*Year :* 2019

# Molecular évolution of RuBisCO subunits

Yamada Kana

Yamada Kana, 2019, Molecular évolution of RuBisCO subunits

Originally published at : Thesis, University of Lausanne

Posted at the University of Lausanne Open Archive http://serval.unil.ch
Document URN : urn:nbn:ch:serval-BIB_D71D8857D8D87

UNIL | Université de Lausanne
Faculté de biologie
et de médecine

**Département de biologie computationnelle**

# Molecular evolution of RuBisCO subunits

**Thèse de doctorat ès sciences de la vie (PhD)**

**présentée à la**

**Faculté de biologie et de médecine**

**de l'Université de Lausanne**

**par**

## Kana Yamada

**Master en biologie de l'Université de Tokyo**

**Jury**

**Prof. Christian Fankhauser**

**Prof. Nicolas Salamin**

**Dr. Luca Fumagalli**

**Dr. Pascal-Antoine Christin**

**Lausanne 2019**

UNIL | Université de Lausanne
Faculté de biologie
et de médecine

Ecole Doctorale
Doctorat ès sciences de la vie

# Imprimatur

Vu le rapport présenté par le jury d'examen, composé de

| | | |
|---|---|---|
| **Président·e** | Monsieur | Prof. Christian **Fankhauser** |
| **Directeur·rice de thèse** | Monsieur | Prof. Nicolas **Salamin** |
| **Experts·es** | Monsieur | Dr Luca **Fumagalli** |
| | Monsieur | Dr Pascal-Antoine **Christin** |

le Conseil de Faculté autorise l'impression de la thèse de

**Madame Kana Yamada**

Master of Agriculture The University of Tokyo, Japon

intitulée

**Molecular evolution of RuBisCO subunits**

Lausanne, le 27 mars 2019

pour le Doyen
de la Faculté de biologie et de médecine

Prof. Christian Fankhauser

# Acknowledgements

Firstly I would like to thanks to my supervisor, Nicolas Salamin for giving me a chance to come to Switzerland and to work on this interesting project. I appreciate deeply to Romain Studer for teaching me the homology modelling and encouraging me to continue. I would like to thanks to Iakov Davydov for giving advice on the positive selection analysis. I am also grateful for his kind personality and open-mind. I would like to thanks to Andrea Komljenovic for giving me some advice about transcriptome analysis. I appreciate Martha Serrano-Serrano for having general discussions about statistical analysis. I am grateful to Dessislava Savova Bianchi and Catherine Berney for giving me some advice about technical issues in wet-lab. I would like to thanks to Guillaume Besnard and Pascal-Antoine Christin for giving me access to precious samples and to Victor Rossier for downloading data and running Exonerate. I appreciate to Pascal-Antoine for his help to improve the writing of my thesis, especially for sections of discussion. I would like to thanks to my friends/colleagues, GuangPeng Ren, and Xavier Meyer for sharing all last years and encouraging me. I would like to thanks to Martha, Xavier Oriane, Daniel, Darren, Phillipe, Rocio, Robert and Sarah for reading some part of my thesis and giving me advice. I would like to thanks to my lab mates for sharing coffee time, Daniele, Sacha, Pablo, Hannes, Talita, and Alexandra. I appreciate my friends of DEE, Madeleine, Darina, Julien G. Eric, Glib, Prof. Nicolas P., Marie, Yan, Kamil, and Katie for sharing nice time together. Part of this work was done at the University of Peradeniya in Sri Lanka. I deeply appreciate to Prof. Sanath Rajapakse and Prof. Suneth Sooriyapathirana for inviting me to Sri Lanka and preparing a great working environment. I would like to thanks to Prof. Chandima Dhanapala, Prof. Ryan Kerney, other professors and demonstrators for sharing a great time. I would like to thanks to my master student, Udari Jayathilake for collaboration about the *rbcS* evolution in carnivorous plants of Sri Lanka.

I appreciate deeply from my heart to my friends who supported me for these long years of Ph.D. Especially to team Pavement, Benjamin Hinterlang, Hélène Joly, Samuel Umbricht. I really loved the time that we shared together. Benji helped me a lot to find a solution for my health issues, especially when I had surgery. He encouraged me for many years to finish my thesis. He never had doubt that I can accomplish. Hélène always listened to me and shared many feelings. I deeply appreciate my boyfriend, Sämi for being always supportive of whatever situations that we had. He is always flexible to any situations and he has simple and positive-mind. He helped me a lot to accomplish. Living in Sri Lanka and traveling around the world was one of the greatest experiences in my life and I appreciate that we could share the precious time together. I would like to thanks to beautiful Sri Lanka and people there for giving me a chance to stay there for a year. I appreciate my family in Sri Lanka, Shanka, Mithma, and Sanath De Silva for taking care of me and giving me the best environment to live. I appreciate to the family of Shanka, Chekha, Sama, Kolitha, Dileepa, Jayandi and P. K. M. B Senanayake, kiriyamma for having me very close to their family. Meeting wonderful people like them is a treasure of my life. I would like to thanks to my best friend, Ingrid Dextra to be always with me and support me. I would like to thanks to Sämi's parents Max and Renate to welcome me to their place and care about me all the time. I thank to Nina, Adi, liebi Sophie and liebi Jäel for sharing nice time together. I would like to thanks to my friends, Giannis, Monica, Volo, Piyarathna of Sithumina, Jayrathna, Hisa, Yumiko, Haruka, Eléna, Mireille, Yvan, Dennis, Charaka, Madame coquillage, Mohamed, Lucas and Anthony for their supports during last years. Finally, I would like to thanks to my mother, Masako and my sister, Rina for helping my decision to come to Switzerland. This Ph.D. thesis is dedicated to Toshiko Matsumura, my dearest grandmother.

# Abstract

The environmental conditions of our planet have been changing since its origin. For species' survival, adaptation to the environment is crucial, for example through the adaptive evolution of photosynthesis. The appearance of the mechanism to concentrate $CO_2$ has given some species a selective advantage under $CO_2$-depleted conditions. C4 plants comprise one of the main groups of such species that have diverged from classical C3 plants and adapted to depletion of $CO_2$ by modifying the cellular structures and biochemical cascades. Ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO), an enzyme which catalyzes the first step of $CO_2$ fixation, has changed cellular location during C4 evolution. RuBisCO of C4 is surrounded by highly concentrated $CO_2$, which prevents the loss of energy and $CO_2$ caused by the affinity of the enzyme for both $O_2$ and $CO_2$. The intercellular gas composition surrounding RuBisCO directly influences the rate of photosynthesis because RuBisCO's slow turnover rate is often the limiting factor for the rate of photosynthesis in higher plants. Therefore, RuBisCO has been considered as the determining factor of the photosynthetic rate and it has been thought to play an important role in plant adaptation to the environmental conditions. In previous studies, the evidence of adaptive evolution of RuBisCO has been detected by positive selection acting on the chloroplast *rbcL* gene encoding large subunits of RuBisCO (RBCL) in independent C4 lineages. The other subunit of RuBisCO, the small subunit (RBCS), has been reported to influence the catalytic efficiency, $CO_2$ specificity, assembly, activity, and stability of RuBisCO. However, the evolution of its encoding nuclear gene *rbcS* is yet poorly studied. Therefore, I aimed to study the molecular evolution of *rbcS* in angiosperms. The *rbcS* gene is a multigene family and the number of gene copies is different between species. The phylogenetic tree of the *rbcS* gene reveals two lineages that may have originated from a duplication event before the divergence of land plants. Copies originating from ancient duplication events seem to have been removed, whereas the copies from recent events appear to be retained. This explains the observation in the *rbcS* tree that gene copies of the same species are more closely related to each other than ones from different species. I hypothesized that each *rbcS* gene copy of the same species may have different characteristics. I compared the interaction of *rbcS* and *rbcL* genes as well as the influence of different encoding RBCS subunits to the stability of RuBisCO by respectively testing coevolution between *rbcS* and each *rbcL* and by homology modelling of RuBisCO composed with a RBCS encoded by different *rbcS* copies. The results suggested that the interaction between RBCS and RBCL, and the influence on the overall stability of the enzyme, are the same among different *rbcS* copies. Therefore, I assumed that all the *rbcS* gene copies cannot be divergent because they need to be structurally compatible with RBCL. In general, when all the gene copies of a multigene family have the same characteristics, multiple gene copies of a species exist to maintain the number of transcripts at the same level as that of a single copy carrying species (dosage effect hypothesis). To test this hypothesis, I estimated the gene expression levels of each gene copy by using published transcriptome data. The results suggest that the gene expression level is similar between species carrying single and multiple copies. The results suggest that species carrying a higher gene copy number have a larger amount of RuBisCO. It has been reported that RuBisCO is degraded or down regulated under specific environmental stress. Thus, I conclude that plants living in such an environmental stress condition may need to synthesize more RuBisCO to prevent a shortage of the enzyme. To understand better the role of RBCS to cope with environmental changes, I tested the positive selection of the *rbcS* gene in species of Poaceae that have different photosynthetic types. Positive selection was detected all over the tree and the signal was not C4-specific. This suggests that the positive selection acting on the *rbcS* gene has not led to the shift of photosynthetic types. I assume that RBCS might be involved in the optimization of RuBisCO after the establishment of C4 photosynthesis type or after migration to new habitats that require different catalytic properties.

# Résumé

Les conditions environnementales de notre planète ne cessent de changer depuis son origine. Pour survivre, il est crucial pour les espèces de s'adapter à leur environnement. Un exemple est l'évolution adaptative de la photosynthèse. L'apparition de mécanismes permettant de concentrer le $CO_2$ a donné à certaines espèces un avantage sélectif lorsqu'elles font face à des conditions appauvries en $CO_2$. Les plantes C4 constituent l'un des principaux groupes d'espèces qui ont divergé des plantes C3 classiques en s'adaptant en modifiant leurs structures cellulaires et cascades biochimiques. La ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO) – une enzyme catalysant la première étape de fixation de $CO_2$ – a changé de localisation cellulaire durant l'évolution du mode de fixation du carbone C4. La RuBisCO des plantes C4 est localisée dans un compartiment caractérisé par une haute concentration en $CO_2$, évitant ainsi la perte d'énergie et de $CO_2$ causée par l'affinité de l'enzyme pour deux substrats: le $CO_2$ et le $O_2$. L'environnement gazeux intracellulaire auquel est confrontée la RuBisCO influence directement le taux de photosynthèse, car son faible taux de renouvellement par rapport à d'autres enzymes photosynthétiques constitue souvent le facteur limitant le taux de photosynthèse chez les plantes supérieures. De ce fait, la RuBisCO est considérée comme le facteur déterminant le taux de photosynthèse et jouant un rôle important dans l'adaptation des plantes aux conditions environnementales. De précédentes études démontrèrent l'évolution adaptative de la RuBisCO par sélection positive agissant sur le gène chloroplastique *rbcL* – qui code pour la grande sous-unité de la RuBisCO (RBCL) – dans des lignées indépendantes de plantes C4. Il a été démontré que l'autre sous-unité de la RuBisCO – la petit sous-unité (RBCS) – influence l'efficacité catalytique, la spécificité de liaison au $CO_2$, l'assemblement, l'activité et la stabilité de la RuBisCO. Néanmoins, l'évolution du gène codant pour cette sous-unité – le gène nucléaire *rbcS* – n'a été que très peu étudiée jusqu'à présent. Par conséquent, le but de mon projet est d'étudier l'évolution moléculaire du gène *rbcS* chez les Angiospermes. Le gène *rbcS* fait partie d'une famille de gènes multiples et son nombre de copies varie selon les espèces. Des arbres phylogénétiques se basant sur *rbcS* ont révélé deux lignées provenant potentiellement d'un événement de duplication ayant eu lieu avant la divergence des plantes terrestres. Les copies provenant d'anciens événements de duplication semblent avoir été éliminées, alors que les copies provenant d'événements récents de duplications paraissent avoir été conservées. Cela explique que les copies de *rbcS* provenant d'une même espèce soient plus proches phylogénétiquement les unes des autres que des copies provenant d'espèces différentes. Je mets en avant l'hypothèse que chaque copie du gène *rbcS* de la même espèce pourrait avoir différentes caractéristiques. J'ai comparé l'interaction entre les gènes *rbcS* et *rbcL* ainsi que l'influence des différentes sous-unités RBCS à la stabilité de la RuBisCO en testant respectivement la coévolution entre *rbcS* et chaque *rbcL* et en modélisant par homologie la RuBisCO composée par une sous-unité RBCS codée par différentes copies du gène *rbcS*. Les résultats suggèrent que l'interaction entre chaque *rbcS* et *rbcL* et l'influence sur la stabilité générale de l'enzyme est similaire entre les différentes copies de *rbcS*. En conséquence, je présume que les différentes copies du gène *rbcS* ne peuvent pas être divergentes car il est nécessaire qu'elles soient compatibles structurellement avec la sous-unité RBCL. En général, lorsque toutes les copies de gènes provenant d'une même famille de gènes multiples ont les mêmes caractéristiques, les différentes copies de gènes permettent de maintenir la même quantité d'éléments transcrits en comparaison avec une espèce ne possédant qu'une copie du gène (hypothèse « d'effet de dosage »). Afin de tester cette hypothèse, j'ai estimé le niveau d'expression pour chaque copie de gène de la même espèce en me basant sur des données transcriptomiques déjà publiées. Les résultats suggèrent que le niveau d'expression des gènes est similaire entre les espèces ayant une ou plusieurs copies du gène. De ce fait, l'hypothèse d'effet de dosage n'est pas applicable dans le cadre de l'évolution de *rbcS*. Les résultats suggèrent que les espèces ayant un plus grand nombre de copies du gène disposent également d'une plus grande quantité de RuBisCO. Il a été rapporté que la RuBisCO se dégrade ou est régulée négativement dans des conditions de stress spécifiques. Par conséquent, je présume que les plantes vivant dans de telles conditions environnementales stressantes doivent synthétiser plus de RuBisCO pour éviter une pénurie de l'enzyme. Pour mieux comprendre le rôle de RBCS face aux changements environnementaux, j'ai testé la sélection positive du gène *rbcS* chez des espèces de Poacées ayant différents mécanismes photosynthétiques. Une sélection positive a été détectée chez toutes les espèces et le signal n'était pas spécifique aux espèces à système C4. Cela suggère que la sélection positive agissant sur le gène *rbcS* n'est pas responsable du changement de type de photosynthèse. Je présume que RBCS ne serait donc pas impliquée dans la transition C3 à C4, mais que cette sous-unité pourrait être impliquée dans l'optimisation de la RuBisCO après l'établissement de la photosynthèse de type C4 ou après la migration vers de nouveaux habitats nécessitant différentes propriétés catalytiques.

