C. & MCCORMICK, A. J. 2022. New horizons for building pyrenoid-based CO2-concentrating mechanisms in plants to improve yields. Plant Physiol, 190, 1609-1627.

ALFREIDER, A. & BOGENSPERGER, T. 2018. Specific detection of form IA RubisCO genes in chemoautotrophic bacteria. J Basic Microbiol, 58, 712-716.

ANDERSSON, I. & BACKLUND, A. 2008. Structure and function of Rubisco. Plant Physiology and Biochemistry, 46, 275-291.

ANDERSSON, I. 2008. Catalysis and regulation in Rubisco. Journal of Experimental Botany, 59, 1555-1568.

ANDRALOJC, PAUL J., MADGWICK, PIPPA J., TAO, Y., KEYS, A., WARD, JANE L., BEALE, MICHAEL H., LOVELAND, JANE E., JACKSON, PHIL J., WILLIS, ANTONY C., GUTTERIDGE, S. & PARRY, MARTIN A. J. 2012. 2-Carboxy-D-arabinitol 1-phosphate (CA1P) phosphatase: evidence for a wider role in plant Rubisco regulation. Biochemical Journal, 442, 733-742.

ANDREWS, T. J. & WHITNEY, S. M. 2003. Manipulating ribulose bisphosphate carboxylase/oxygenase in the chloroplasts of higher plants. Archives of Biochemistry and Biophysics, 414, 159-169.

ANDREWS, T. J. 1996. The bait in the Rubisco mousetrap. Nature structural biology, 3, 3-7.

ASHIDA, H., DANCHIN, A. & YOKOTA, A. 2005. Was photosynthetic RuBisCO recruited by acquisitive evolution from RuBisCO-like proteins involved in sulfur metabolism? Research in microbiology, 156, 611-618.

ASSMY, P., EHN, J. K., FERNÁNDEZ-MÉNDEZ, M., HOP, H., KATLEIN, C., SUNDFJORD, A., BLUHM, K., DAASE, M., ENGEL, A. & FRANSSON, A. 2013. Floating ice-algal aggregates below melting Arctic sea ice. PLoS One, 8, e76599.

ATKINSON, N., LEITÃO, N., ORR, D. J., MEYER, M. T., CARMO-SILVA, E., GRIFFITHS, H., SMITH, A. M. & MCCORMICK, A. J. 2017. Rubisco small subunits from the unicellular green alga Chlamydomonas complement Rubisco-deficient mutants of Arabidopsis. New Phytol, 214, 655-667.

BADGER, M. R. & BEK, E. J. 2008. Multiple Rubisco forms in proteobacteria: their functional significance in relation to CO2 acquisition by the CBB cycle. Journal of Experimental Botany, 59, 1525-1541.

BALTAR, F., MARTÍNEZ-PÉREZ, C., AMANO, C., VIAL, M., ROBAINA-ESTÉVEZ, S., REINTHALER, T., HERNDL, G. J., ZHAO, Z., LOGARES, R., MORALES, S. E. & GONZÁLEZ, J. M. 2023. A ubiquitous

gammaproteobacterial clade dominates expression of sulfur oxidation genes across the mesopelagic ocean. Nature Microbiology, 8, 1137-1148.

BANDA, D. M., PEREIRA, J. H., LIU, A. K., ORR, D. J., HAMMEL, M., HE, C., PARRY, M. A., CARMO-SILVA, E., ADAMS, P. D. & BANFIELD, J. F. 2020. Novel bacterial clade reveals origin of form I Rubisco. Nature plants, 6, 1158-1166.

BAR-ON, Y. M. & MILO, R. 2019. The biomass composition of the oceans: a blueprint of our blue planet. Cell, 179, 1451-1454.

BARTON, S., JENKINS, J., BUCKLING, A., SCHAUM, C.-E., SMIRNOFF, N., RAVEN, J. A. & YVON-DUROCHER, G. 2020. Evolutionary temperature compensation of carbon fixation in marine phytoplankton. Ecology letters, 23, 722-733.

BASU, S., GLEDHILL, M., DE BEER, D., PRABHU MATONDKAR, S. & SHAKED, Y. 2019. Colonies of marine cyanobacteria Trichodesmium interact with associated bacteria to acquire iron from dust. Communications biology, 2, 284.

BENOISTON, A.-S., IBARBALZ, F. M., BITTNER, L., GUIDI, L., JAHN, O., DUTKIEWICZ, S. & BOWLER, C. 2017. The evolution of diatoms and their biogeochemical functions. Philosophical Transactions of the Royal Society B: Biological Sciences, 372, 20160397.

BERTAGNOLLI, A. D. & STEWART, F. J. 2018. Microbial niches in marine oxygen minimum zones. Nature Reviews Microbiology, 16, 723-729.

BERTOLINO, L. T., CAINE, R. S. & GRAY, J. E. 2019. Impact of stomatal density and morphology on water-use efficiency in a changing world. Frontiers in plant science, 10, 225.

BHATIA, M. P., KUJAWINSKI, E. B., DAS, S. B., BREIER, C. F., HENDERSON, P. B. & CHARETTE, M. A. 2013. Greenland meltwater as a significant and potentially bioavailable source of iron to the ocean. Nature Geoscience, 6, 274-278.

BLICHER, A., WODZINSKA, K., FIDORRA, M., WINTERHALTER, M. & HEIMBURG, T. 2009. The temperature dependence of lipid membrane permeability, its quantized nature, and the influence of anesthetics. Biophysical journal, 96, 4581-4591.

BÖHNKE, S. & PERNER, M. 2017. Unraveling RubisCO Form I and Form II Regulation in an Uncultured Organism from a Deep-Sea Hydrothermal Vent via Metagenomic and Mutagenesis Studies. Frontiers in Microbiology, 8.

BOLAY, P., SCHLÜTER, S., GRIMM, S., RIEDIGER, M., HESS, W. R. & KLÄHN, S. 2022. The transcriptional regulator RbcR controls ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO) genes in the cyanobacterium Synechocystis sp. PCC 6803. New Phytologist, 235, 432-445.

BONAN, G. 2019. Climate change and terrestrial ecosystem modeling, Cambridge University Press.

BOUVIER, J. W., EMMS, D. M., RHODES, T., BOLTON, J. S., BRASNETT, A., EDDERSHAW, A., NIELSEN, J. R., UNITT, A., WHITNEY, S. M. & KELLY, S. 2021. Rubisco Adaptation Is More Limited by Phylogenetic Constraint Than by Catalytic Trade-off. Molecular Biology and Evolution, 38, 2880-2896.

BRACHER, A., WHITNEY, S. M., HARTL, F. U. & HAYER-HARTL, M. 2017. Biogenesis and Metabolic Maintenance of Rubisco. Annual Review of Plant Biology, 68, 29-60.

BRINKER, A., PFEIFER, G., KERNER, M. J., NAYLOR, D. J., HARTL, F. U. & HAYER-HARTL, M. 2001. Dual Function of Protein Confinement in Chaperonin-Assisted Protein Folding. Cell, 107, 223-233.

BROWN, M. R. 1991. The amino-acid and sugar composition of 16 species of microalgae used in mariculture. Journal of experimental marine biology and ecology, 145, 79-99.

BUCHFINK, B., XIE, C. & HUSON, D. H. 2015. Fast and sensitive protein alignment using DIAMOND. Nature Methods, 12, 59-60.

BUSCH, F. A. 2020. Photorespiration in the context of Rubisco biochemistry, CO2 diffusion and metabolism. The Plant Journal, 101, 919-939.

CABELLO-YEVES, P. J., SCANLAN, D. J., CALLIERI, C., PICAZO, A., SCHALLENBERG, L., HUBER, P., RODA-GARCIA, J. J., BARTOSIEWICZ, M., BELYKH, O. I. & TIKHONOVA, I. V. 2022. α-cyanobacteria possessing form IA RuBisCO globally dominate aquatic habitats. The ISME Journal, 16, 2421-2432.

CAI, F., HEINHORST, S., SHIVELY, J. M. & CANNON, G. C. 2008. Transcript analysis of the Halothiobacillus neapolitanus cso operon. Archives of Microbiology, 189, 141-150.

CAINE, R. S., YIN, X., SLOAN, J., HARRISON, E. L., MOHAMMED, U., FULTON, T., BISWAL, A. K., DIONORA, J., CHATER, C. C. & COE, R. A. 2019. Rice with reduced stomatal density conserves water and has improved drought tolerance under future climate conditions. New Phytologist, 221, 371-384.

CAO, S., ZHANG, W., DING, W., WANG, M., FAN, S., YANG, B., MCMINN, A., WANG, M., XIE, B.-B., QIN, Q.-L., CHEN, X.-L., HE, J. & ZHANG, Y.-Z. 2020. Structure and function of the Arctic and Antarctic marine microbiota as revealed by metagenomics. Microbiome, 8, 47.