# Index

# List of figures

## Chapter 1



## Chapter 2



## Chapter 3

# List of tables

## Chapter 1



## Chapter 2



## Chapter 3

# General introduction

The environmental conditions of the surface of our planet have been constantly changing since its origin (Beerling & Royer, 2011; Edwards, Osborne, Strömberg, Smith, & Consortium, 2010; Pearson, Foster, & Wade, 2009). This changing environment has affected the survival of species and, for some species, led to adaptation to new environmental conditions by changing morphological characteristics, cellular structures, biochemical cascades, and ecological niches of the organisms (Miller, Ota, Sumaila, Cisneros-Montemayor, & Cheung, 2018; Sage, Christin, & Edwards, 2011; Zhang, Zhang, & Rosenberg, 2002; Zhong et al., 2013).

An example of such a process is the adaptive evolution of photosynthesis linked to environmental changes (Christin et al., 2008a; Edwards, Still, & Donoghue, 2007; Horn et al., 2014; Kapralov, Smith, & Filatov, 2012; Mckown & Dengler, 2007; Studer, Christin, Williams, & Orengo, 2014; Yamori & Von Caemmerer, 2009). Ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO) is an enzyme of the photosynthetic pathway that is estimated to have appeared billions of years ago when the atmosphere did not include $O_2$ (Bauwe, Hagemann, Kern, & Timm, 2012). It catalyzes the first reaction to fix $CO_2$ to sugar in the cellular cycle known as the Calvin-Benson cycle in plants. After the $O_2$ has started to be produced in the atmosphere, RuBisCO has started to intake not only $CO_2$ but also $O_2$ as substrates by catalyzing both photosynthesis and photorespiration. Because photorespiration is the reverse reaction of photosynthesis (Spreitzer & Salvucci, 2002), it causes waste of energy and $CO_2$ (Parry, Andralojc, Mitchell, Madgwick, & Keys, 2003; von Caemmerer & Quick, 2000). As atmospheric $CO_2$ decreased drastically in the Oligocene (Beerling & Royer, 2011; Edwards et al., 2010; Pearson et al., 2009), the supply of $CO_2$ to RuBisCO came to be insufficient.

To solve this problem, some plants have evolved to carry the $CO_2$-concentrating mechanism (CCM) of the photosynthetic pathway as the adaptive evolution to the depletion of $CO_2$ (Christin et al., 2008b; Edwards et al., 2010; Sage, Christin, & Edwards, 2011; Vicentini, Barber, & Aliscioni, 2008). The $CO_2$-concentrating mechanism is a cellular-structural or temporal mechanism to concentrate the atmospheric $CO_2$ (Simpson, 1953). In the classical C3 type of photosynthesis, $CO_2$ is taken into mesophyll cells and fixed in the Calvin-Benson cycle in the same cells. However, a new type of photosynthesis that has diverged from C3 type (Sage, 2004), called C4 type, has CCM, whereby $CO_2$ is taken into mesophyll cells but is transferred to bundle-sheath cells and fixed in the Calvin-Benson cycle in bundle-sheath cells (Leegood, 2002; Sage, 2004). These new modifications enable RuBisCO of C4 plants to receive highly concentrated $CO_2$; thus, the catalytic efficiency of RuBisCO is higher in C4 plants than in C3 plants (Kubien, Whitney, moore, & Jesson, 2008; Sage, 2002; Seemann, Badger, & Berry, 1984; Wessinger, Edwards, & Ku, 1989; Yeoh, Badger, & Watson, 1980; et al., 1980; Yeoh, Badger, & Watson, 1981). Conversely, the $CO_2$ specificity of RuBisCO is higher in C3 plants than in C4 plants (Jordan & Ogren, 1983; Kubien et al., 2008; Sage, 2002, Seemann et al., 1984; von Caemmerer & Quick, 2000) because high $CO_2$ specificity is less important in $CO_2$-rich cells of C4 plants (Tcherkez, Farquhar, Andrews, & Lorimer, 2006).

In the transition from C3 to C4 type, it is not only the cellular location of the Calvin-Benson cycle and cellular structures that have changed, but new enzymes have also been synthesized. The main pathway of C4 photosynthesis can be explained simply as follows. Firstly, the atmospheric $CO_2$ is fixed by Beta-carbonic anhydrase (Beta-CA) and phosphoenolpyruvate carboxylase (PEPC) in mesophyll cells. Fixed carbon compounds are transferred to bundle-sheath cells by the involvement of multiple enzymes. Then, $CO_2$ is released by decarboxylating enzymes such as NAD-malic enzymes (NAD-ME), NADP-malic enzymes (NADP-ME), or/and phosphoenolpyruvate carboxykinase (PCK). Then, released $CO_2$ is fixed in the Calvin-Benson cycle of bundle-sheath cells (Badger & Price, 1994; Christin et al., 2013; Furbank, Hatch, & Jenkins, 2000; Grula & Hudspeth, 1987; Kanai &

Edwards, 1999; Ku, Kano-Murakami, & Matsuoka, 1996; Matsuoka, 1995). Indeed, C4-type photosynthesis is one of the most complicated processes that many enzymes and biochemical reactions are involved in. However, not all the C4-specific enzymes are newly synthesized from a sketch. In most of cases, genes encoding C4-specific enzymes have already existed in C3 type, and co-option of these genes has led the transition from C3 to C4 type (Christin et al., 2013). For example, a gene family encoding NADP-ME already existed in C3 type and the recruitment of specific gene lineages by up-regulation of expression resulted in C4-specific NADP-ME (Christin et al., 2013). Another example is the switch of the predominant isoforms of Beta-CA from C3 type to C4 type, that functions and intracellular locations are different. The switch of predominant isoforms is determined by regulation of gene expression levels of encoding genes of different isoforms (Ludwig, 2012, 2016; Tanz, Tetu, Vella, & Ludwig, 2009).

A C4-specific enzyme is often encoded by a gene family which includes multiple gene copies (members), known as a multigene family. These play important roles in organizing the novel or modified functions that are required during adaptation (Mcglothlin et al., 2016; Nei, Gu, & Sitnikova, 1997; Niimura, 2009; Ohta 1991). The members of multigene families can differ in function, cellular localization of encoding protein, stability, and/or expression levels (Clark, Sessions, Eastburn, & Roux, 2001; Hudsona, Dengler, Hattersleya, & Dengler, 1992; Ku et al., 1996; Niimura, 2009; Petter, Bonow, & Klinkert, 2008). Also, the evolutionary history of each gene copy of a multigene family can be highly diverse (Nei et al., 1997; Ohta, 1991). Evolutionary processes such as duplication or the selection of advantageous mutations occur independently in specific copies, and the genetic information can be exchanged between the gene copies by crossing over, recombination, and/or gene conversions (Dumont & Elchler, 2013; Mano & Innan, 2008; Nei & Rooney, 2005; Ohta, 1977, 1979, 1983). Therefore, tracking the evolution of multigene families is extremely complex (Christin et al., 2013; Ohta, 1991), and our knowledge of these processes is limited (Benton, 2015; Eyun, 2013).

The *rbcS* multigene family encodes the small subunits (RBCS) of RuBisCO. RuBisCO is composed of eight RBCS and eight large subunits (RBCL) (Andersson, 2008). Since the catalytic sites of RuBisCO are part of RBCL (Andersson, 2008; Spreitzer & Salvucci, 2002), RBCL has been considered as the main subunits that determine the catalytic properties and adaptive evolution of RuBisCO. Several studies using the *rbcL* gene have offered the evidence that RuBisCO was involved in the adaptive evolution of photosynthesis by detecting positive selection on *rbcL* genes in independent C4 lineages (Christin et al., 2008a; Kapralov & Filatov, 2007). The evidence of evolution of the photosynthesis has been intensively examined by studying the evolution of *rbcL*. However, previous scientific studies have suggested that RBCS has influences on changing the catalytic efficiency, $CO_2$ specificity, activity, quantity, assembly, and stability of RuBisCO (Andrews & Ballment, 1983; Bracher, Starling-Windhof, Ulrich Hartl, & Hayer-Hartl, 2011; Furbank et al., 2000; Genkov & Spreitzer, 2009; Genkov, Meyer, Griffiths, & Spreitzer, 2010; Spreitzer, 2003). Besides that, it has been suggested that RBCS has regions that have high affinity to $CO_2$ on the surface which enable captured $CO_2$ to migrate to the catalytic sites of the closest RBCL (van Lun, Hub, van der Spoel, & Andersson, 2014). Also, it has been discussed that the availability of RBCS up-regulates the transcript levels of *rbcL* and increases the amount of RBCL (Suzuki & Makino, 2012). These studies suggest the importance of RBCS in the functions, regulations, and protein structure of RuBisCO; however, the evolution of RBCS has been poorly studied. To understand the evolution of RBCS, it is necessary to improve our understanding about the evolution of the encoding *rbcS* gene. This thesis address three key questions: 1) how the *rbcS* multigene family evolved, 2) if the *rbcS* gene was involved in the evolution of photosynthesis, and 3) how the *rbcS* gene copies differ.

In Chapter 1, I study the phylogenetic relationships of the *rbcS* gene copies among angiosperms to understand the dynamics of *rbcS* evolution. The focus of previous studies about the evolution of *rbcS* is limited within genera (Shown in *Flaveria, Musa, Triticum aestivum, Zea mays* and Solanaceae; Kapralov, Kubien, Andersson, & Filatov, 2011; O'Neal, Pokalsky, Kiehne, & Shewmaker, 1987; Pichersky & Cashmore, 1986; Sasanuma,

2001; Thomas-Hall et al., 2007; Wolter, Fritz, Willmitzer, Schell, & Schreier, 1988). Therefore, I expand the knowledge of the *rbcS* evolution among genera. It has already been observed that the number of gene copies is different between species (Kapralov et al., 2011; Picersky & Cashmore, 1986; Thomas-Hall et al., 2007; Wolter et al., 1988). However, there has been minimal discussion of the causes that may have led to the copy number variation and the differences of characteristics between the *rbcS* gene copies within and between species. The previous study of Yang et al. (2016) has shown that different isoforms can interact differently with other proteins. Therefore, firstly, I hypothesized that each *rbcS* copy may encode different isoforms of RBCS and each isoform may interact differently with RBCL. I tested this hypothesis by comparing the pattern of coevolution between each *rbcS* and *rbcL*, because the coevolution test can detect sites interacting between proteins (i.e. subunits) (Hakes, Lovell, Oliver, & Robertson, 2007). Secondly, I hypothesized that a specific isoform of RBCS encoded by a specific *rbcS* gene copy may increase the activity of RuBisCO by formatting the structure of the enzyme that is optimal for active sites of RBCL. I tested this hypothesis by modelling a RuBisCO structure composed of all RBCS encoded by a single *rbcS* gene copy. Then, I compared the stability between modelled structures. I assumed that the differences of stability of modelled structures help to infer the differences of activities of the enzyme because the trade-off between activity and stability of RuBisCO has already been shown (Studer et al., 2014).

In Chapter 2, I hypothesized that the *rbcS* gene has been involved in the shift of photosynthetic types, because it has been suggested that RBCS may influence the catalytic properties of RuBisCO (Andrews & Ballment, 1983; Genkov & Spreitzer, 2009; Genkov et al., 2010, Spreitzer, 2003) and also the *rbcL* gene encoding the counterpart subunits of RBCS has undergone positive selection during C4 evolution. The advantageous mutations of genes that are fixed by positive selection play an important role in adaptation evolution (Uecker & Hermisson, 2011). Thus, I tested positive selection acting on *rbcS* genes in plants of different types of photosynthesis.

In Chapter 3, I made two hypotheses about the gene expressions of *rbcS* gene as follows: i) each copy of the same species expresses differently; ii) multiple copies of *rbcS* in a species may exist to maintain the amount of gene products at the same level as for single copy carrying species (the dosage effect hypothesis, as explained in Rice & McLysaght, 2017; Zuo et al., 2016). Previously, the total expression of *rbcS* has already been tested in different conditions (e.g. temperature, $CO_2$ concentration, water deficit, light regulation), tissues, developmental stages, and cellular localizations (Cavanagh & Kubien, 2014; Dean, Pichersky, & Dunsmuir, 1989; Hudsona et al., 1992; Manzara, Carrasco, & Gruissem, 1991; Morita, Hatanaka, Misoo, & Fukuyama, 2014; Thomas-Hall et al., 2007; Wanner & Guissem, 1991; Zhang et al., 2013). However, the comparison of expression levels between gene copies of the same species has been tested only in *Arabidopsis thaliana* at different temperatures and $CO_2$ concentrations (Cheng, Moore, & Seemann, 1998; Yoon et al., 2001). I used published transcriptome data to test my hypotheses. To investigate the first hypothesis I compared the gene expression levels between gene copies within species under control conditions in several tissues, and at different environmental conditions. To test the second hypothesis, I compared the gene expression levels between species under control conditions.

In my Ph.D. thesis, I fill the gap in scholarly knowledge about the evolution of *rbcS.* Testing the phylogenetic relationships of *rbcS*, the involvement of *rbcS* in C4 evolution, the difference between *rbcS* gene copies in coevolution, protein stability, and gene expression levels will give better insights about the dynamics of *rbcS* evolution. Also, it will contribute to a better understanding of the evolution of RuBisCO.

# Chapter 1. Evolutionary history of *rbcS* and the interaction of *rbcS* and *rbcL* in angiosperms

## Introduction

Gene duplication is one of the main mechanisms that can create novel features at the molecular level during evolution (Flagel & Wendel, 2009). The functional role played by duplicated genes has been discussed in detail (Hughes, 1994; Lynch & Force, 2000) and the mechanisms at work in this process are now relatively well understood (Hughes, 1994; Innan & Kondrashov, 2010; Rensing, 2014; Roulin et al., 2012; Studer, Penel, Duret, & Robinson-Rechavi, 2008). At the molecular level, it was initially proposed that relaxation of the selective constraints on one of the gene copies following gene duplication allows an accumulation of mutations that can permit the evolution of novel or sub-gene function or lead to a total loss of function (Moore & Purugganan, 2005; Ohta, 1988; Wagner, 1998). However, the advantages brought by gene duplication could not only stem from the effects of mutations but also from the protection it provides against deleterious mutations or the mechanisms of the dosage effect (Cheeseman et al., 2016; Kafri, Dahan, Levy, & Pilpel, 2008; Papp, Pál, & Hurst, 2003).

The creation of new gene copies by duplication is further affected by species divergence and the evolutionary history of the resulting gene family. Members of most gene families are therefore connected by a complex history of duplication and speciation events that have produced paralogous and orthologous gene copies. Paralogous copies are diverged from a single ancestor by duplication. Orthologous copies are the copies that have diverged from a common ancestral gene by speciation. The identification of the proper sets of orthologous genes is not an easy task (Altenhoff, Schneider, Gonnet, & Dessimoz, 2011). Correct identification of relationships of gene copies is further complicated by the presence of gene

conversion that will alter the origin of similarities between homologous regions (Mansai & Innan, 2010; Song et al., 2012).

A multigene family is a group of genes in which each member may have experienced evolutionary processes such as duplication and/or selection of advantageous mutations. Their genetic information can be exchanged between them by crossing over, recombination, and/or gene conversions (Dumont & Elchler, 2013; Mano & Innan, 2008; Nei & Rooney, 2005; Ohta 1977, 1979, 1983). Each gene copy can differ not only by the evolutionary process but also by function, cellular localization of encoding protein, stability, and/or expression levels (Clark et al., 2001; Hudsona et al., 1992; Ku et al., 1996; Niimura, 2009; Petter et al., 2008). Different gene copies of a multigene family can play a core role in organizing the novel or modified functions that are often required during adaptive evolution (Mcglothlin et al., 2016; Nei et al., 1997; Niimura, 2009; Ohta, 1991).

An example of this adaptive evolution is the evolution of photosynthesis. Atmospheric $CO_2$ drastically decreased in the Oligocene (Beerling & Royer, 2011; Edwards et al., 2010; Pearson et al., 2009). To adapt to depleted $CO_2$ concentration, some species have evolved to carry a mechanism to concentrate $CO_2$ by modifying the biochemical cascade and the cellular structures (Sage, 2004). One of the main groups of such species is comprised of C4 plants that have diverged from classical C3 plants. For the C4 type of photosynthesis, new enzymes are required. Most of the C4-specific enzymes are encoded by multigene families and the co-option of pre-existing genes of C3 type plays an important role during the transition from C3 type to C4 type (Christin et al., 2013).

Ribulose-bisphosphate carboxylate/oxygenase small subunit (*rbcS*) is a multigene family encoding a small subunit (RBCS) of RuBisCO, the first enzyme of the Calvin-Benson cycle to fix $CO_2$ to sugar (Hatch & Slack, 1968; Kanai & Edwards, 1999). RuBisCO has slower catalytic efficiency than other photosynthetic enzymes because it has the affinity to both $O_2$ and $CO_2$ (Rawsthorne, 1992) that results in a loss of energy and $CO_2$ (Kubien et al. 2008; Peterhansel et al. 2010). Thus, it has been considered to be the limiting factor of the

photosynthetic rate in higher plants (Hudson, Evans, von Caemmerer, Arvidsson, & Andrews, 1992; Von Caemmerer, Millgate, Farquhar, & Furbank, 1997). In C4 plants, the $CO_2$-concentrating mechanism (CCM) enabled RuBisCO to be surrounded by highly concentrated $CO_2$. Thus, the catalytic efficiency of RuBisCO became better in C4 plants than C3 plants (Badger & Andrews, 1987; Sage & Coleman, 2001; von Caemmerer & Quick, 2000). Therefore, RuBisCO has been considered as the key enzyme in the adaptive evolution of photosynthesis.

The evidence for the adaptive evolution of RuBisCO has been shown through the study of the evolution of the chloroplast *rbcL* gene encoding RBCL, the other subunit of RuBisCO. Positive selection for *rbcL* has been detected in independent C4 lineages (Christin et al., 2008a; Kapralov & Filatov, 2007). The signal of positive selection of the *rbcL* gene is almost 20 times stronger than that detected for *rbcS* in *Flaveria* (Kapralov, Kubien, Andersson, & Filatov, 2011). The RBCL subunit is considered to determine the catalytic properties of RuBisCO because it contains the catalytic site of the enzyme (Andersson, 2008). Therefore, the evolution of the *rbcL* gene has attracted greater research attention than that of the *rbcS* gene. However, RBCS has been reported to have an influence on the catalytic efficiency, $CO_2$ specificity, activity, quantity, assembly, and stability of RuBisCO (Andrews & Ballment, 1983; Bracher, Starling-Windhof, Ulrich Harti, & Hayer-Harti, 2011; Furbank et al., 2000; Genkov & Spreitzer, 2009, Genkov, Meyer, Griffiths, & Spreitzer, 2010; Spreitzer, 2003). Studer et al. (2014) have suggested that some positively selected codons encoding amino acid residues that are located at the interface between RBCL and RBCS may affect the stability and the catalytic properties of RuBisCO. All these studies suggest that the interaction between RBCS and RBCL, and the *rbcS* gene itself may play important roles in the evolution of RuBisCO.

A better understanding of the evolutionary history of *rbcS* is thus essential to obtain a deeper insight into the evolution of RuBisCO. We extracted the *rbcS* sequences from available full genomes of angiosperms and reconstructed the phylogenetic relationships of the *rbcS* gene

copies. We then tested the positive selection acting on the *rbcS* gene in angiosperms. Positive selection of *rbcS* has already been tested within some genera but has never been tested on a wider sample range of plant groups. Therefore, we aim to elucidate the differences between gene copies of *rbcS* in higher plants and to infer their respective evolutionary histories. Firstly, we hypothesized that each *rbcS* copy may have a different interaction with *rbcL.* We investigated this hypothesis by testing coevolution between *rbcS* and *rbcL*. Secondly, we hypothesized that RBCS encoded by different *rbcS* gene copies may have a different degree of influence on the stability of RuBisCO. We tested this by modelling a RuBisCO structure with eight RBCSs encoded by a unique *rbcS* copy. We did the same for each *rbcS* copy and compared the stability between models. Our study provides new insights into the evolutionary mechanism of the *rbcS* multigene family and sheds light on its influence on RuBisCO evolution.

# Materials and Methods

## Phylogenetic tree of *rbcS* among angiosperms

We downloaded the genomic data of all angiosperms available in Phytozome v10 (https://phytozome.jgi.doe.gov/pz/portal.html). All the *rbcS* gene copies annotated for *Arabidopsis thaliana* were extracted from UniProt. We then used the Exonerate software (Slater & Birney, 2005) to retrieve, from the genomic data of each angiosperm species, the gene regions that were homologous to the *rbcS* of *A. thaliana*. We used DNA for both the query and target sequences and used the “coding2genome” algorithm available in Exonerate. According to the lengths of exons and introns of each gene copy on Phytozme database, we set the maximum length of introns to 1,000 base pairs. According to the similarities between gene copies on Phytozome database, we filtered the sequences that have more than 50 % of identity between the query and the target sequence, then, we selected only the best ten hits among filtered sequences. We translated nucleotide sequences to amino acid sequences using the translate tool of ExPaSy

(http://web.expasy.org/translate/). We aligned the sequences obtained using MAFFT (Katoh & Standley, 2013) and removed unreliable sequences that were poorly aligned using GUIDANCE2 with default settings (http://guidance.tau.ac.il/ver2/; Sela, Ashkenazy, Katoh, & Pupko, 2015). We then converted these amino acid alignments back into codon alignment using PAL2NAL (http://www.bork.embl.de/pal2nal/#RunP2N) to obtain the final nucleotide alignment of 90 *rbcS* gene copies for 33 angiosperm species. The GTR+G model of substitution was identified as the best model using Jmodeltest 2.1.4 (Darriba, Taboada, Doallo, & Posada, 2012). We reconstructed the phylogenetic tree with PhyML version 3.0 (Guindon & Gascuel, 2003) using the BEST algorithm for tree rearrangement, while estimating all parameters of the GTR+G model and the branch lengths. Branch support values were estimated based on 1,000 bootstrap replicates.

## Gene conversion

We tested for the signatures of recombination and gene conversion in the *rbcS* gene copies using the Recombination Detection Program v4.56 software (RDP4; Martin, Murrell, Golden, Khoosal, & Muhire, 2015). We used Chimaera, 3seq, GENECONV, MaxChi, and SiScan with their default parameters. The nucleotide alignment created for the phylogenetic reconstruction was used as an input for the gene conversion analyses.

## Selection

Positive selection analysis in *rbcS* was performed using the mixed effects model of evolution (MEME) implemented in HyPhy version 2.2.6 (Pond, Frost, & Muse, 2005). We used the MG94 codons substitution base model (Muse, Gaut, & Carolina, 1994) and false discovery rate (FDR; Benjamini et al., 1995) with a threshold of 0.1 to correct for multiple testing. We selected the MEME model because it is more suitable than CodeML of PAML (Yang, 2007) for estimating site-specific probabilities (Lu et al., 2013). Positions under positive selection were plotted on the known protein structure of *Oryza sativa* (Chain C of 1WDD in the Protein Data Bank; Matsumura et al., 2012) using the software PyMol version 1.3 (Schrödinger, LLC, 2015).

## Coevolution between *rbcS* and *rbcL*

We used the *rbcL* sequences from the alignment of Christin et al. (Christin et al., 2008a), but retained only the 33 species for which both *rbcS* and *rbcL* sequences were available. The resulting alignment was 1,342 base pairs long.

Coevolution analysis of *rbcS* and *rbcL* was performed using the maximum likelihood implementation of model Coev (Dib, Silverstro, & Salamin, 2014; Dib et al., 2015). For each pair of sites we compared the dependent and independent models of substitution using the Akaike Information Criterion (AIC). The difference in AIC (dAIC) between the two models varies depending on the tree structure and characteristics of the alignment. Null distribution of dAIC was obtained by simulating sequences under independent substitution model (for the details, see Dib et al., 2014). Only sites with a ratio between the parameters s and d (i.e. s/d) > 10 were used (Dib et al., 2014). Since *rbcL* has a single gene copy per species and *rbcS* shows a variable copy number between species, we duplicated the *rbcL* sequences to match the *rbcS* copy number. The *rbcS* and *rbcL* alignments were concatenated into one matrix and conserved positions with the identity more than 95% were removed to minimize the number of computations according to the method explained in the developer's articles (Dib et al., 2014, 2015). The final concatenated alignment of *rbcL*/*rbcS* contained 541 nucleotide positions (330 bp of *rbcL* and 211 bp of *rbcS*), which led to a total of 69,630 tests of coevolution for pairs. In every pair, one of the sites belonged to *rbcL*, while the other belonged to *rbcS*. Coevolving profiles were visualized using the R package qgraph (Epskamp, Cramer, Waldorp, Schmittmann, & Borsboom, 2012). Pairs of positions that passed the dAIC and s/d ratio thresholds were plotted on the known protein structure of *O. sativa* (1WDD of Protein Data Bank: Matsumura et al., 2012) using PyMol version 1.3 (Schrödinger, LLC, 2015).

## Protein stability of RuBisCO structure

The RuBisCO quaternary structure is a hexadecamer composed of eight subunits of RBCL and eight subunits of RBCS. Since RBCL is encoded by a single gene, the eight RBCL subunits are always the same for a species. On the other hand, the exact combination of the eight RBCS subunits is not known. We assumed here that, for a given RuBisCO protein, the eight RBCS subunits are encoded by the same copy of *rbcS*. This assumption made the modelling of protein stability feasible by limiting the number of combinations and allows us to study differences between gene copies.

We performed homology modelling and estimated the Gibbs free energy to compare the stability of the whole RuBisCO structures. Gibbs free energy indicates differences of energy during a chemical reaction. We used in our case Gibbs free energy as the difference of thermodynamic stability between the folded and unfolded states of a protein. When this measure is below 0, the folded state is preferred over the unfolded state and protein models with smaller value of Gibbs free energy can be considered to be more stable. To model the RuBisCO stability in angiosperms, the RBCS and RBCL amino acid sequences of several species of Brassicaceae and Poaceae were downloaded from UniProt (UniProt Consortium, 2015). We selected these two clades because they are well defined in the *rbcS* phylogenetic tree and are representative of the evolution of *rbcS* (see results). To create RBCS encoded by a single gene copy, we duplicated eight times the *rbcS* sequence in each pair protein structure file. However, when different gene copies of the same species differed only by synonymous substitutions or when amino acids differ in region outside the crystallized structure, only one complex was tested for these gene copies since amino acid sequences were identical (e.g. *Setaria italica* copies 4 and 5 in Table 4). Homology modelling was performed using Modeller 9.17 (Eswar, Eramian, Webb, Shen, & Sali, 2008). The RuBisCO structure of *O. sativa* (1WDD of Protein Data Bank; Matsumura et al., 2012) was used as a template. The homology modelling was run 100 times for each structural complex of *rbcL*/*rbcS* and the best model (the one with the lowest DOPE score) was selected for further

analyses. These models were then repaired with FoldX 4.0 (Schymkowitz et al., 2005) using the RepairPDB function. The repair step is mandatory for removing potential bad contacts (i.e. Van der Waals clashes) in the structures, which may cause instability of modelled protein. Also using FoldX 4.0, we predicted the differences of Gibbs free energy between maximum likelihood model and null model (ΔG) of each estimated structure using the "Stability" function, with default parameters. Three-dimensional structures were visualized with PyMol 1.3 (Schrödinger, LLC, 2015). Estimated ΔG for the Brassicaceae and Poaceae were visualized on their respective *rbcS* gene trees using the function phenogram of the R package phytools (Revell, 2012).