CAPÓ-BAUÇÀ, S., IÑIGUEZ, C., AGUILÓ-NICOLAU, P. & GALMÉS, J. 2022. Correlative adaptation between Rubisco and CO(2)-concentrating mechanisms in seagrasses. Nat Plants, 8, 706-716.

CAPONE, D. G., SUBRAMANIAM, A., MONTOYA, J. P., VOSS, M., HUMBORG, C., JOHANSEN, A. M., SIEFERT, R. L. & CARPENTER, E. J. 1998. An extensive bloom of the N2-fixing cyanobacterium Trichodesmium erythraeum in the central Arabian Sea. Marine Ecology Progress Series, 172, 281-292.

CARMO-SILVA, A. E. & SALVUCCI, M. E. 2013. The regulatory properties of Rubisco activase differ among species and affect photosynthetic induction during light transitions. Plant physiology, 161, 1645-1655.

CASPARI, O. D., MEYER, M. T., TOLLETER, D., WITTKOPP, T. M., CUNNIFFE, N. J., LAWSON, T., GROSSMAN, A. R. & GRIFFITHS, H. 2017. Pyrenoid loss in Chlamydomonas reinhardtii causes limitations in CO2 supply, but not thylakoid operating efficiency. J Exp Bot, 68, 3903-3913.

CAVANAGH, A. P., SLATTERY, R. & KUBIEN, D. S. 2023. Temperature-induced changes in Arabidopsis Rubisco activity and isoform expression. J Exp Bot, 74, 651-663.

CAVANAGH, A. P., SOUTH, P. F., BERNACCHI, C. J. & ORT, D. R. 2022. Alternative pathway to photorespiration protects growth and productivity at elevated temperatures in a model crop. Plant Biotechnology Journal, 20, 711-721.

CHEN, M. & BLANKENSHIP, R. E. 2011. Expanding the solar spectrum used by photosynthesis. Trends in Plant Science, 16, 427-431.

CHEN, T., FANG, Y., JIANG, Q., DYKES, G. F., LIN, Y., PRICE, G. D., LONG, B. M. & LIU, L. N. 2022. Incorporation of Functional Rubisco Activases into Engineered Carboxysomes to Enhance Carbon Fixation. ACS Synth Biol, 11, 154-161.

CHEN, T., RIAZ, S., DAVEY, P., ZHAO, Z., SUN, Y., DYKES, G. F., ZHOU, F., HARTWELL, J., LAWSON, T., NIXON, P. J., LIN, Y. & LIU, L.-N. 2022. Producing fast and active Rubisco in tobacco to enhance photosynthesis. The Plant Cell, 35, 795-807.

CLEMENT, R., DIMNET, L., MABERLY, S. C. & GONTERO, B. 2016. The nature of the CO2-concentrating mechanisms in a marine diatom, Thalassiosira pseudonana. New Phytologist, 209, 1417-1427.

COHEN, N. R. 2022. Mixotrophic plankton foraging behaviour linked to carbon export. Nature Communications, 13, 1302.

DAVIDI, D., SHAMSHOUM, M., GUO, Z., BAR-ON, Y. M., PRYWES, N., OZ, A., JABLONSKA, J., FLAMHOLZ, A., WERNICK, D. G., ANTONOVSKY, N., DE PINS, B., SHACHAR, L., HOCHHAUSER, D., PELEG, Y., ALBECK, S., SHARON, I., MUELLER-CAJAR, O. & MILO, R. 2020. Highly active rubiscos discovered by systematic interrogation of natural sequence diversity. The EMBO Journal, n/a, e104081.

DE CORTE, D., MUCK, S., TIROCH, J., MENA, C., HERNDL, G. J. & SINTES, E. 2021. Microbes mediating the sulfur cycle in the Atlantic Ocean and their link to chemolithoautotrophy. Environmental Microbiology, 23, 7152-7167.

DEGEN, G. E., ORR, D. J. & CARMO-SILVA, E. 2021. Heat-induced changes in the abundance of wheat Rubisco activase isoforms. New Phytologist, 229, 1298-1311.

DELMONT, T. O., PIERELLA KARLUSICH, J. J., VESELI, I., FUESSEL, J., EREN, A. M., FOSTER, R. A., BOWLER, C., WINCKER, P. & PELLETIER, E. 2022. Heterotrophic bacterial diazotrophs are more abundant than their cyanobacterial counterparts in metagenomes covering most of the sunlit ocean. The ISME journal, 16, 927-936.

DELMONT, T. O., QUINCE, C., SHAIBER, A., ESEN, Ö. C., LEE, S. T. M., RAPPÉ, M. S., MCLELLAN, S. L., LÜCKER, S. & EREN, A. M. 2018. Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. Nature Microbiology, 3, 804-813.

DEVA, C. R., URBAN, M. O., CHALLINOR, A. J., FALLOON, P. & SVITÁKOVA, L. 2020. Enhanced leaf cooling is a pathway to heat tolerance in common bean. Frontiers in plant science, 11, 19.

DEVOS, N., INGOUFF, M., LOPPES, R. & MATAGNE, R. F. 1998. RUBISCO ADAPTATION TO LOW TEMPERATURES: A COMPARATIVE STUDY IN PSYCHROPHILIC AND MESOPHILIC UNICELLULAR ALGAE. Journal of Phycology, 34, 655-660.

DEVRIES, T. 2022. The Ocean Carbon Cycle. Annual Review of Environment and Resources, 47, 317-341.

DOW, G. J., BERGMANN, D. C. & BERRY, J. A. 2014. An integrated model of stomatal development and leaf physiology. New Phytologist, 201, 1218-1226.

DUNN, J., HUNT, L., AFSHARINAFAR, M., MESELMANI, M. A., MITCHELL, A., HOWELLS, R., WALLINGTON, E., FLEMING, A. J. & GRAY, J. E. 2019. Reduced stomatal density in bread wheat leads to increased water-use efficiency. Journal of Experimental Botany, 70, 4737-4748.

EBDEN, M. 2008. Gaussian processes for regression: A quick introduction. The Website of Robotics Research Group in Department on Engineering Science, University of Oxford, 91, 424-436.

ERB, T. J. & ZARZYCKI, J. 2018. A short history of RubisCO: the rise and fall (?) of Nature's predominant CO2 fixing enzyme. Current Opinion in Biotechnology, 49, 100-107.

ESQUIVEL, M. G., GENKOV, T., NOGUEIRA, A. S., SALVUCCI, M. E. & SPREITZER, R. J. 2013. Substitutions at the opening of the Rubisco central solvent channel affect holoenzyme stability and CO 2/O 2 specificity but not activation by Rubisco activase. Photosynthesis research, 118, 209-218.

FELLER, G. & GERDAY, C. 2003. Psychrophilic enzymes: hot topics in cold adaptation. Nature Reviews Microbiology, 1, 200-208.

FELLER, U., ANDERS, I. & MAE, T. 2008. Rubiscolytics: fate of Rubisco after its enzymatic function in a cell is terminated. Journal of Experimental Botany, 59, 1615-1624.

FISHER, N. L., CAMPBELL, D. A., HUGHES, D. J., KUZHIUMPARAMBIL, U., HALSEY, K. H., RALPH, P. J. & SUGGETT, D. J. 2020. Divergence of photosynthetic strategies amongst marine diatoms. PLoS One, 15, e0244252.

FLAMHOLZ, A. I., PRYWES, N., MORAN, U., DAVIDI, D., BAR-ON, Y. M., OLTROGGE, L. M., ALVES, R., SAVAGE, D. & MILO, R. 2019. Revisiting Trade-offs between Rubisco Kinetic Parameters. Biochemistry, 58, 3365-3376.

FLÜGEL, F., TIMM, S., ARRIVAULT, S., FLORIAN, A., STITT, M., FERNIE, A. R. & BAUWE, H. 2017. The photorespiratory metabolite 2-phosphoglycolate regulates photosynthesis and starch accumulation in Arabidopsis. The Plant Cell, 29, 2537-2551.

FLYNN, K. J., MITRA, A., ANESTIS, K., ANSCHÜTZ, A. A., CALBET, A., FERREIRA, G. D., GYPENS, N., HANSEN, P. J., JOHN, U., MARTIN, J. L., MANSOUR, J. S., MASELLI, M., MEDIĆ, N., NORLIN, A., NOT, F., PITTA, P., ROMANO, F., SAIZ, E., SCHNEIDER, L. K., STOLTE, W. & TRABONI, C. 2019. Mixotrophic protists and a new paradigm for marine ecology: where does plankton research go now? Journal of Plankton Research, 41, 375-391.