# Results

## Phylogenetic tree of *rbcS* among angiosperms

We downloaded all the available genomes from Phytozome v10 and extracted *rbcS* copies of each angiosperm species present in the database using the four *rbcS* genes of *A. thaliana* as target sequences. The phylogenetic tree of the 90 *rbcS* gene sequences available for 33 species is shown in Figure 1. Each plant family is well defined with subtending branches well supported (bootstrap support > 79%; Figure 2), except for the two families Caricaceae and Rosaceae. The relationships obtained within each family or subfamily is further well supported. Globally, the topology of the gene tree follows the expected species tree of angiosperms (e.g. clear division between monocots and eudicots; see Magallón, Gómez-Acevedo, Sánchez-Reyes, & Hernández-Hernández, 2015) but the relationships between several plant families in eudicots (i.e. Rosaceae, Linaceae, Salicaceae, Malvaceae, Solanaceae, and Faboideae) were not supported by high bootstrap values (Figure 2). The low support obtained could be due to short branch lengths and the peculiar evolutionary history of the *rbcS* gene (see below). The *rbcS* gene tree estimated by PhyML shows a particular topology with the gene copies of the same species clustering together with high bootstrap support (Figure 1). The phylogenetic analyses showed also deeper duplication

events in several plant families (e.g. Brassicaceae, Rosaceae) and there are a few exceptions, such as *Ricinus communis* and *Eucalyptus grandis*, that have gene copies widely spread across the tree. This same pattern was observed for the trees based on the amino acid data (Figure 3) and the third codon position (Figure 4).

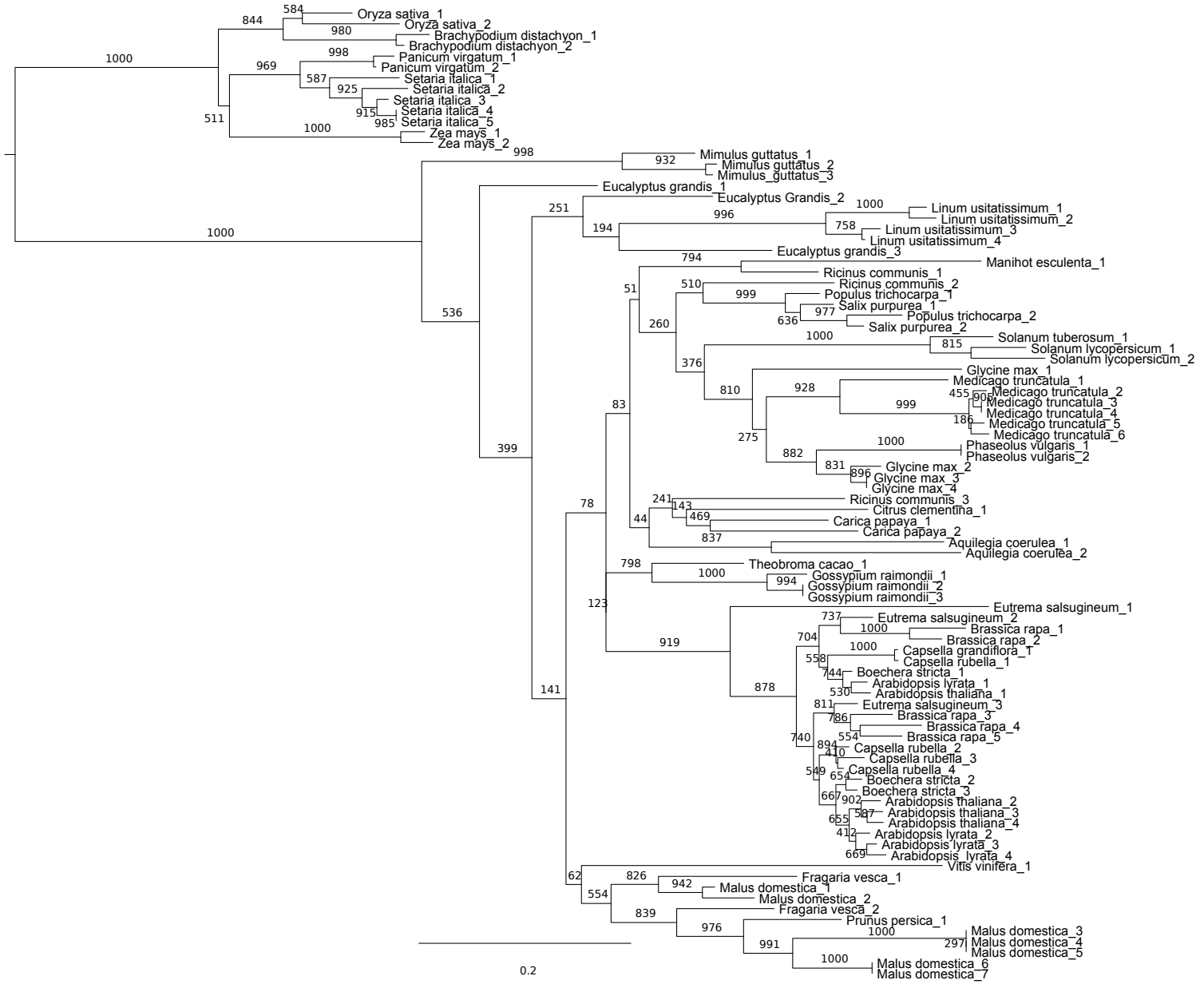



**Figure 1. Maximum likelihood tree of *rbcS* in angiosperms**

The phylogenetic tree was reconstructed in PhyML3.0 (Guindon et al., 2003) using a GTR+G model. Branch support was estimated using 1,000 bootstraps replicates. Each gene copy of a given species is identified by the species name and distinguished by a number. The scale bar is shown below the tree.

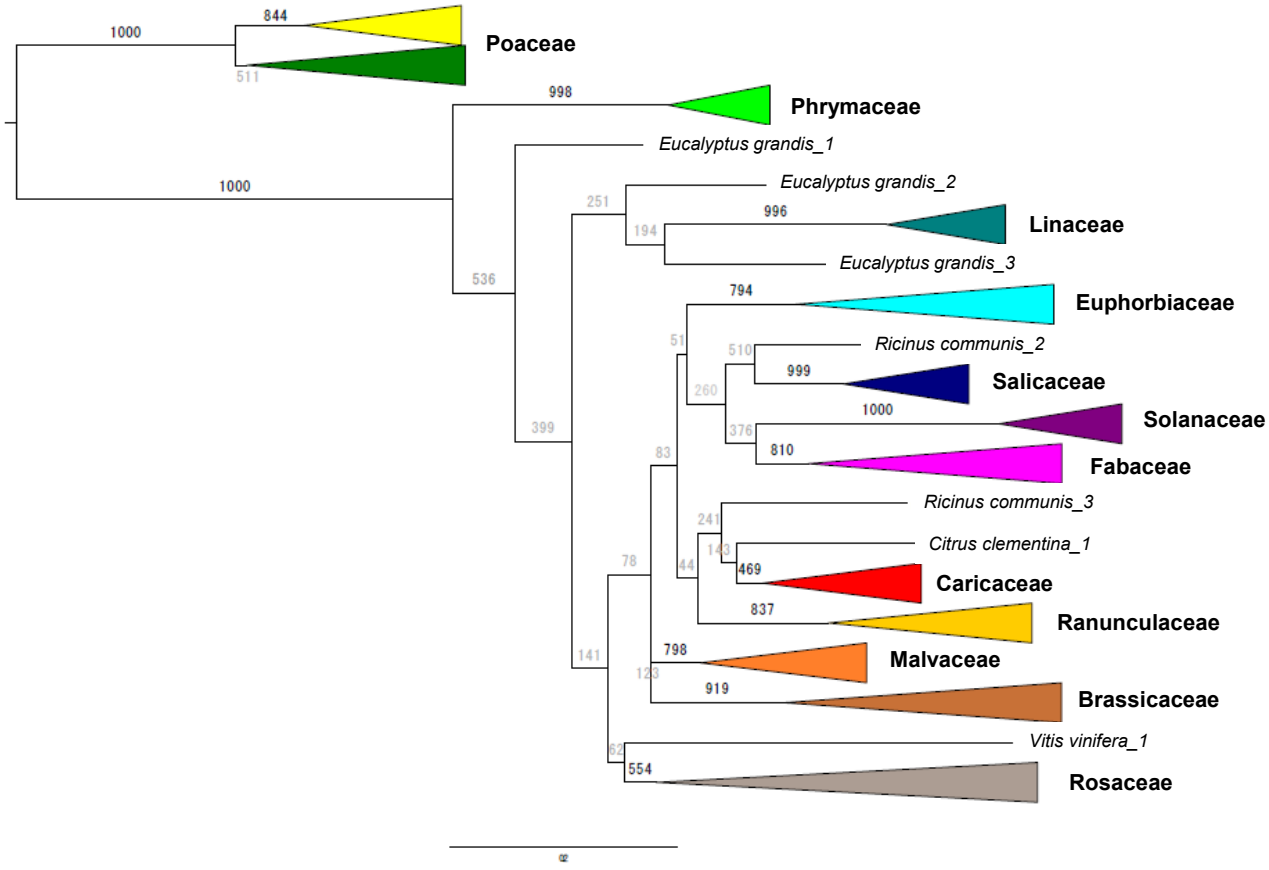



**Figure 2. Collapsed maximum likelihood tree of *rbcS* in angiosperms**
The phylogenetic tree was reconstructed in PhyML3.0 (Guindon et al., 2003) using a GTR+G model. Branch support was estimated using 1,000 bootstraps. Highly supported branches are shown in black and weakly supported branches are shown in gray. The clade of each family is collapsed. A distance scale of tree is shown below the phylogeny.

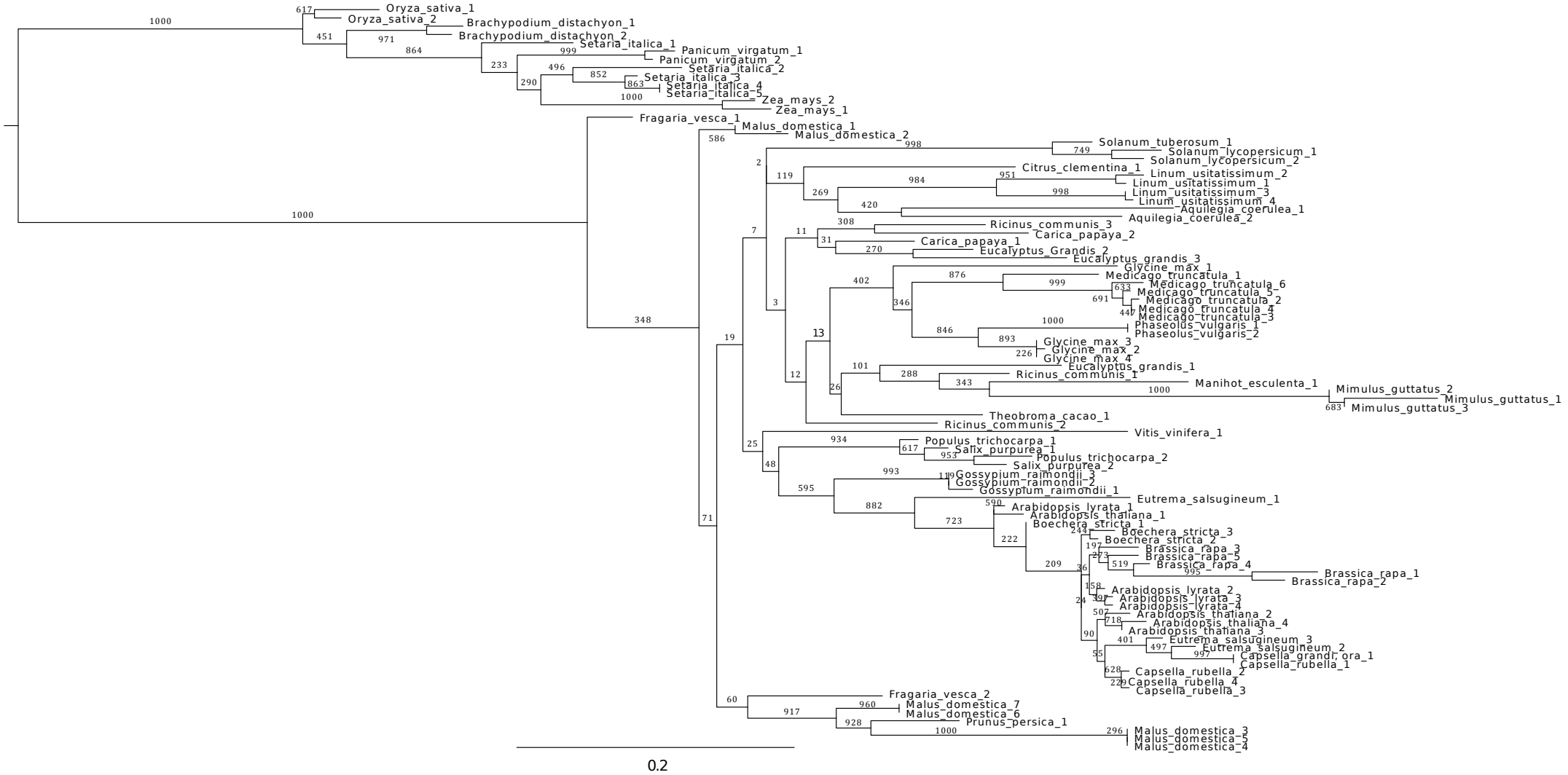



**Figure 3. Maximum likelihood tree of *rbcS* based on translated amino acid sequences**

The alignment of *rbcS* used to reconstruct Figure 1 was translated into amino acid sequences. The phylogenetic tree based on amino acid sequences was reconstructed using PhyML3.0 (Guindon et al., 2003) with a LG model. Branch support was estimated using 1,000 bootstrap replicates. The scale bar is shown below the tree.

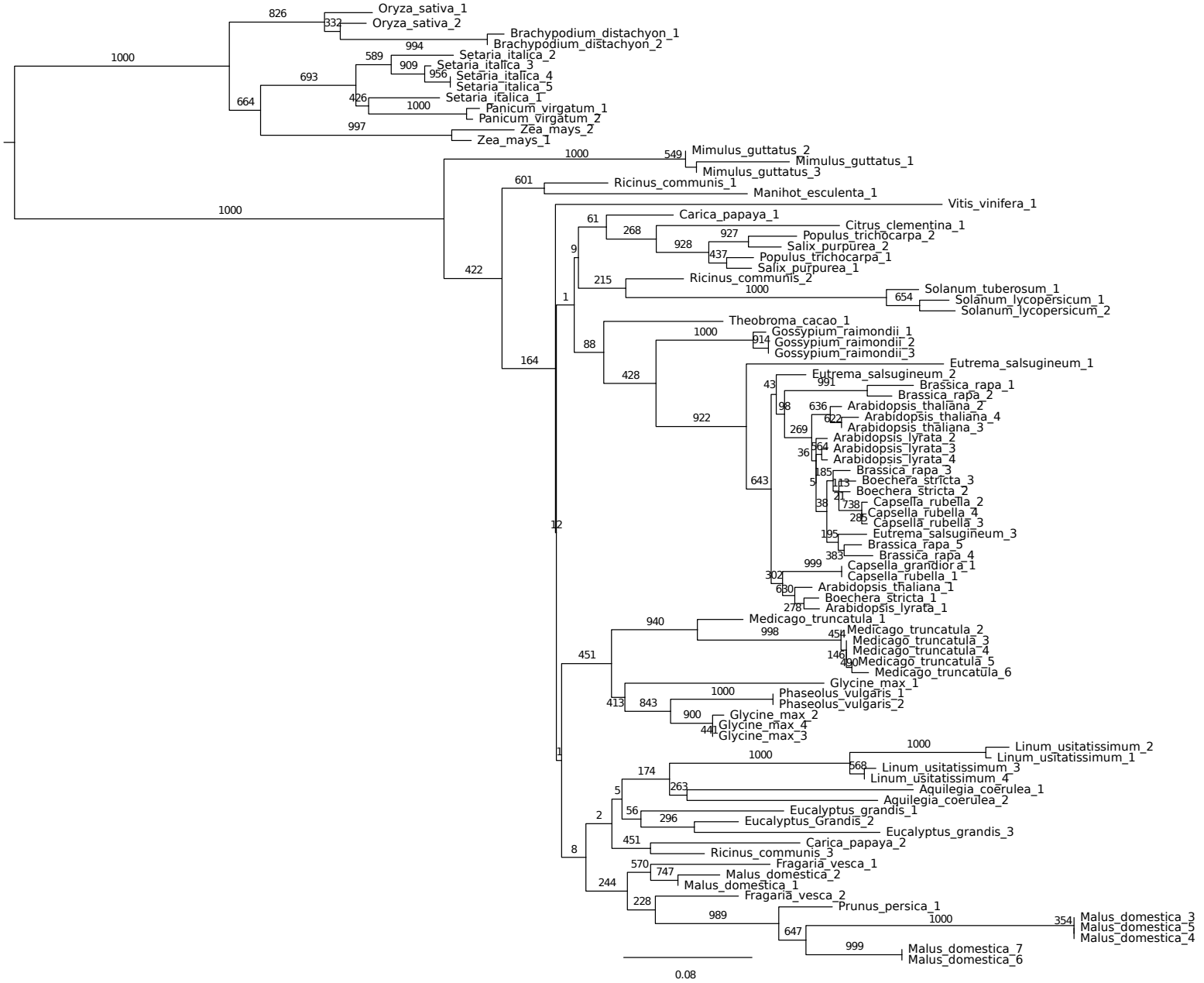



**Figure 4. Maximum likelihood tree of *rbcS* based on nucleotides and excluding 3rd codon positions from the**

We removed the 3rd codon positions from the alignment used in Figure 1 and we reconstructed the phylogenetic tree using PhyML3.0 (Guindon et al., 2003) with a GTR+G model. Branch support was estimated using 1,000 bootstraps replicates. The scale bar is shown below the tree.

We identified two *rbcS* lineages (*rbcS*-lineage1 and *rbcS*-lineage2) that are supposed to have originated from a duplication event before the divergence of eudicots and monocots. One gene lineage includes genes that cluster together with a known expressed gene in *OsRbcS2* (Morita et al., 2014) in photosynthetic organs; we refer to this gene lineage as *rbcS*-lineage1 (Figure 1). The second gene lineage includes gene copies expressed in non-photosynthetic organs such as *OsRbcS1* (Morita et al., 2014); we refer to this as *rbcS*-lineage2 (Figure 5). We excluded 20 sequences of *rbcS*-lineage2 from further analysis because our focus is on the molecular evolution of the gene copies involved in photosynthesis.

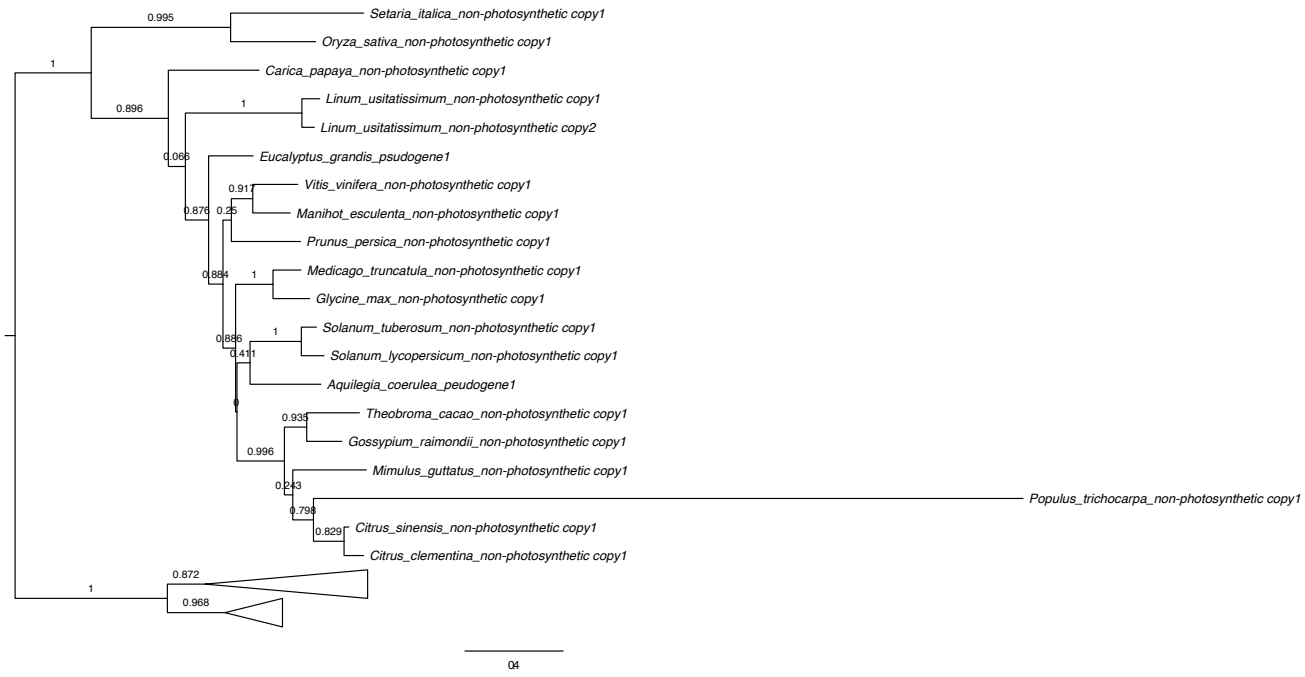



**Figure 5. Phylogenetic relationships of *rbcS*-lineage2**
We reconstructed phylogenetic relationships with gene copies including unusual RBCS. Collapsed clades are functional copies of eudicots (below) and monocots (above), respectively. Details of these collapsed clades are shown in Figure 1.

Minimum numbers of gene copies per species are shown in Table 1. The number of *rbcS* gene copies varies depending on the species. We detected a minimum of one to a maximum of seven gene copies per species using our method. However, we should note that there is a limitation of our method to detect the exact copy number because some existing gene copies such as tandem copies may have been missed during the process of assembly (Panchy, Lehti-Shiu, & Shiu, 2016). Further, the reported number of gene copies may change by using newly released genomes with an improved method of assembly.

**Table1. Minimum number of *rbcS* gene copies per species in angiosperms**

| Species name | Minimum number of *rbcS* copies |
|---|---|
| *Aquilegia coerulea* | 2 |
| *Arabidopsis lyrata* | 4 |
| *Arabidopsis thaliana* | 4 |
| *Boechera stricta* | 3 |
| *Brachypodium distachyon* | 2 |
| *Brassica rapa* | 5 |
| *Capsella grandiflora* | 1 |
| *Capsella rubella* | 4 |
| *Carica papaya* | 2 |
| *Citrus clementine* | 1 |
| *Eucalyptus grandis* | 3 |
| *Eutrema salsugineum* | 3 |
| *Fragaria vesca* | 2 |
| *Glycine max* | 4 |
| *Gossypium raimondii* | 3 |
| *Linum usitatissimum* | 4 |
| *Malus domestica* | 7 |
| *Manihot esculenta* | 1 |
| *Medicago truncatula* | 6 |
| *Mimulus guttatus* | 3 |
| *Oryza sativa* | 2 |
| *Panicum virgatum* | 2 |
| *Phaseolus vulgaris* | 2 |
| *Populus trichocarpa* | 2 |
| *Prunus persica* | 1 |
| *Ricinus communis* | 3 |
| *Salix purpurea* | 2 |
| *Setaria italica* | 5 |
| *Solanum lycopersicum* | 2 |
| *Solanum tuberosum* | 1 |
| *Theobroma cacao* | 1 |
| *Vitis vinifera* | 1 |
| *Zea mays* | 2 |

## Positive selection

We tested *rbcS* sequences for the signs of positive selection using the MEME model of Hyphy (Pond, Frost, & Muse, 2005). A strong signal of positive selection was detected in 15 sites (Table 2; Figure 6). The P-value and the q-value, the adjusted p-values using an optimized FDR approach of each site are shown in Table 2. The episodes of positive selection were not associated with specific branches or duplication events.

**Table 2. Nucleotide sites of *rbcS* under positive selection and corresponding amino acid residues of RBCS**

| Sites in our nucleotide alignment | Corresponding amino acid residues in RuBisCO structure of *Oryza sativa* (1WDD of Protein Data Bank) | p-value | q-value |
|---|---|---|---|
| 190 | 5 | 9.56E-05 | 3.01E-03 |
| 223 | 16 | 2.34E-04 | 3.14E-03 |
| 238 | 21 | 5.78E-04 | 5.20E-03 |
| 241 | 22 | 1.89E-02 | 8.50E-02 |
| 271 | 32 | 3.60E-03 | 2.06E-02 |
| 274 | 33 | 2.16E-02 | 9.07E-02 |
| 286 | 37 | 1.03E-02 | 4.98E-02 |
| 331 | 52 | 1.92E-04 | 3.14E-03 |
| 394 | 73 | 2.89E-03 | 1.82E-02 |
| 415 | 80 | 7.85E-04 | 5.49E-03 |
| 418 | 81 | 2.99E-04 | 3.14E-03 |
| 433 | 86 | 2.52E-04 | 3.14E-03 |
| 439 | 88 | 7.30E-04 | 5.49E-03 |
| 484 | 103 | 1.54E-05 | 9.71E-04 |
| 499 | 108 | 5.23E-03 | 2.74E-02 |

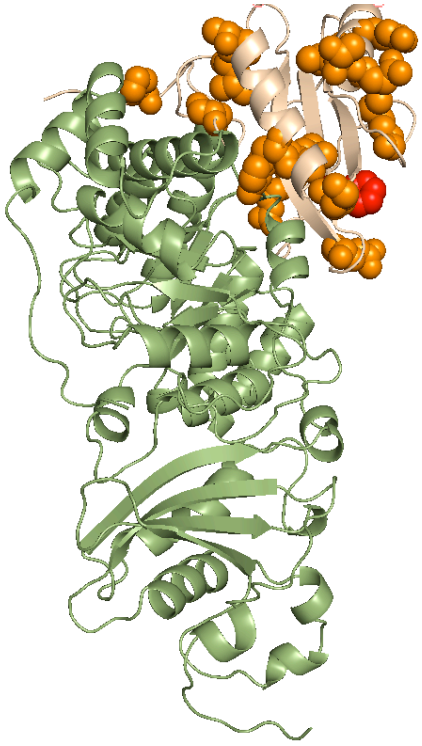


**Figure 6. RBCS residues under positive selection**

Fifteen positions of *rbcS* showed strong signals of positive selection. We plotted corresponding amino acid residues to RuBisCO structure of *Oryza sativa* (1WDD of Protein Data Bank). Pink cartoon ribbons indicate RBCS, chain C of 1WDD. Green cartoon ribbons indicate RBCL, chain A of 1WDD. Orange spheres are positions under positive selection. Red spheres show positions under positive selection and also under coevolution with *rbcL*.