FORRYAN, A., NAVEIRA GARABATO, A. C., VIC, C., NURSER, A. J. G. & HEARN, A. R. 2021. Galápagos upwelling driven by localized wind–front interactions. Scientific Reports, 11, 1277.

FRACHEBOUD, Y., JOMPUK, C., RIBAUT, J., STAMP, P. & LEIPNER, J. 2004. Genetic analysis of cold-tolerance of photosynthesis in maize. Plant molecular biology, 56, 241-253.

FROLOV, E. N., KUBLANOV, I. V., TOSHCHAKOV, S. V., LUNEV, E. A., PIMENOV, N. V., BONCH-OSMOLOVSKAYA, E. A., LEBEDINSKY, A. V. & CHERNYH, N. A. 2019. Form III RubisCO-mediated transaldolase variant of the Calvin cycle in a chemolithoautotrophic bacterium. Proceedings of the National Academy of Sciences, 116, 18638-18646.

FRY, B. & WAINRIGHT, S. C. 1991. Diatom sources of ¹³ C-rich carbon in marine food webs. Marine Ecology Progress Series, 76, 149-157.

GALMÉS, J., KAPRALOV, M. V., ANDRALOJC, P. J., CONESA, M. À., KEYS, A. J., PARRY, M. A. J. & FLEXAS, J. 2014. Expanding knowledge of the Rubisco kinetics variability in plant species: environmental and evolutionary trends. Plant, Cell & Environment, 37, 1989-2001.

GAO, F., CHEN, C., ARAB, D. A., DU, Z., HE, Y. & HO, S. Y. 2019. EasyCodeML: A visual tool for analysis of selection using CodeML. Ecology and Evolution, 9, 3891-3898.

GEE, C. W. & NIYOGI, K. K. 2017. The carbonic anhydrase CAH1 is an essential component of the carbon-concentrating mechanism in Nannochloropsis oceanica. Proceedings of the National Academy of Sciences, 114, 4537-4542.

GILLOOLY, J. F., BROWN, J. H., WEST, G. B., SAVAGE, V. M. & CHARNOV, E. L. 2001. Effects of size and temperature on metabolic rate. science, 293, 2248-2251.

GONZÁLEZ-BENÍTEZ, N., GARCÍA-CORRAL, L. S., MORÁN, X. A. G., MIDDELBURG, J. J., PIZAY, M. D. & GATTUSO, J.-P. 2019. Drivers of microbial carbon fluxes variability in two oligotrophic Mediterranean coastal systems. Scientific Reports, 9, 17669.

GOUDET, M. M. M., ORR, D. J., MELKONIAN, M., MÜLLER, K. H., MEYER, M. T., CARMO-SILVA, E. & GRIFFITHS, H. 2020. Rubisco and carbon-concentrating mechanism co-evolution across chlorophyte and streptophyte green algae. New Phytologist, 227, 810-823.

GOUDRIAAN, J. & VAN LAAR, H. 2012. Modelling potential crop growth processes: textbook with exercises, Springer Science & Business Media.

GREENE, R. M., GEIDER, R. J. & FALKOWSKI, P. G. 1991. Effect of iron limitation on photosynthesis in a marine diatom. Limnology and Oceanography, 36, 1772-1782.

GUNN, L. H., MARTIN AVILA, E., BIRCH, R. & WHITNEY, S. M. 2020. The dependency of red Rubisco on its cognate activase for enhancing plant photosynthesis and growth. Proceedings of the National Academy of Sciences, 117, 25890-25896.

HAMMEL, A., SOMMER, F., ZIMMER, D., STITT, M., MÜHLHAUS, T. & SCHRODA, M. 2020. Overexpression of Sedoheptulose-1,7-Bisphosphatase Enhances Photosynthesis in Chlamydomonas reinhardtii and Has No Effect on the Abundance of Other Calvin-Benson Cycle Enzymes. Front Plant Sci, 11, 868.

HANSON, T. E. & TABITA, F. R. 2001. A ribulose-1,5-bisphosphate carboxylase/oxygenase (RubisCO)-like protein from Chlorobium tepidum that is involved with sulfur metabolism and the response to oxidative stress. Proc Natl Acad Sci U S A, 98, 4397-402.

HARA, K., KAJITA, R., TORII, K. U., BERGMANN, D. C. & KAKIMOTO, T. 2007. The secretory peptide gene EPF1 enforces the stomatal one-cell-spacing rule. Genes & development, 21, 1720-1725.

HARPEL, M. R., SERPERSU, E. H., LAMERDIN, J. A., HUANG, Z.-H., GAGE, D. A. & HARTMAN, F. C. 1995. Oxygenation Mechanism of Ribulose-Bisphosphate Carboxylase/Oxygenase. Structure and Origin of 2-Carboxytetritol 1,4-Bisphosphate, a Novel O2-Dependent Side Product Generated by a Site-Directed Mutant. Biochemistry, 34, 11296-11306.

HAUSER, T., BHAT, J. Y., MILIČIĆ, G., WENDLER, P., HARTL, F. U., BRACHER, A. & HAYER-HARTL, M. 2015. Structure and mechanism of the Rubisco-assembly chaperone Raf1. Nature structural & molecular biology, 22, 720-728.

HENSON, S. A., CAEL, B. B., ALLEN, S. R. & DUTKIEWICZ, S. 2021. Future phytoplankton diversity in a changing climate. Nature Communications, 12, 5372.

HERMIDA-CARRERA, C., FARES, M. A., FERNÁNDEZ, Á., GIL-PELEGRÍN, E., KAPRALOV, M. V., MIR, A., MOLINS, A., PEGUERO-PINA, J. J., ROCHA, J., SANCHO-KNAPIK, D. & GALMÉS, J. 2017. Positively selected amino acid replacements within the RuBisCO enzyme of oak trees are associated with ecological adaptations. PLoS One, 12, e0183970.

HERMIDA-CARRERA, C., KAPRALOV, M. V. & GALMÉS, J. 2016. Rubisco Catalytic Properties and Temperature Response in Crops. Plant Physiology, 171, 2549-2561.

HOUBORG, R., CESCATTI, A. & MIGLIAVACCA, M. Constraining model simulations of GPP using satellite retrieved leaf chlorophyll. 2012 IEEE International Geoscience and Remote Sensing Symposium, 2012. IEEE, 6455-6458.

HÜGLER, M. & SIEVERT, S. M. 2011. Beyond the Calvin cycle: autotrophic carbon fixation in the ocean. Annual review of marine science, 3, 261-289.

HUNER, N. & MACDOWALL, F. 1978. Evidence for an in vivo conformational change in ribulose bisphosphate carboxylase–oxygenase from Puma rye during cold adaptation. Canadian journal of biochemistry, 56, 1154-1161.

IQBAL, W. A., LISITSA, A. & KAPRALOV, M. V. 2023. Predicting plant Rubisco kinetics from RbcL sequence data using machine learning. Journal of Experimental Botany, 74, 638-650.

IQBAL, W. A., MILLER, I. G., MOORE, R. L., HOPE, I. J., COWAN-TURNER, D. & KAPRALOV, M. V. 2021. Rubisco substitutions predicted to enhance crop performance through carbon uptake modelling. Journal of Experimental Botany, 72, 6066-6075.

JENKS, A. & GIBBS, S. P. 2000. Immunolocalization and distribution of Form II Rubisco in the pyrenoid and chloroplast stroma of Amphidinium carterae and Form I Rubisco in the symbiont-derived plastids of Peridinium foliaceum (Dinophyceae). Journal of Phycology, 36, 127-138.

JOHNSON, J. & BERRY, J. 2021. The role of cytochrome b6f in the control of steady-state photosynthesis: a conceptual and quantitative model. Photosynthesis Research, 148, 101-136.

JOSHI, J., MUELLER-CAJAR, O., TSAI, Y.-C. C., HARTL, F. U. & HAYER-HARTL, M. 2015. Role of Small Subunit in Mediating Assembly of Red-type Form I Rubisco. Journal of Biological Chemistry, 290, 1066-1074.

JUMPER, J., EVANS, R., PRITZEL, A., GREEN, T., FIGURNOV, M., RONNEBERGER, O., TUNYASUVUNAKOOL, K., BATES, R., ŽÍDEK, A., POTAPENKO, A., BRIDGLAND, A., MEYER, C., KOHL, S. A. A., BALLARD, A. J., COWIE, A., ROMERA-PAREDES, B., NIKOLOV, S., JAIN, R., ADLER, J., BACK, T., PETERSEN, S., REIMAN, D., CLANCY, E., ZIELINSKI, M., STEINEGGER, M., PACHOLSKA, M., BERGHAMMER, T., BODENSTEIN, S., SILVER, D., VINYALS, O., SENIOR, A. W., KAVUKCUOGLU, K., KOHLI, P. & HASSABIS, D. 2021. Highly accurate protein structure prediction with AlphaFold. Nature, 596, 583-589.

KAPRALOV, M. V. & FILATOV, D. A. 2007. Widespread positive selection in the photosynthetic Rubisco enzyme. BMC evolutionary biology, 7, 1-10.

KAPRALOV, M. V., KUBIEN, D. S., ANDERSSON, I. & FILATOV, D. A. 2011. 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.

KAPRALOV, M. V., SMITH, J. A. C. & FILATOV, D. A. 2012. Rubisco evolution in C4 eudicots: an analysis of Amaranthaceae sensu lato. PloS one, 7, e52974.

KATTGE, J. & KNORR, W. 2007. Temperature acclimation in a biochemical model of photosynthesis: a reanalysis of data from 36 species. Plant, cell & environment, 30, 1176-1190.

KEYS, A. J. 1986. Rubisco: Its Role in Photorespiration. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 313, 325-336.

KIKUTANI, S., NAKAJIMA, K., NAGASATO, C., TSUJI, Y., MIYATAKE, A. & MATSUDA, Y. 2016. Thylakoid luminal θ-carbonic anhydrase critical for growth and photosynthesis in the marine diatom Phaeodactylum tricornutum. Proceedings of the National Academy of Sciences, 113, 9828-9833.

KOLESINSKI, P., GOLIK, P., GRUDNIK, P., PIECHOTA, J., MARKIEWICZ, M., TARNAWSKI, M., DUBIN, G. & SZCZEPANIAK, A. 2013. Insights into eukaryotic Rubisco assembly - crystal structures of RbcX chaperones from Arabidopsis thaliana. Biochim Biophys Acta, 1830, 2899-906.

KOSAKOVSKY POND, S. L., POON, A. F., VELAZQUEZ, R., WEAVER, S., HEPLER, N. L., MURRELL, B., SHANK, S. D., MAGALIS, B. R., BOUVIER, D. & NEKRUTENKO, A. 2020. HyPhy 2.5—a customizable platform for evolutionary hypothesis testing using phylogenies. Molecular biology and evolution, 37, 295-299.

KRAUSE-JENSEN, D. & DUARTE, C. M. 2016. Substantial role of macroalgae in marine carbon sequestration. Nature Geoscience, 9, 737-742.

KUPRIYANOVA, E. V., SINETOVA, M. A., MARKELOVA, A. G., ALLAKHVERDIEV, S. I., LOS, D. A. & PRONINA, N. A. 2011. Extracellular β-class carbonic anhydrase of the alkaliphilic cyanobacterium Microcoleus chthonoplastes. Journal of Photochemistry and Photobiology B: Biology, 103, 78-86.

LAMPARD, G. R., MACALISTER, C. A. & BERGMANN, D. C. 2008. Arabidopsis stomatal initiation is controlled by MAPK-mediated regulation of the bHLH SPEECHLESS. Science, 322, 1113-1116.

LI, Y., HALLERMAN, E. M., WU, K. & PENG, Y. 2020. Insect-Resistant Genetically Engineered Crops in China: Development, Application, and Prospects for Use. Annual Review of Entomology, 65, 273-292.

LI, Z., XIN, X., XIONG, B., ZHAO, D., ZHANG, X. & BI, C. 2020. Engineering the Calvin–Benson–Bassham cycle and hydrogen utilization pathway of Ralstonia eutropha for improved autotrophic growth and polyhydroxybutyrate production. Microbial Cell Factories, 19, 228.

LIANG, F. & LINDBLAD, P. 2017. Synechocystis PCC 6803 overexpressing RuBisCO grow faster with increased photosynthesis. Metab Eng Commun, 4, 29-36.

LIN, M. T., OCCHIALINI, A., ANDRALOJC, P. J., PARRY, M. A. & HANSON, M. R. 2014. A faster Rubisco with potential to increase photosynthesis in crops. Nature, 513, 547-550.

LIN, Z., AKIN, H., RAO, R., HIE, B., ZHU, Z., LU, W., COSTA, A. D. S., FAZEL-ZARANDI, M., SERCU, T., CANDIDO, S. & RIVES, A. 2022. Language models of protein sequences at the scale of evolution enable accurate structure prediction. bioRxiv, 2022.07.20.500902.

LIU, A. K., PEREIRA, J. H., KEHL, A. J., ROSENBERG, D. J., ORR, D. J., CHU, S. K., BANDA, D. M., HAMMEL, M., ADAMS, P. D. & SIEGEL, J. B. 2022. Structural plasticity enables evolution and innovation of RuBisCO assemblies. Science advances, 8, eadc9440.

LIU, Y., QIN, L., HAN, L., XIANG, Y. & ZHAO, D. 2015. Overexpression of maize SDD1 (ZmSDD1) improves drought resistance in Zea mays L. by reducing stomatal density. Plant Cell, Tissue and Organ Culture (PCTOC), 122, 147-159.

LOBO, A. K. M., ORR, D. J., GUTIERREZ, M. O., ANDRALOJC, P. J., SPARKS, C., PARRY, M. A. J. & CARMO-SILVA, E. 2019. Overexpression of ca1pase Decreases Rubisco Abundance and Grain Yield in Wheat1 [CC-BY]. Plant Physiology, 181, 471-479.

LOGANATHAN, N., TSAI, Y.-C. C. & MUELLER-CAJAR, O. 2016a. Characterization of the heterooligomeric red-type rubisco activase from red algae. Proceedings of the National Academy of Sciences, 113, 14019-14024.

LOGANATHAN, N., TSAI, Y.-C. C. & MUELLER-CAJAR, O. 2016b. Characterization of the heterooligomeric red-type rubisco activase from red algae. Proceedings of the National Academy of Sciences of the United States of America, 113, 14019-14024.

LØNBORG, C., CARREIRA, C., JICKELLS, T. & ÁLVAREZ-SALGADO, X. A. 2020. Impacts of global change on ocean dissolved organic carbon (DOC) cycling. Frontiers in Marine Science, 7, 466.

LOSH, J. L., YOUNG, J. N. & MOREL, F. M. M. 2013. Rubisco is a small fraction of total protein in marine phytoplankton. New Phytologist, 198, 52-58.

LOVE, M., ANDERS, S. & HUBER, W. 2014. Differential analysis of count data–the DESeq2 package. Genome Biol, 15, 10-1186.

MACKENZIE, T. D., JOHNSON, J. M., COCKSHUTT, A. M., BURNS, R. A. & CAMPBELL, D. A. 2005. Large reallocations of carbon, nitrogen, and photosynthetic reductant among phycobilisomes, photosystems, and Rubisco during light acclimation in Synechococcus elongatus strain PCC7942 are constrained in cells under low environmental inorganic carbon. Archives of Microbiology, 183, 190-202.

MACKINDER, L. C., MEYER, M. T., METTLER-ALTMANN, T., CHEN, V. K., MITCHELL, M. C., CASPARI, O., FREEMAN ROSENZWEIG, E. S., PALLESEN, L., REEVES, G. & ITAKURA, A. 2016. A repeat protein links Rubisco to form the eukaryotic carbon-concentrating organelle. Proceedings of the National Academy of Sciences, 113, 5958-5963.

MAO, Y., CATHERALL, E., DÍAZ-RAMOS, A., GREIFF, G. R., AZINAS, S., GUNN, L. & MCCORMICK, A. J. 2023. The small subunit of Rubisco and its potential as an engineering target. Journal of Experimental Botany, 74, 543-561.

MATSUDA, Y. & KROTH, P. G. 2014. Carbon fixation in diatoms, Springer.

MATSUMURA, H., MIZOHATA, E., ISHIDA, H., KOGAMI, A., UENO, T., MAKINO, A., INOUE, T., YOKOTA, A., MAE, T. & KAI, Y. 2012. Crystal structure of rice Rubisco and implications for activation induced by positive effectors NADPH and 6-phosphogluconate. Journal of molecular biology, 422, 75-86.

MCKAY, R., GIBBS, S. P. & VAUGHN, K. 1991. RuBisCo activase is present in the pyrenoid of green algae. Protoplasma, 162, 38-45.

MEDLYN, B., DREYER, E., ELLSWORTH, D., FORSTREUTER, M., HARLEY, P., KIRSCHBAUM, M., LE ROUX, X., MONTPIED, P., STRASSEMEYER, J. & WALCROFT, A. 2002. Temperature response of parameters of a biochemically based model of photosynthesis. II. A review of experimental data. Plant, Cell & Environment, 25, 1167-1179.