## Coevolution between *rbcS* and *rbcL*

We tested a total of 69,630 pairs of sites to detect coevolution between *rbcS* and *rbcL*. Signal of coevolution was, as expected, pervasive between these two genes and 12,410 pairs of sites had a dAIC value between the null and alternative model higher than the threshold of 6.85 estimated by simulations (Figure 7-a). Among these 12,410 pairs, we further looked at the strength of the signal by considering the ratio of the parameters s and d, which indicates a strong signal if its value is higher than 10 (Dib et al., 2014, 2015). The distribution of s/d ratios is shown in Figure 7-b and we identified 15 pairs with an s/d ratio higher than 10 (Table 3). We found that four of these 15 positions along the *rbcS* sequence (positions 66, 75, 87, and 441; Table 3) were each coevolving with multiple positions of *rbcL* and these multiple positions were mostly spread to the whole region of *rbcL* (Figure 8). In contrast, only one position of *rbcL* (position 463; Table 3) was found to be coevolving with multiple positions of *rbcS*.

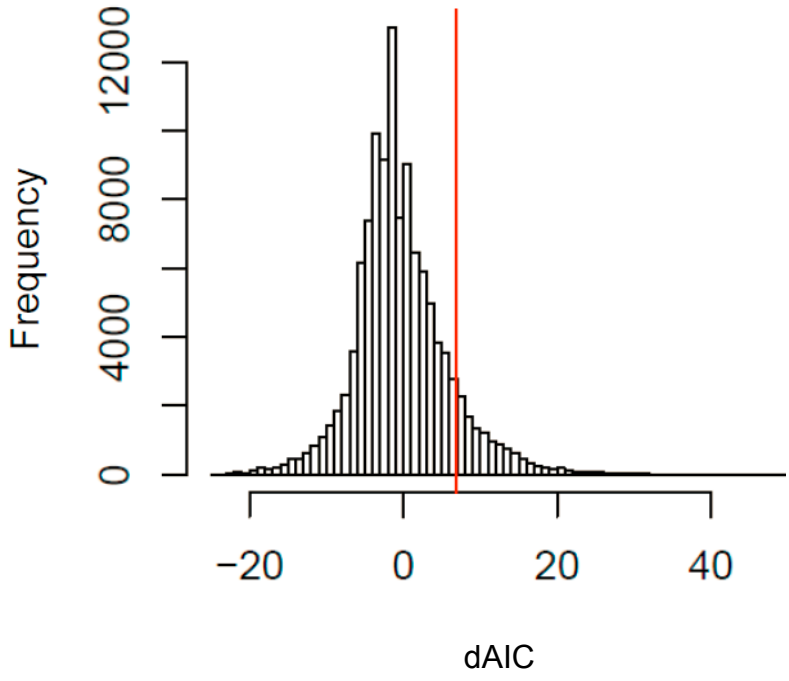



**Figure 7-a. dAIC distribution of frequency of coevolving profiles by Coev model**
We tested coevolution on 69,630 pairs of positions between *rbcS* and *rbcL* by Coev (Lib et al., 2014, 2015). A total of 12,410 pairs passed the threshold of significant dAIC (6.85; red line) estimated by the Coev model.

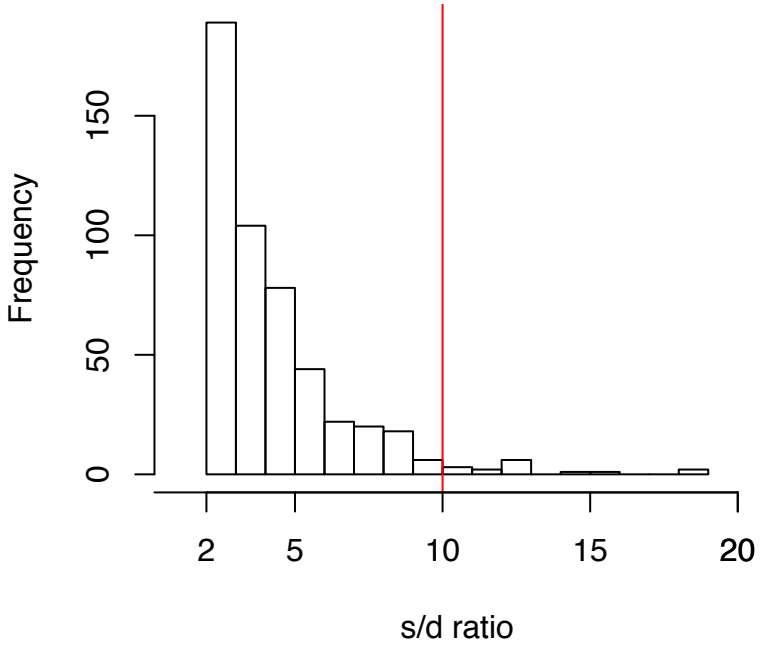



**Figure 7-b. s/d ratio distribution of frequency of coevolving profiles by Coev model**

We tested coevolution on 69,630 pairs of positions between *rbcS* and *rbcL*. A total of 12,410 pairs passed the threshold of significant dAIC, 6.85. The s/d ratio over 10 suggests a strong signal of coevolution (Lib et al., 2014, Lib et al., 2015). Pairs with s/d ratios of greater than 2 are shown in this figure.

**Table 3. Coevolving sites between *rbcS* and *rbcL***

| *rbcS* | *rbcL* | Profile | dAIC (difference of the AIC values between coev model and null model) | s/d |
|---|---|---|---|---|
| 66 | 463 | CA,TC | 8.97848 | 12.05963 |
| 66 | 1195 | AA,GC | 8.97848 | 12.05963 |
| 75 | 1321 | CA,TG | 11.27012 | 11.2801 |
| 75 | 94 | CA,TG | 11.27012 | 11.2801 |
| 87 | 46 | CC,TA | 11.97856 | 10.91736 |
| 87 | 142 | AC,GA | 7.88714 | 10.93568 |
| 87 | 1195 | AA,GC | 7.32326 | 18.18245 |
| 87 | 463 | CA,TC | 7.32326 | 18.18245 |
| 87 | 56 | CC,GA | 7.1302 | 12.40002 |
| 232 | 814 | AC,GT | 12.91864 | 15.57228 |
| 441 | 1001 | CC,TA | 8.7363 | 10.97703 |
| 441 | 662 | AG,GA | 6.87918 | 12.03816 |
| 441 | 85 | AG,GA | 6.87918 | 12.03816 |
| 441 | 796 | CA,TG | 6.87918 | 12.03816 |
| 484 | 1192 | CA,TC | 12.10548 | 14.60993 |

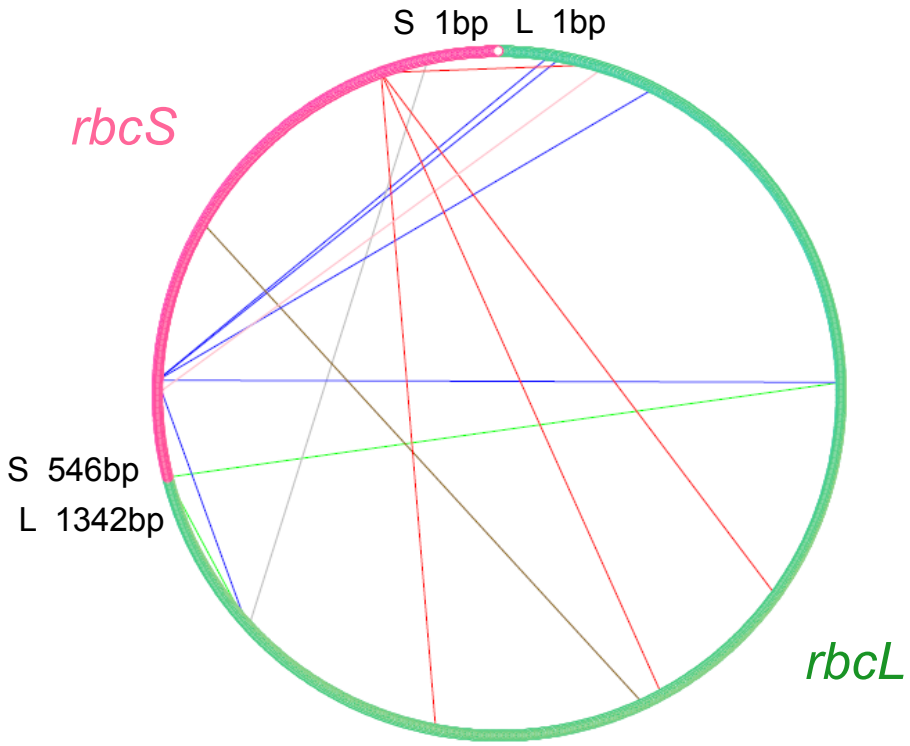
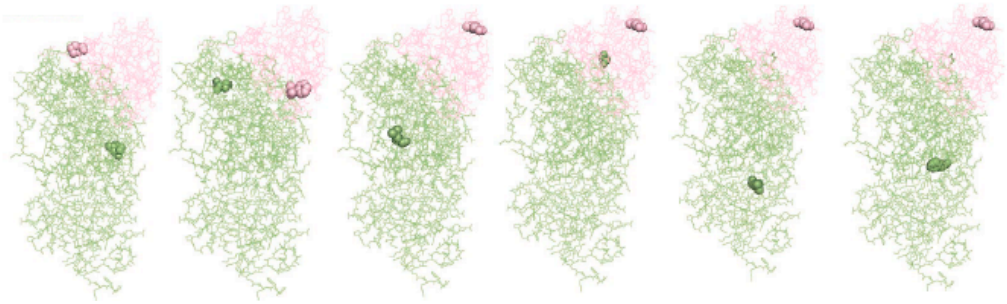



**Figure 8. Coevolving sites between *rbcS* and *rbcL***
Coevolution between paired combinations of *rbcS* sites and *rbcL* sites was estimated by a maximum-likelihood implementation of Coev and dependent model. The differences of AIC between pairs of models (dAIC) were calculated. AIC of the null model (=6.85) was used as threshold. We then filtered further with the s/d ratio threshold (=10) according to the previous studies (Lib et al., 2014, 2015). Fifteen profiles met the criteria and these sites were plotted using the qgraph function of R. Nucleotide sites of *rbcS* are shown in the pink bar and those of *rbcL* in the deep green bar. A coevolving profile set is connected with a line. Coevolving pairs that include the same *rbcS* sites are drawn in lines of the same colour.

Among the 15 pairs of coevolving sites, only six occurred in the part of the reference sequence that is present in the available protein structure of RuBisCO 1WDD (Figure 9). The largest dAIC for coevolution test was found at position 232 of *rbcS* (position 19 of reference sequence 1WDD) and position 814 of *rbcL* (position 280 of reference sequence 1WDD). Both residues were on the surface of each subunit. The second-largest dAIC belonged to position 484 of *rbcS* (position 103 of reference sequence 1WDD) and position 1,192 of *rbcL* (position 406 of reference sequence 1WDD). These residues were inside of each subunit. Both positions of RBCL (positions 814 and 1192) were listed as residues in the circular core between helix and strands in the alpha/beta-barrel in the large subunits of spinach RuBisCO structure (Knight, Andersson, & Brändén, 1990). The position 484 of *rbcS* (position 103 of reference sequence 1WDD) was detected under positive selection and is coevolving with the *rbcL* gene.



| | | | | | | |
|---|---|---|---|---|---|---|
| Sites in *rbcS* nucleotide sequence | 232 | 484 | 441 | 441 | 441 | 441 |
| Corresponding amino acid residues of RBCS (Chain C) in 1WDD of Protein Data Bank | 19 | 103 | 88 | 88 | 88 | 88 |
| Sites in *rbcL* nucleotide sequence | 814 | 1192 | 1001 | 662 | 85 | 796 |
| Corresponding amino acid residues of RBCL (Chain A) in 1WDD of Protein Data Bank | 280 | 406 | 343 | 230 | 37 | 274 |

**Figure 9. Coevolving positions of RBCS and RBCL plotted to RuBisCO protein structure (1WDD of PDB)**

Fifteen pairs of positions of *rbcS* and *rbcL* were detected as the sites undergoing coevolution. Sites 1–48 of our nucleotide alignment are not in the protein coding region, so they are not shown in this figure. Green ribbons indicate RBCL, chain A of 1WDD of Protein data bank. Pink ribbons indicate RBCS, chain C of 1WDD of Protein Data Bank. Spheres indicate coevolving positions. The table indicates the corresponding sites and residues in alignment.

## Protein stability of RuBisCO structure

Our phylogenetic analyses indicated that at least two plant families (Rosaceae and Brassicaceae; Figure 1) had old duplication events during their evolutionary history. In contrast, the Poaceae family did not show any signs of old duplication events (Figure 1). The large sequence divergence between gene copies in Brassicaceae could lead to a variable stability of the heterodimers formed with the single RBCL protein when different *rbcS* gene copies are involved. We therefore compared the characteristics of each gene copy from both the Brassicaceae and Poaceae by estimating the Gibbs free energy of the RuBisCO structure (Table 4).

In Poaceae, the Gibbs free energy values estimated were similar for gene copies of the same species (Figure 10; Table 4). There was also a clear distinction between the values for the Pooideae, represented by *Brachypodium distachyon*, and representatives of the PACMAD clade (*Z. mays* and *S. italica*). *O. sativa* was not included in our analysis because the amino acid sequences of each gene copy of *rbcS*-lineage1 are identical. In Brassicaceae, we expected differences of Gibbs free energy values between gene copies because their duplication is relatively old, having taken place during the early steps of diversification of the family. However, the estimated values showed a clear clustering by species with paralogous sequences having similar Gibbs free energy values (Figure 10; Table 4). This shows that stabilities for the RuBisCO complex within the species are consistent, despite different evolutionary histories of the paralogous gene copies.

Corresponding sites and residues of RuBisCO genes and subunits between our analyses and public database are shown in Table 5 to help future studies.