MEI, H., LIAO, Z. H., ZHOU, Y. & LI, S. Z. 2005. A new set of amino acid descriptors and its application in peptide QSARs. Peptide Science: Original Research on Biomolecules, 80, 775-786.

MINODA, A., WEBER, A. P. M., TANAKA, K. & MIYAGISHIMA, S.-Y. 2010. Nucleus-Independent Control of the Rubisco Operon by the Plastid-Encoded Transcription Factor Ycf30 in the Red Alga Cyanidioschyzon merolae. Plant Physiology, 154, 1532-1540.

MIRZABAEV, A., WU, J., EVANS, J., GARCIA-OLIVA, F., HUSSEIN, I. A. G., IQBAL, M. H., KIMUTAI, J., KNOWLES, T., MEZA, F., NEDJROAOUI, D., TENA, F., TÜRKEŞ, M., VÁZQUEZ, R. J. & WELTZ, M. 2019. Desertification. In: SHUKLA, P. R., SKEG, J., CALVO BUENDIA, E., MASSON-DELMOTTE, V., PÖRTNER, H. O., ROBERTS, D. C., ZHAI, P., SLADE, R., CONNORS, S., VAN DIEMEN, S., FERRAT, M., HAUGHEY, E., LUZ, S., PATHAK, M., PETZOLD, J., PORTUGAL PEREIRA, J., VYAS, P., HUNTLEY, E., KISSICK, K., BELKACEMI, M. & MALLEY, J. (eds.) Climate Change and Land: an IPCC special report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems.

MOELLER, H. V., LAUFKÖTTER, C., SWEENEY, E. M. & JOHNSON, M. D. 2019. Light-dependent grazing can drive formation and deepening of deep chlorophyll maxima. Nature communications, 10, 1978.

MUELLER-CAJAR, O. 2017. The Diverse AAA+ Machines that Repair Inhibited Rubisco Active Sites. Frontiers in Molecular Biosciences, 4.

MUELLER-CAJAR, O., STOTZ, M., WENDLER, P., HARTL, F. U., BRACHER, A. & HAYER-HARTL, M. 2011. Structure and function of the AAA+ protein CbbX, a red-type Rubisco activase. Nature, 479, 194-9.

MURRELL, B., WERTHEIM, J. O., MOOLA, S., WEIGHILL, T., SCHEFFLER, K. & KOSAKOVSKY POND, S. L. 2012. Detecting individual sites subject to episodic diversifying selection. PLoS genetics, 8, e1002764.

NI, G., ZIMBALATTI, G., MURPHY, C. D., BARNETT, A. B., ARSENAULT, C. M., LI, G., COCKSHUTT, A. M. & CAMPBELL, D. A. 2017. Arctic Micromonas uses protein pools and non-photochemical quenching to cope with temperature restrictions on Photosystem II protein turnover. Photosynthesis Research, 131, 203-220.

NI, T., JIANG, Q., NG, P. C., SHEN, J., DOU, H., ZHU, Y., RADECKE, J., DYKES, G. F., HUANG, F., LIU, L.-N. & ZHANG, P. 2023. Intrinsically disordered CsoS2 acts as a general molecular thread for α-carboxysome shell assembly. Nature Communications, 14, 5512.

NI, T., SUN, Y., BURN, W., AL-HAZEEM, M. M., ZHU, Y., YU, X., LIU, L.-N. & ZHANG, P. 2022. Structure and assembly of cargo Rubisco in two native α-carboxysomes. Nature Communications, 13, 4299.

OH, Z. G., ANG, W. S. L., POH, C. W., LAI, S.-K., SZE, S. K., LI, H.-Y., BHUSHAN, S., WUNDER, T. & MUELLER-CAJAR, O. 2023. A linker protein from a red-type pyrenoid phase separates with Rubisco via oligomerizing sticker motifs. Proceedings of the National Academy of Sciences, 120, e2304833120.

OHBA, H., STEWARD, N., KAWASAKI, S., BERBERICH, T., IKEDA, Y., KOIZUMI, N., KUSANO, T. & SANO, H. 2000. Diverse response of rice and maize genes encoding homologs of WPK4, an SNF1-related protein kinase from wheat, to light, nutrients, low temperature and cytokinins. Molecular and General Genetics MGG, 263, 359-366.

ORR, D. J., ROBIJNS, A. K. J., BAKER, C. R., NIYOGI, K. K. & CARMO-SILVA, E. 2022. Dynamics of Rubisco regulation by sugar phosphate derivatives and their phosphatases. Journal of Experimental Botany, 74, 581-590.

PARRY, M., MADGWICK, P., PARMAR, S., CORNELIUS, M. & KEYS, A. 1992. Mutations in loop six of the large subunit of ribulose-1, 5-bisphosphate carboxylase affect substrate specificity. Planta, 187, 109-112.

PARTO, S. & LARTILLOT, N. 2018. Molecular adaptation in Rubisco: Discriminating between convergent evolution and positive selection using mechanistic and classical codon models. PLOS ONE, 13, e0192697.

PEARCE, F. G. 2006. Catalytic by-product formation and ligand binding by ribulose bisphosphate carboxylases from different phylogenies. Biochemical Journal, 399, 525-534.

PEERS, G. & PRICE, N. M. 2006. Copper-containing plastocyanin used for electron transport by an oceanic diatom. Nature, 441, 341-344.

PENG, Z., LIU, G. & HUANG, K. 2021. Cold Adaptation Mechanisms of a Snow Alga Chlamydomonas nivalis During Temperature Fluctuations. Frontiers in Microbiology, 11.

PEREIRA, M. B., WALLROTH, M., JONSSON, V. & KRISTIANSSON, E. 2018. Comparison of normalization methods for the analysis of metagenomic gene abundance data. BMC Genomics, 19, 274.

PESANT, S., NOT, F., PICHERAL, M., KANDELS-LEWIS, S., LE BESCOT, N., GORSKY, G., IUDICONE, D., KARSENTI, E., SPEICH, S., TROUBLÉ, R., DIMIER, C., SEARSON, S., ACINAS, S. G., BORK, P., BOSS, E., BOWLER, C., DE VARGAS, C., FOLLOWS, M., GORSKY, G., GRIMSLEY, N., HINGAMP, P., IUDICONE, D., JAILLON, O., KANDELS-LEWIS, S., KARP-BOSS, L., KARSENTI, E., KRZIC, U., NOT, F., OGATA, H., PESANT, S., RAES, J., REYNAUD, E. G., SARDET, C., SIERACKI, M., SPEICH, S., STEMMANN, L., SULLIVAN, M. B., SUNAGAWA, S., VELAYOUDON, D., WEISSENBACH, J., WINCKER, P. & TARA OCEANS CONSORTIUM, C. 2015. Open science resources for the discovery and analysis of Tara Oceans data. Scientific Data, 2, 150023.

PIERELLA KARLUSICH, J. J., BOWLER, C. & BISWAS, H. 2021. Carbon dioxide concentration mechanisms in natural populations of marine diatoms: insights from Tara Oceans. Frontiers in Plant Science, 12, 659.

PINS, B. D., GREENSPOON, L., BAR-ON, Y. M., SHAMSHOUM, M., BEN-NISSAN, R., MILSHTEIN, E., DAVIDI, D., SHARON, I., MUELLER-CAJAR, O., NOOR, E. & MILO, R. 2023. Systematic exploration of bacterial form I rubisco maximal carboxylation rates. bioRxiv, 2023.07.27.550689.

POUDEL, S., PIKE, D. H., RAANAN, H., MANCINI, J. A., NANDA, V., RICKABY, R. E. & FALKOWSKI, P. G. 2020. Biophysical analysis of the structural evolution of substrate specificity in RuBisCO. Proceedings of the National Academy of Sciences, 117, 30451-30457.

PRICE, G. D., PENGELLY, J. J., FORSTER, B., DU, J., WHITNEY, S. M., VON CAEMMERER, S., BADGER, M. R., HOWITT, S. M. & EVANS, J. R. 2013. The cyanobacterial CCM as a source of genes for improving photosynthetic CO2 fixation in crop species. J Exp Bot, 64, 753-68.

RAMAGE, R. T., READ, B. A. & TABITA, F. R. 1998. Alteration of the α helix region of cyanobacterial ribulose 1, 5-bisphosphate carboxylase/oxygenase to reflect sequences found in high substrate specificity enzymes. Archives of biochemistry and biophysics, 349, 81-88.

RANTANEN, M., KARPECHKO, A. Y., LIPPONEN, A., NORDLING, K., HYVÄRINEN, O., RUOSTEENOJA, K., VIHMA, T. & LAAKSONEN, A. 2022. The Arctic has warmed nearly four times faster than the globe since 1979. Communications Earth & Environment, 3, 168.

RAVEN, J. A. & JOHNSTON, A. M. 1991. Mechanisms of inorganic-carbon acquisition in marine phytoplankton and their implications for the use of other resources. Limnology and Oceanography, 36, 1701-1714.

RAVEN, J. A. 2013. Rubisco: still the most abundant protein of Earth? New Phytologist, 198, 1-3.

RAZA, A., RAZZAQ, A., MEHMOOD, S. S., ZOU, X., ZHANG, X., LV, Y. & XU, J. 2019. Impact of Climate Change on Crops Adaptation and Strategies to Tackle Its Outcome: A Review. Plants, 8, 34.

REINFELDER, J. R. 2011. Carbon concentrating mechanisms in eukaryotic marine phytoplankton. Annual review of marine science, 3, 291-315.

RIEBESELL, U., WOLF-GLADROW, D. & SMETACEK, V. 1993. Carbon dioxide limitation of marine phytoplankton growth rates. Nature, 361, 249-251.

ROGERS, A., MEDLYN, B. E., DUKES, J. S., BONAN, G., VON CAEMMERER, S., DIETZE, M. C., KATTGE, J., LEAKEY, A. D., MERCADO, L. M. & NIINEMETS, Ü. 2017. A roadmap for improving the representation of photosynthesis in Earth system models. New Phytologist, 213, 22-42.

ROMERO, P. A., KRAUSE, A. & ARNOLD, F. H. 2013. Navigating the protein fitness landscape with Gaussian processes. Proceedings of the National Academy of Sciences, 110, E193-E201.

ROYO-LLONCH, M., SÁNCHEZ, P., RUIZ-GONZÁLEZ, C., SALAZAR, G., PEDRÓS-ALIÓ, C., SEBASTIÁN, M., LABADIE, K., PAOLI, L., M. IBARBALZ, F. & ZINGER, L. 2021. Compendium of 530 metagenome-assembled bacterial and archaeal genomes from the polar Arctic Ocean. Nature Microbiology, 6, 1561-1574.

RYAN, P., FORRESTER, T. J., WROBLEWSKI, C., KENNEY, T. M., KITOVA, E. N., KLASSEN, J. S. & KIMBER, M. S. 2019. The small RbcS-like domains of the β-carboxysome structural protein CcmM bind RubisCO at a site distinct from that binding the RbcS subunit. Journal of Biological Chemistry, 294, 2593-5195.

SABBATINI, S., ARRIGA, N., BERTOLINI, T., CASTALDI, S., CHITI, T., CONSALVO, C., NJAKOU DJOMO, S., GIOLI, B., MATTEUCCI, G. & PAPALE, D. 2016. Greenhouse gas balance of cropland conversion to bioenergy poplar short-rotation coppice. Biogeosciences, 13, 95-113.

SABINE, C. L., HEIMANN, M., ARTAXO, P., BAKKER, D. C., CHEN, C.-T. A., FIELD, C. B., GRUBER, N., LE QUÉRÉ, C., PRINN, R. G. & RICHEY, J. E. 2004. Current status and past trends of the global carbon cycle. Scope-scientific committee on problems of the environment international council of scientific unions, 62, 17-44.

SAGE, R. F. 2002. Variation in the k cat of Rubisco in C3 and C4 plants and some implications for photosynthetic performance at high and low temperature. Journal of Experimental Botany, 53, 609-620.

SAITO, Y., OIKAWA, M., NAKAZAWA, H., NIIDE, T., KAMEDA, T., TSUDA, K. & UMETSU, M. 2018. Machine-Learning-Guided Mutagenesis for Directed Evolution of Fluorescent Proteins. ACS Synthetic Biology, 7, 2014-2022.

SALAZAR, G., PAOLI, L., ALBERTI, A., HUERTA-CEPAS, J., RUSCHEWEYH, H.-J., CUENCA, M., FIELD, C. M., COELHO, L. P., CRUAUD, C., ENGELEN, S., GREGORY, A. C., LABADIE, K., MAREC, C., PELLETIER, E., ROYO-LLONCH, M., ROUX, S., SÁNCHEZ, P., UEHARA, H., ZAYED, A. A., ZELLER, G., CARMICHAEL, M., DIMIER, C., FERLAND, J., KANDELS, S., PICHERAL, M., PISAREV, S., POULAIN, J., ACINAS, S. G., BABIN, M., BORK, P., BOSS, E., BOWLER, C., COCHRANE, G., DE VARGAS, C., FOLLOWS, M., GORSKY, G., GRIMSLEY, N., GUIDI, L., HINGAMP, P., IUDICONE, D., JAILLON, O., KANDELS-LEWIS, S., KARP-BOSS, L., KARSENTI, E., NOT, F., OGATA, H., PESANT, S., POULTON, N., RAES, J., SARDET, C., SPEICH, S., STEMMANN, L., SULLIVAN, M. B., SUNAGAWA, S., WINCKER, P., ACINAS, S. G., BABIN, M., BORK, P.,

BOWLER, C., DE VARGAS, C., GUIDI, L., HINGAMP, P., IUDICONE, D., KARP-BOSS, L., KARSENTI, E., OGATA, H., PESANT, S., SPEICH, S., SULLIVAN, M. B., WINCKER, P. & SUNAGAWA, S. 2019. Gene Expression Changes and Community Turnover Differentially Shape the Global Ocean Metatranscriptome. Cell, 179, 1068-1083.e21.

SALESSE-SMITH, C. E., SHARWOOD, R. E., BUSCH, F. A., KROMDIJK, J., BARDAL, V. & STERN, D. B. 2018. Overexpression of Rubisco subunits with RAF1 increases Rubisco content in maize. Nature Plants, 4, 802-810.

SALVUCCI, M. E., PORTIS, A. R. & OGREN, W. L. 1985. A soluble chloroplast protein catalyzes ribulosebisphosphate carboxylase/oxygenase activation in vivo. Photosynthesis research, 7, 193-201.

SATO, T., ATOMI, H. & IMANAKA, T. 2007. Archaeal type III RuBisCOs function in a pathway for AMP metabolism. Science, 315, 1003-1006.

SCALES, J. C., PARRY, M. A. & SALVUCCI, M. E. 2014. A non-radioactive method for measuring Rubisco activase activity in the presence of variable ATP: ADP ratios, including modifications for measuring the activity and activation state of Rubisco. Photosynth Res, 119, 355-65.

SERRANO, O., GÓMEZ-LÓPEZ, D. I., SÁNCHEZ-VALENCIA, L., ACOSTA-CHAPARRO, A., NAVAS-CAMACHO, R., GONZÁLEZ-CORREDOR, J., SALINAS, C., MASQUE, P., BERNAL, C. A. & MARBÀ, N. 2021. Seagrass blue carbon stocks and sequestration rates in the Colombian Caribbean. Scientific Reports, 11, 11067.

SHARWOOD, R. E., GHANNOUM, O., KAPRALOV, M. V., GUNN, L. H. & WHITNEY, S. M. 2016. Temperature responses of Rubisco from Paniceae grasses provide opportunities for improving C3 photosynthesis. Nature Plants, 2, 1-9.

SHEN, T., LI, W., PAN, W., LIN, S., ZHU, M. & YU, L. 2017. Role of bacterial carbonic anhydrase during CO2 capture in the CO2-H2O-carbonate system. Biochemical Engineering Journal, 123, 66-74.

SHINWARI, Z. K., JAN, S. A., NAKASHIMA, K. & YAMAGUCHI-SHINOZAKI, K. 2020. Genetic engineering approaches to understanding drought tolerance in plants. Plant Biotechnology Reports, 14, 151-162.

SIGMAN, D. M., JACCARD, S. L. & HAUG, G. H. 2004. Polar ocean stratification in a cold climate. Nature, 428, 59-63.

SKENDŽIĆ, S., ZOVKO, M., ŽIVKOVIĆ, I. P., LEŠIĆ, V. & LEMIĆ, D. 2021. The Impact of Climate Change on Agricultural Insect Pests. Insects, 12, 440.

SOMERO, G. N. 1995. Proteins and Temperature. Annual Review of Physiology, 57, 43-68.

SONG, Y., CHEN, Q., CI, D., SHAO, X. & ZHANG, D. 2014. Effects of high temperature on photosynthesis and related gene expression in poplar. BMC plant biology, 14, 1-20.

SPREITZER, R. J., PEDDI, S. R. & SATAGOPAN, S. 2005. Phylogenetic engineering at an interface between large and small subunits imparts land-plant kinetic properties to algal Rubisco. Proceedings of the National Academy of Sciences, 102, 17225-17230.

STOECKER, D. K. & LAVRENTYEV, P. J. 2018. Mixotrophic Plankton in the Polar Seas: A Pan-Arctic Review. Frontiers in Marine Science, 5.