**Table 4. Delta Gibbs free energy of modelled RuBisCO structure**

| | Species names | Name of each gene copy | Differences of Gibbs free energy between maximum likelihood model and null model |
|---|---|---|---|
| Brassicaceae | *Arabidopsis lyrata* | *Aly1* | -205.782 |
| | | *Aly2* | -186.133 |
| | | *Aly3* | -216.206 |
| | *Arabidopsis thaliana* | *Ath1* | -203.385 |
| | | *Ath2* | -234.677 |
| | | *Ath3* | -218.445 |
| | | *Ath4* | -227.182 |
| | *Brassica rapa* | *Bra1* | -186.759 |
| | | *Bra2* | -209.734 |
| | | *Bra3* | -182.142 |
| | | *Bra4* | -213.151 |
| | | *Bra5* | -193.802 |
| | *Capsella rubella* | *Cru1* | -226.842 |
| | | *Cru2* | -204.699 |
| | | *Cru3* | -206.047 |
| | | *Cru4* | -244.759 |
| | *Eutrema solsugineum* | *Esa1* | 23.9451 |
| | | *Esa2* | 20.741 |
| | | *Esa3* | 28.6711 |
| Poaceae | *Brachypodium distachyon* | *Bdi1* | -332.978 |
| | | *Bdi2* | -310.927 |
| | *Setaria italica* | *Sit1* | -231.832 |
| | | *Sit2* | -207.691 |
| | | *Sit3* | -221.541 |
| | | *Sit4/Sit5* | -204.342 |
| | *Zea mays* | *Zma1* | -231.694 |
| | | *Zma2* | -228.502 |

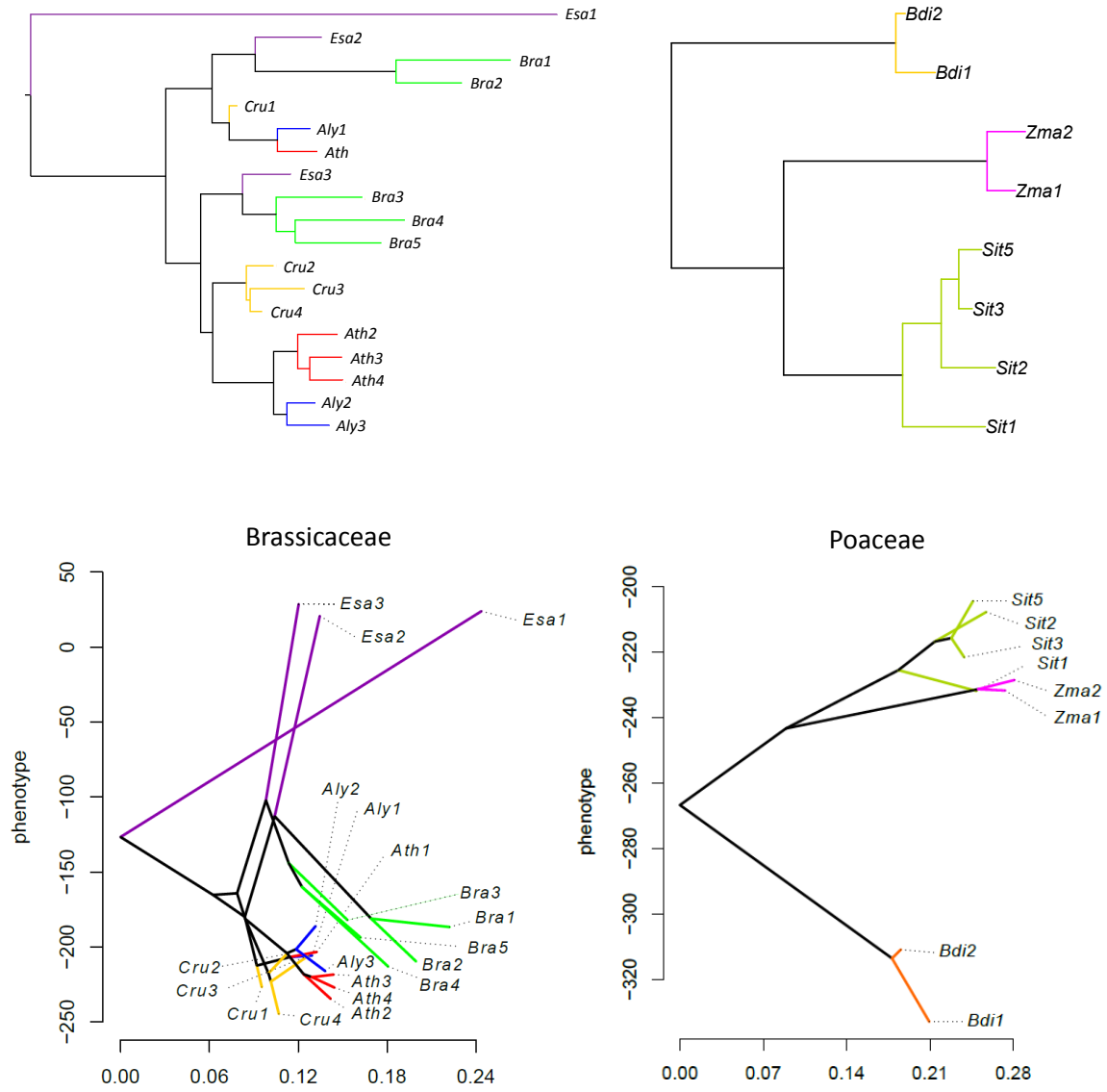



**Figure 10. Stability of modelled RuBisCO structure**

The phylogenetic trees of *rbcS* in Brassicaceae and Poaceae are shown in the first row. RuBisCO protein structures with RBCS encoded by each *rbcS* were modelled by homology modelling of Modeller (Eswar et al., 2008). The modelled structure was repaired by the RepairPDB function of FoldX4 (Schymkowitz et al., 2005). The stability of the whole RuBisCO was estimated using the "Stability" function of FoldX4. Then, the result of protein stability was taken as a trait and phylogenetic relationships were given as input trees. We then drew a phenogram using the "phytools" package (Revell, 2012) in R (Figures in the second row). *Sit5* is shown as representative of *Sit4/Sit5* because of synonymous substitutions.

## Table 5. Correspondance of positions between different databases

| Analysis | Coevolution | | | | | | Positive selection | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Gene name | *rbcL* | | *rbcS* | | | | *rbcS* | | | |
| Nucleotide/ Amino acid | Nucleotide | Amino acid. | Nucleotide | | Amino acid | | Nucleotide | | Amino acid | |
| Profile | Our alignment | 1WDD | Our alignment | NM_105379.3 | P10795 | 1WDD | Our alignment | NM_105379.3 | P10795 | 1WDD |
| Database | - | PDB | - | NCBI | UniProt | PDB | - | NCBI | UniProt | PDB |
| Positions | 463 | 163 | 66 | 241 | 22 | - | 190 | 365 | 64 | 5 |
| | 1195 | 407 | 66 | 241 | 22 | - | 223 | 398 | 75 | 16 |
| | 1321 | 449 | 75 | 250 | 25 | - | 238 | 413 | 80 | 21 |
| | 94 | 40 | 75 | 250 | 25 | - | 241 | 416 | 81 | 22 |
| | 46 | 24 | 87 | 262 | 29 | - | 271 | 446 | 91 | 32 |
| | 142 | 56 | 87 | 262 | 29 | - | 274 | 449 | 92 | 33 |
| | 1195 | 407 | 87 | 262 | 29 | - | 286 | 461 | 96 | 37 |
| | 463 | 163 | 87 | 262 | 29 | - | 331 | 506 | 111 | 52 |
| | 56 | 28 | 87 | 262 | 29 | - | 394 | 569 | 132 | 73 |
| | 814 | 280 | 232 | 407 | 78 | 19 | 415 | 590 | 139 | 80 |
| | 1001 | 343 | 441 | 616 | 147 | 88 | 418 | 593 | 140 | 81 |
| | 662 | 230 | 441 | 616 | 147 | 88 | 433 | 608 | 145 | 86 |
| | 85 | 37 | 441 | 616 | 147 | 88 | 439 | 614 | 147 | 88 |
| | 796 | 274 | 441 | 616 | 147 | 88 | 484 | 659 | 162 | 103 |
| | 1192 | 406 | 484 | 659 | 162 | 103 | 499 | 674 | 167 | 108 |

# Discussion

In this study, we investigated the evolution of the small subunit of the RuBisCO protein in 33 species of angiosperms. We characterized the differences between each *rbcS* gene copy by testing coevolution between *rbcS* and *rbcL* and the influence of each copy on the stability of the enzyme.

We reconstructed the phylogenetic relationships of the *rbcS* gene copies, and this showed a pattern whereby gene copies of the same species were more closely related to each other than those of different species. We did not detect a significant signal of gene conversion but found extensive coevolution between the two RBCS and RBCL subunits. Although the presence of coevolution between these two genes that encode tightly linked proteins was expected, our analyses showed that the coevolution between *rbcS* and *rbcL* did not involve specific *rbcS* gene copies, but represented rather a pervasive process throughout the evolution of these genes. We finally identified several sites that are evolving under positive selection in *rbcS* and showed through homology modelling, that the incorporation of any of the *rbcS* sequence for a given species does not affect significantly the stability of the RuBisCO protein.

### Phylogenetic reconstruction of the *rbcS* gene family

The topology of the *rbcS* gene tree within each angiosperm family mostly follows the topology of the expected species tree. In most species of angiosperms, gene copies of the same species were more closely related than those of different species. This pattern has already been reported within some species of the same genera such as *Solanum* and *Flaveria* (Kapralov et al., 2011; Pichersky & Cashmore, 1986). Our analysis is, however, the first to show that this pattern is not restricted to specific genera and is present across all angiosperms. We also found family-specific duplication events in Brassicaceae and Rosaceae.

In general, the evolution of multigene families is affected by a number of processes that involve either divergent, concerted, or birth-and-death evolution (Nei & Rooney, 2005). Nei & Rooney (2005) defined divergent evolution as a mechanism that gene copies of the common ancestral species are retained for long-term after speciation in descendant species. However, we observed copy number variation between species and also pseudogenized copies; thus, divergent evolution is unlikely to be the main process of *rbcS* evolution.

Gene copies of the same species were more similar than gene copies of different species. Such similarity between gene copies within species is often the result of frequent gene conversions between gene copies during concerted evolution. Sugita and colleagues (Sugita, Manzara, Pichersky, Cashmore, & Gruissem, 1987) have suggested that the high similarity of paralogous *rbcS* copies of *Solanum lycopersicum* is more likely to be explained by gene conversion. We tested for gene conversion using RDP4 (Martin et al., 2015) and CHAP2 (Song et al., 2012). However, we could not detect any significant signal of gene conversion across angiosperms. This result is congruent with the results of Miller (2014) who could not find clear evidence of gene conversion between *rbcS* gene copies of species from Solanaceae. Additionally, we observed that gene copies of the same species are separated by long branches, such as those found in *Linum usitatissimum* or *Mimulus guttatus*. These genes are unlikely to be affected by concerted evolution because gene copies should be less genetically distant by frequent gene conversions or crossing-over.

Finally, we considered the possibility that *rbcS* evolved following a birth-and-death process (Nei & Rooney, 2005). The observed pattern of the *rbcS* tree may have occurred by frequent recent duplications followed by pseudogenization and gene loss.

## Retention rate of duplicates and two lineages of *rbcS*

The pattern of the *rbcS* tree suggests that gene copies that may have originated from the ancient duplication events have been removed (except the event that created *rbcS*-lineage1 and 2, and ones before the divergence of Brassicaceae and Rosaceae), while gene copies that may have originated from recent events have been retained. We found two *rbcS* lineages (*rbcS*-lineage1 and *rbcS*-lineage2) that might have originated from a duplication event before the divergence of monocots and dicots. *RbcS*-lineage1 (shown in Figure 1) includes gene copies that are expressed in photosynthetic organs (Cheng et al., 1998; Yoon et al., 2001). *RbcS*-lineage2 (shown in Figure 5) includes gene copies that are expressed in non-photosynthetic organs such as *OsRbcS1* (Morita, Hatanaka, Misoo, & Fukayama, 2016). Some gene copies of the *rbcS*-lineage2 include stop codons. All the species carry gene copies of *rbcS*-lineage1, but only a few species carry copies of *rbcS*-lineage2. Considering the time passed from the divergence of monocots and dicots, gene copies of *rbcS*-lineage2 may have been kept because they may have a different function from that of *rbcS*-lineage1. The incorporation of *OsRbcS1* to RuBisCO has increased the catalytic turnover rate of RuBisCO (Morita et al., 2014). More investigation is required to understand why gene copies of *rbcS*-lineage2 that may contribute to the improvement of catalytic properties of RuBisCO do not exist in all the species.

## Positive selection and coevolution analyses

Our second goal was to estimate the selective pressure acting on *rbcS* and uncover the coevolution between *rbcS* and *rbcL* encoding the subunits of the RuBisCO protein by estimating the coevolution between pairs of sites from these two genes. We detected positive selection in 15 positions along the *rbcS* sequence (Table 2), which indicates that the evolution of the *rbcS* gene is affected by episodic events of positive selection and that the adaptation of the RuBisCO protein, which has been previously attributed mainly to the evolution of *rbcL* (Christin, et al., 2008a; Kapralov & Filatov, 2007), could also be mediated by changes occurring within the gene encoding for the small subunit. We also detected extensive signals of coevolution between the two subunits, which reinforces our

understanding of the tight interaction between the two subunits. One of the coevolving positions (*rbcL* position 1321; Table 3) is part of a codon that is highly conserved between higher plants and *Chlamydomonas* algae (Marin-Navarro & Moreno, 2006). The substitution of this amino acid from a cysteine to a serine has been shown to drastically increase the degradation of the RuBisCO in *Chlamydomonas* (Marin-Navarro & Moreno, 2006). The corresponding position on *rbcS* (position 75; Table 3) further coevolves with another *rbcL* position (position 94; Table 3) that was also described as important for the degradation of the RuBisCO (Kokubun, Ishida, Makino, & Mae, 2002). Our results could indicate that the position 75 on the small subunit may also be involved in the protection against the degradation of the RuBisCO.