STOTZ, M., MUELLER-CAJAR, O., CINIAWSKY, S., WENDLER, P., HARTL, F. U., BRACHER, A. & HAYER-HARTL, M. 2011. Structure of green-type Rubisco activase from tobacco. Nature Structural & Molecular Biology, 18, 1366-1370.

SUN, J.-L., SUI, X.-L., HUANG, H.-Y., WANG, S.-H., WEI, Y.-X. & ZHANG, Z.-X. 2014. Low light stress down-regulated Rubisco gene expression and photosynthetic capacity during cucumber (Cucumis sativus L.) leaf development. Journal of Integrative Agriculture, 13, 997-1007.

SUN, Y., GU, L. & DICKINSON, R. E. 2012. A numerical issue in calculating the coupled carbon and water fluxes in a climate model. Journal of Geophysical Research: Atmospheres, 117.

SUN, Y., HARMAN, V. M., JOHNSON, J. R., BROWNRIDGE, P. J., CHEN, T., DYKES, G. F., LIN, Y., BEYNON, R. J. & LIU, L.-N. 2022. Decoding the absolute stoichiometric composition and structural plasticity of α-carboxysomes. Mbio, 13, e03629-21.

SUNAGAWA, S., ACINAS, S. G., BORK, P., BOWLER, C., ACINAS, S. G., BABIN, M., BORK, P., BOSS, E., BOWLER, C., COCHRANE, G., DE VARGAS, C., FOLLOWS, M., GORSKY, G., GRIMSLEY, N., GUIDI, L., HINGAMP, P., IUDICONE, D., JAILLON, O., KANDELS, S., KARP-BOSS, L., KARSENTI, E., LESCOT, M., NOT, F., OGATA, H., PESANT, S., POULTON, N., RAES, J., SARDET, C., SIERACKI, M., SPEICH, S., STEMMANN, L., SULLIVAN, M. B., SUNAGAWA, S., WINCKER, P., EVEILLARD, D., GORSKY, G., GUIDI, L., IUDICONE, D., KARSENTI, E., LOMBARD, F., OGATA, H., PESANT, S., SULLIVAN, M. B., WINCKER, P., DE VARGAS, C. & TARA OCEANS, C. 2020. Tara Oceans: towards global ocean ecosystems biology. Nature Reviews Microbiology, 18, 428-445.

TABITA, F. R., HANSON, T. E., SATAGOPAN, S., WITTE, B. H. & KREEL, N. E. 2008a. Phylogenetic and evolutionary relationships of RubisCO and the RubisCO-like proteins and the functional lessons provided by diverse molecular forms. Philosophical Transactions of the Royal Society B: Biological Sciences, 363, 2629-2640.

TABITA, F. R., SATAGOPAN, S., HANSON, T. E., KREEL, N. E. & SCOTT, S. S. 2008b. Distinct form I, II, III, and IV Rubisco proteins from the three kingdoms of life provide clues about Rubisco evolution and structure/function relationships. Journal of experimental botany, 59, 1515-1524.

TANAKA, Y., NAKATSUMA, D., HARADA, H., ISHIDA, M. & MATSUDA, Y. 2005. Localization of soluble β-carbonic anhydrase in the marine diatom Phaeodactylum tricornutum. Sorting to the chloroplast and cluster formation on the girdle lamellae. Plant Physiology, 138, 207-217.

TAYLOR, T. C., BACKLUND, A., BJORHALL, K., SPREITZER, R. J. & ANDERSSON, I. 2001. First crystal structure of Rubisco from a green alga, Chlamydomonas reinhardtii. J Biol Chem, 276, 48159-64.

TAYLOR, T. C., BACKLUND, A., BJORHALL, K., SPREITZER, R. J. & ANDERSSON, I. 2001. First crystal structure of Rubisco from a green alga, Chlamydomonas reinhardtii. Journal of Biological Chemistry, 276, 48159-48164.

TISON, J.-L., MAKSYM, T., FRASER, A. D., CORKILL, M., KIMURA, N., NOSAKA, Y., NOMURA, D., VANCOPPENOLLE, M., ACKLEY, S. & STAMMERJOHN, S. 2020. Physical and biological properties of early winter Antarctic sea ice in the Ross Sea. Annals of Glaciology, 61, 241-259.

TOMMASI, I. C. 2021. The mechanism of Rubisco catalyzed carboxylation reaction: chemical aspects involving acid-base chemistry and functioning of the molecular machine. Catalysts, 11, 813.

TOYODA, K., YOSHIZAWA, Y., ISHII, M. & ARAI, H. 2022. Regulation of the high-specificity Rubisco genes by the third CbbR-type regulator in a hydrogen-oxidizing bacterium Hydrogenovibrio marinus. Journal of Bioscience and Bioengineering, 134, 496-500.

TSAI, Y. C., LIEW, L., GUO, Z., LIU, D. & MUELLER-CAJAR, O. 2022b. The CbbQO-type rubisco activases encoded in carboxysome gene clusters can activate carboxysomal form IA rubiscos. J Biol Chem, 298, 101476.

TSAI, Y. C., MUELLER-CAJAR, O., SASCHENBRECKER, S., HARTL, F. U. & HAYER-HARTL, M. 2012. Chaperonin cofactors, Cpn10 and Cpn20, of green algae and plants function as hetero-oligomeric ring complexes. J Biol Chem, 287, 20471-81.

TSAI, Y.-C. C., LAPINA, M. C., BHUSHAN, S. & MUELLER-CAJAR, O. 2015a. Identification and characterization of multiple rubisco activases in chemoautotrophic bacteria. Nature communications, 6, 8883-8883.

TSAI, Y.-C. C., LAPINA, M. C., BHUSHAN, S. & MUELLER-CAJAR, O. 2015b. Identification and characterization of multiple rubisco activases in chemoautotrophic bacteria. Nature communications, 6, 8883.

TSAI, Y.-C. C., LIEW, L., GUO, Z., LIU, D. & MUELLER-CAJAR, O. 2022a. The CbbQO-type rubisco activases encoded in carboxysome gene clusters can activate carboxysomal form IA rubiscos. Journal of Biological Chemistry, 298.

TSAI, Y.-C. C., YE, F., LIEW, L., LIU, D., BHUSHAN, S., GAO, Y.-G. & MUELLER-CAJAR, O. 2020. Insights into the mechanism and regulation of the CbbQO-type Rubisco activase, a MoxR AAA+ ATPase. Proceedings of the National Academy of Sciences, 117, 381-387.

TSENG, C.-H. & TANG, S.-L. 2014. Marine microbial metagenomics: from individual to the environment. International Journal of Molecular Sciences, 15, 8878-8892.

VALEGÅRD, K., ANDRALOJC, P. J., HASLAM, R. P., PEARCE, F. G., ERIKSEN, G. K., MADGWICK, P. J., KRISTOFFERSEN, A. K., VAN LUN, M., KLEIN, U. & EILERTSEN, H. C. 2018a. Structural and functional analyses of Rubisco from arctic diatom species reveal unusual posttranslational modifications. Journal of Biological Chemistry, 293, 13033-13043.

VALEGÅRD, K., HASSE, D., ANDERSSON, I. & GUNN, L. H. 2018. Structure of Rubisco from Arabidopsis thaliana in complex with 2-carboxyarabinitol-1,5-bisphosphate. Acta Crystallogr D Struct Biol, 74, 1-9.

VALEGÅRD, K., HASSE, D., ANDERSSON, I. & GUNN, L. H. 2018b. Structure of Rubisco from Arabidopsis thaliana in complex with 2-carboxyarabinitol-1,5-bisphosphate. Acta Crystallogr D Struct Biol, 74, 1-9.

VAN DER MAATEN, L. & HINTON, G. 2008. Visualizing data using t-SNE. Journal of machine learning research, 9.

VERNETTE, C., LECUBIN, J., SÁNCHEZ, P., COORDINATORS, T. O., SUNAGAWA, S., DELMONT, T. O., ACINAS, S. G., PELLETIER, E., HINGAMP, P. & LESCOT, M. 2022. The Ocean Gene Atlas v2.0: online exploration of the biogeography and phylogeny of plankton genes. Nucleic Acids Research, 50, W516-W526.

VINCENT, W. F. 2002. Cyanobacterial Dominance in the Polar Regions. In: WHITTON, B. A. & POTTS, M. (eds.) The Ecology of Cyanobacteria: Their Diversity in Time and Space. Dordrecht: Springer Netherlands.

VITLIN GRUBER, A., NISEMBLAT, S., AZEM, A. & WEISS, C. 2013. The complexity of chloroplast chaperonins. Trends Plant Sci, 18, 688-94.

VOROBEV, A., DUPOUY, M., CARRADEC, Q., DELMONT, T. O., ANNAMALÉ, A., WINCKER, P. & PELLETIER, E. 2020. Transcriptome reconstruction and functional analysis of eukaryotic marine plankton communities via high-throughput metagenomics and metatranscriptomics. Genome research, 30, 647-659.

WACHTER, R. M., SALVUCCI, M. E., CARMO-SILVA, A. E., BARTA, C., GENKOV, T. & SPREITZER, R. J. 2013. Activation of interspecies-hybrid Rubisco enzymes to assess different models for the Rubisco–Rubisco activase interaction. Photosynthesis Research, 117, 557-566.

WANG, H., YAN, X., AIGNER, H., BRACHER, A., NGUYEN, N. D., HEE, W. Y., LONG, B. M., PRICE, G. D., HARTL, F. U. & HAYER-HARTL, M. 2019. Rubisco condensate formation by CcmM in β-carboxysome biogenesis. Nature, 566, 131-135.

WANG, J. 2020. An intuitive tutorial to Gaussian processes regression. arXiv preprint arXiv:2009.10862.

WANG, L.-M., SHEN, B.-R., LI, B.-D., ZHANG, C.-L., LIN, M., TONG, P.-P., CUI, L.-L., ZHANG, Z.-S. & PENG, X.-X. 2020. A Synthetic Photorespiratory Shortcut Enhances Photosynthesis to Boost Biomass and Grain Yield in Rice. Molecular Plant, 13, 1802-1815.

WERTHEIM, J. O., MURRELL, B., SMITH, M. D., KOSAKOVSKY POND, S. L. & SCHEFFLER, K. 2015. RELAX: detecting relaxed selection in a phylogenetic framework. Molecular biology and evolution, 32, 820-832.

WHEATLEY, N. M., SUNDBERG, C. D., GIDANIYAN, S. D., CASCIO, D. & YEATES, T. O. 2014. Structure and identification of a pterin dehydratase-like protein as a ribulose-bisphosphate carboxylase/oxygenase (RuBisCO) assembly factor in the α-carboxysome. J Biol Chem, 289, 7973-81.

WHITEHEAD, L., LONG, B. M., PRICE, G. D. & BADGER, M. R. 2014. Comparing the in vivo function of α-carboxysomes and β-carboxysomes in two model cyanobacteria. Plant Physiol, 165, 398-411.

WHITNEY, S. M., HOUTZ, R. L. & ALONSO, H. 2011. Advancing Our Understanding and Capacity to Engineer Nature's CO_2-Sequestering Enzyme, Rubisco. Plant Physiology, 155, 27-35.

WIDDERICH, N., HOEPPNER, A., PITTELKOW, M., HEIDER, J., SMITS, S. H. & BREMER, E. 2014. Biochemical properties of ectoine hydroxylases from extremophiles and their wider taxonomic distribution among microorganisms. PloS one, 9, e93809.

WIETRZYNSKI, W., TRAVERSO, E., WOLLMAN, F.-A. & WOSTRIKOFF, K. 2021. The state of oligomerization of Rubisco controls the rate of synthesis of the Rubisco large subunit in Chlamydomonas reinhardtii. The Plant Cell, 33, 1706-1727.

WU, A., BRIDER, J., BUSCH, F. A., CHEN, M., CHENU, K., CLARKE, V. C., COLLINS, B., ERMAKOVA, M., EVANS, J. R. & FARQUHAR, G. D. 2023. A cross-scale analysis to understand and quantify the effects of photosynthetic enhancement on crop growth and yield across environments. Plant, Cell & Environment, 46, 23-44.

WU, A., HAMMER, G. L., DOHERTY, A., VON CAEMMERER, S. & FARQUHAR, G. D. 2019. Quantifying impacts of enhancing photosynthesis on crop yield. Nature plants, 5, 380-388.

XIE, C., ZHANG, R., QU, Y., MIAO, Z., ZHANG, Y., SHEN, X., WANG, T. & DONG, J. 2012. Overexpression of MtCAS31 enhances drought tolerance in transgenic Arabidopsis by reducing stomatal density. New Phytologist, 195, 124-135.

XIE, J., XU, Z., ZHOU, S., PAN, X., CAI, S., YANG, L. & MEI, H. 2013. The VHSE-based prediction of proteasomal cleavage sites. PLoS One, 8, e74506.

YAMORI, W., SUZUKI, K., NOGUCHI, K., NAKAI, M. & TERASHIMA, I. 2006. Effects of Rubisco kinetics and Rubisco activation state on the temperature dependence of the photosynthetic rate in spinach leaves from contrasting growth temperatures. Plant, cell & environment, 29, 1659-1670.

YANG, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Molecular biology and evolution, 24, 1586-1591.

YANG, Z., WONG, W. S. W. & NIELSEN, R. 2005. Bayes Empirical Bayes Inference of Amino Acid Sites Under Positive Selection. Molecular Biology and Evolution, 22, 1107-1118.

YOUNG, J. N. & SCHMIDT, K. 2020. It's what's inside that matters: physiological adaptations of high-latitude marine microalgae to environmental change. New Phytologist, 227, 1307-1318.

YOUNG, J. N., GOLDMAN, J. A. L., KRANZ, S. A., TORTELL, P. D. & MOREL, F. M. M. 2015. Slow carboxylation of Rubisco constrains the rate of carbon fixation during Antarctic phytoplankton blooms. New Phytologist, 205, 172-181.

YOUNG, J. N., GOLDMAN, J. A. L., KRANZ, S. A., TORTELL, P. D. & MOREL, F. M. M. 2015a. Slow carboxylation of Rubisco constrains the rate of carbon fixation during Antarctic phytoplankton blooms. New Phytologist, 205, 172-181.

YOUNG, J. N., HEUREUX, A. M. C., SHARWOOD, R. E., RICKABY, R. E. M., MOREL, F. M. M. & WHITNEY, S. M. 2016. Large variation in the Rubisco kinetics of diatoms reveals diversity among their carbon-concentrating mechanisms. Journal of Experimental Botany, 67, 3445-3456.

YOUNG, J. N., KRANZ, S. A., GOLDMAN, J. A. L., TORTELL, P. D. & MOREL, F. M. M. 2015b. Antarctic phytoplankton down-regulate their carbon-concentrating mechanisms under high CO2 with no change in growth rates. Marine Ecology Progress Series, 532, 13-28.

YOUNG, J. N., RICKABY, R. E. M., KAPRALOV, M. V. & FILATOV, D. A. 2012. Adaptive signals in algal Rubisco reveal a history of ancient atmospheric carbon dioxide. Philosophical Transactions of the Royal Society B: Biological Sciences, 367, 483-492.

ZARZYCKI, J., AXEN, S. D., KINNEY, J. N. & KERFELD, C. A. 2012. Cyanobacterial-based approaches to improving photosynthesis in plants. Journal of Experimental Botany, 64, 787-798.

ZHAN, Y., MARCHAND, C. H., MAES, A., MAURIES, A., SUN, Y., DHALIWAL, J. S., UNIACKE, J., ARRAGAIN, S., JIANG, H., GOLD, N. D., MARTIN, V. J. J., LEMAIRE, S. D. & ZERGES, W. 2018. Pyrenoid functions revealed by proteomics in Chlamydomonas reinhardtii. PLoS One, 13, e0185039.

ZHANG, N., KALLIS, R. P., EWY, R. G. & PORTIS JR, A. R. 2002. Light modulation of Rubisco in Arabidopsis requires a capacity for redox regulation of the larger Rubisco activase isoform. Proceedings of the National Academy of Sciences, 99, 3330-3334.

ZHOU, Y. & WHITNEY, S. 2019. Directed Evolution of an Improved Rubisco; In Vitro Analyses to Decipher Fact from Fiction. Int J Mol Sci, 20.

ZHOU, Y., GUNN, L. H., BIRCH, R., ANDERSSON, I. & WHITNEY, S. M. 2023. Grafting Rhodobacter sphaeroides with red algae Rubisco to accelerate catalysis and plant growth. Nature Plants, 9, 978-986.

ZHU, X.-G., PORTIS JR, A. R. & LONG, S. P. 2004. Would transformation of C3 crop plants with foreign Rubisco increase productivity? A computational analysis extrapolating from kinetic properties to canopy photosynthesis. Plant, Cell & Environment, 27, 155-165.

ZORZ, J. K., ALLANACH, J. R., MURPHY, C. D., ROODVOETS, M. S., CAMPBELL, D. A. & COCKSHUTT, A. M. 2015. The RUBISCO to photosystem II ratio limits the maximum photosynthetic rate in picocyanobacteria. Life, 5, 403-417.