We also detected some sites of *rbcL* (positions 56 and 1,321; Table 3) as being part of a coevolving pair with sites of *rbcS*. These two positions of *rbcL* were reported as positively selected in previous studies (Kapralov et al., 2011; Sen et al., 2011), which could suggest that the *rbcS* substitutions might be reacting to functional changes on the large subunits. Some coevolving positions, in particular positions 463 and 662 of *rbcL* (positions 163 and 230 of reference sequence 1WDD) and positively selected positions 19, 25, and 111 of *rbcS*, are on the interface of RBCS and RBCL (Knight & Andersson 1990). Further, position 662 (position 230 of reference sequence 1WDD) has also been reported to locate where RBCS and RBCL are hydrogen-bonded (Knight & Andersson 1990). Kapralov and Filatov (2007) have suggested that widespread positive selection of *rbcL* may help the plant to adjust to changes of environmental conditions. In our study, we show positive selection acting on the *rbcS* gene and positively selected *rbcS* sites that are coevolving with *rbcL*. Position 484 of *rbcS* is both under positive selection and coevolving with *rbcL*. These results may suggest the substitution of amino acid of RBCL may coordinatedly substitute amino acid of RBCS, and vice versa. Chakrabarti and Panchenko (2010) have suggested that functionally important sites undergo coevolution. Some of the positively selected sites or coevolving sites are on the interface of RBCS and RBCL. We suppose that the evolutionary processes of RBCS and RBCL are profoundly influenced to each other. These reported positively selected

positions of *rbcS* and coevolving positions of *rbcS* with *rbcL* may be important sites for the structure and the function of RBCS and these results may help to elucidate the function of RBCS.

## Protein stability of RuBisCO structure

Another goal was to understand the differences of stability between different gene copies. The composition of the RBCS subunits within the RuBisCO complex in vivo is not known. Structural stability is an important feature in an enzyme, which tends to evolve in a narrow range of stability. RuBisCO is no exception and it was observed that some amino acid substitutions under positive selection can slightly shift the stability during adaptation, in order to improve the catalytic efficiency while keeping the global fold intact (Studer et al., 2014). We were interested to see if the differences in the multiple copies of *rbcS* could significantly impact the stability of the RuBisCO complex.

Our protein stability modelling suggests that gene copies of the same species may have similar functions in spite of their different evolutionary histories. Sasanuma (Sasanuma, 2001) investigated the fate of newly duplicated *rbcS* genes in *Triticum* spp. and found evidence of homogenization and pseudogenized genes. However, no evidence of gaining new functions was detected. Therefore, multiple gene copies may exist for robustness (Plata & Vitkup, 2014; Andreas Wagner, 2005) to maintain the important function of protein.

Like Sasanuma's, our results suggest the robustness of the *rbcS* gene in terms of the dosage effect. RuBisCO is necessary for plants to survive. The robustness of *rbcS* can assist plants adaptation to drastic environmental changes or loss of gene copies. Further investigation is required if we are to understand *rbcS* evolution in more detail. The evolutionary history of *rbcS* is complex to track but we suppose that studying *rbcS* will allow for a deeper understanding of the multigene family.

# Conclusions

Investigating the mechanisms that have shaped the evolution of the RuBisCO complex is important for understanding the function of this key enzyme in photosynthesis. This is usually done by looking at the plastid gene *rbcL*, but this approach only provides half of the picture and it is important to consider the evolution of the smaller subunit encoded by the nuclear gene family *rbcS*. Although *rbcS* has a more complex evolutionary history than *rbcL*, involving the appearance of multiple paralogous gene copies, there are strong connections between the two subunits, as detected in the coevolution analysis of *rbcS* and *rbcL*. Some coevolving or positively selected positions are on the interface of RBCS and RBCL. A striking example is the position 484 of *rbcS,* which is both under positive selection and coevolving with *rbcL*. These results suggest substantial interactions between the subunits. However, the coevolution is not occurring between specific gene copies of *rbcS* and *rbcL*. Further, the differences of evolutionary history of each of the gene copies do not lead to differences in the stability of the RuBisCO. We thus propose: i) that *rbcS* gene copies are created under neutral evolutionary processes, or ii) that different copies are kept by selective pressure that allows plants to cope with different environmental conditions or to express differently in each organ. We need to further investigate the mechanism and the rate of gain and loss of *rbcS*. Transcriptome data of *rbcS* on different organs and different conditions (temperature, aridity) may help to understand if these copies are playing a role in maintaining stoichiometry.

# Chapter 2. Evolution of the *rbcS* gene and adaptive evolution of photosynthesis in Poaceae

## Introduction

Adaptation to the changing environment is crucial for species to survive. After the depletion of atmospheric $CO_2$ in the Oligocene, some species have evolved to have a mechanism to concentrate $CO_2$ (CCM) by cellular-structural or temporal separation (Christin et al., 2008b; Edwards et al., 2010; Sage et al., 2011; Vicentini et al., 2008). The C4 plants are such a group of species that have diverged from the classical C3 plants by modifying the cellular structure and biochemical cascade (Sage, 2004). The C4 plants have evolved more than 60 times independently in multiple lineages across angiosperms (Edwards et al., 2010; Sage et al., 2011; Vicentini et al., 2008).

In the C4 plants, atmospheric $CO_2$ is fixed in mesophyll cells, from which it is transported to bundle-sheath cells where the cycle to fix $CO_2$ to sugar (i.e. the Calvin-Benson cycle) is located (Hatch & Slack, 1968; Kanai & Edwards, 1999). The Calvin-Benson cycle relocated from mesophyll cells to bundle-sheath cells during the transition from C3 to C4 type. RuBisCO has the affinity to both $O_2$ and $CO_2$ as substrates (Rawsthorne, 1992) and it causes loss of energy and $CO_2$ especially in $CO_2$-depleted conditions (Kubien et al., 2008; Peterhansel et al., 2010). The $CO_2$-concentrating mechanism of C4 plants enables Ribulose-1,5- bisphosphate carboxylase/oxygenase (RuBisCO), the first enzyme of the Calvin-Benson cycle that fixes $CO_2$ to sugar, to be surrounded by highly concentrated $CO_2$. As a result, the catalytic efficiency of RuBisCO is better in C4 plants than in C3 plants (Badger & Andrews, 1987; Sage & Coleman, 2001; von Caemmerer & Quick, 2000). Therefore, RuBisCO has been considered as the key enzyme of adaptive evolution of photosynthesis.

Evidence for the adaptive evolution of RuBisCO has already been shown in the positive selection of the plastid *rbcL* gene encoding large subunits of RuBisCO (RBCL) in independent C4 lineages (Christin, et al., 2008a; Kapralov & Filatov, 2007). RBCL consists RuBisCO with small subunits (RBCS) encoded by nuclear *rbcS* genes. Because catalytic sites are part of RBCL (Andersson, 2008), RBCL has attracted more scholarly attention than RBCS; however, RBCS has been reported to be involved in the catalytic efficiency, $CO_2$ specificity, assembly, stability, and activity of RuBisCO (Andrews & Ballment, 1983; Bracher et al., 2011; Furbank et al., 2000; Genkov & Spreitzer, 2009; Genkov et al., 2010; Spreitzer, 2003). Thus, RBCS also seems to play important roles in the evolution of RuBisCO. In particular, catalytic efficiency and $CO_2$ specificity are the key differences between the RuBisCO of C3 and C4 plants. It is therefore reasonable to suppose that RBCS may have been involved in the shift of the photosynthetic types.

To understand the adaptive evolution of RBCS, studying the evolution of the encoding *rbcS* multigene family is necessary. However, the evolution and actual role of *rbcS* genes have been studied very little. The *rbcS* gene has a different number of gene copies per species. High similarities between gene copies of the same species in comparison with those of different species have been reported within genera (Kapralov et al., 2011). Kapralov and his colleagues (2011) have suggested that two distant lineages of *rbcS* are distinguished by the lengths of introns that exist in genus of *Flaveria*. *Flaveria* is known to include both C3 and C4 type of plants within the same genus and Kapralov and his colleagues have detected a weak signal of positive selection for *rbcS* of C4 lineages. However, the signal was almost 20 times weaker than that of *rbcL* and it was not significant. The test was performed using 15 species of the same genus, so it was not comprehensive enough to understand the pattern of positive selection across genera.

Therefore, I aimed to test the involvement of RBCS in the adaptive evolution of photosynthesis by testing positive selection on the *rbcS* genes across genera. I hypothesized that selective pressures acting on *rbcS* genes were shifted by the evolution of

C4 photosynthesis. To test the hypothesis, I sequenced *rbcS* to obtain the larger sampling of C3 and C4 grasses. First, I used 454 sequencing of the PCR product to isolate the *rbcS* gene from species selected considering the taxonomic and photosynthetic diversity of Poaceae. Then, I developed a new pipeline to assemble the sequenced reads into gene models that were representative of the main copies existing in each genome. This dataset was used to build a detailed phylogenetic tree of *rbcS* genes, which enabled inferences about the gene duplication events that might have occurred before the divergence of the included species. Finally, the developed phylogenetic tree allowed me to specifically test for the occurrence of positive selection on C4-specific branches (hypothesis of C4-specific positive selection) and across all branches. It should be noted that some gene copies may have been missed during PCR or/and 454 sequencing. However, my method could detect the main gene copies per species, so it could be used to infer about older duplication events and to test positive selection on the phylogenetic tree.

# Materials and Methods

## Selection of samples

I selected 60 species from Poaceae representing each subfamily and photosynthetic type (Table 1; Grey labels indicate the species that were not used to build the phylogenetic tree, see the section of Sorting and clustering of the 454 reads for reasons). Of the selected species, 13 belongs to the BEP clade, and 47 belongs to the PACMAD clade. The number of species per type of photosynthesis was: 32 from C3, two from C3–C4 intermediate, and 26 from C4.

**Table 1. Selection of samples**

| ID of plants | Name of Subfamily | Name of Species | | Type of photosynthesis |
|---|---|---|---|---|
| 1 | Ehrhatoideae | *Leersia* | *hexandra* | C3 |
| 2 | | *Humbertochloa* | *bambusiuscula* | C3 |
| 3 | Bambusoideae | *Nastus* | *elongates* | C3 |
| 4 | | *Arundinaria* | *marojejyensis* | C3 |
| 5 | | *Pariana* | *modesta* | C3 |
| 6 | | *Olyra* | *latifolia* | C3 |
| 7 | | *Pariana* | *radiciflora* | C3 |
| 8 | Pooideae | *Brachypodium* | *madagascariense* | C3 |
| 9 | | *Agrostis* | *elliottii* | C3 |
| 10 | | *Poa* | *cenisia* | C3 |
| 11 | | *Alopecurus* | *alpinus* | C3 |
| 12 | | *Festuca* | *paniculata* | C3 |
| 13 | | *Helictotrichon* | *sempervirens* | C3 |
| 14 | Early diverginc grass lineages | *Leptophis* | *cochleata* | C3 |
| 15 | Micrairoideae | *Isachne* | *mauritiana* | C3 |
| 16 | | *Coelachne* | *africana* | C3 |
| 17 | Arundinoideae | *Phragmites* | *mauritianus* | C3 |
| 18 | | *Molinia* | *caerulea* | C3 |
| 19 | Danthonoideae | *Scutachne* | *hitchcock* | C3 |
| 20 | | *Merxmuellera* | *tsaratananensis* | C3 |
| 21 | Aristidoideae | *Aristida* | *rhiniochloa* | C4 |
| 22 | | *Aristida* | *adscensionis* | C4 |
| 59 | | *Stipagrostis* | *sp.* | C4 |
| 23 | Chloridoideae | *Eragrostis* | *hildebrandtii* | C4 |
| 24 | | *Eragrostis* | *capensis* | C4 |
| 25 | | *Eragrostis* | *pectinacea* | C4 |
| 26 | | *Sporobolus* | *virginicus* | C4 |
| 27 | | *Sporobolus* | *pyramidalis* | C4 |
| 28 | | *Perotis* | *patens* | C4 |
| 29 | | *Ctenium* | *concinnum* | C4 |
| 30 | | *Craspedorhachis* | *africana* | C4 |
| 31 | | *Neyraudia* | *arundinacea* | C4 |
| 32 | Panicoideae | *Arundinella* | *nepalensis* | C4 |
| 33 | | *Elionurus* | *tristis* | C4 |
| 34 | | *Chrysopogon* | *serrulatus* | C4 |
| 35 | | *Hemarthria* | *natans* | C4 |
| 36 | | *Streptostachys* | *asperifolia* | C3 |
| 37 | | *Ichnanthus* | *pallens* | C3 |
| 38 | | *Axonopus* | *ramosus* | C4 |
| 39 | | *Homolepis* | *aturensis* | C3-C4 |
| 40 | | *Centotheca* | *lappacea* | C3 |
| 41 | | *Steinchisma* | *laxa* | C3-C4 |
| 43 | | *Tristachya* | *betsileensis* | C4 |
| 46 | | *Centotheca* | *lappacea* | C3 |
| 47 | | *Lecomtella* | *madagascariensis* | C3 |
| 48 | | *Yvesia* | *madagascariensis* | C4 |
| 49 | | *Sacciolepis* | *indica* | C3 |
| 50 | | *Cytococcum* | *deltoideus* | C3 |
| 51 | | *Poecilostachys* | *bakeri* | C3 |
| 52 | | *Alloteropsis* | *cimicina* | C4 |

| | | | | |
|---|---|---|---|---|
| 53 | | *Echinochloa* | *frumentacea* | C4 |
| 54 | | *Digitaria* | *radicosa* | C4 |
| 55 | | *Panicum* | *capillare* | C4 |
| 56 | | *Panicum* | *hymeniochilum* | C3 |
| 57 | | *Panicum* | *pleianthum* | C3 |
| 58 | | *Panicum* | *dichotomiflorum* | C3 |
| 60 | | *Cenchrus* | *spinifex* | C4 |
| 42 | Outlying Panicoideae | *Loudetia* | *simplex* | C4 |
| 44 | | *Trichopteryx* | *dregeana* | C4 |
| 45 | | *Magastachya* | *mucronata* | C3 |

## Design of primers and protocol

According to the available sequences of *rbcS* of species from Poaceae on the Phytozome v.12 database, I designed primers for *rbcS* to specifically target and amplify only the *rbcS* gene, but simultaneously to be universal enough to amplify the *rbcS* gene in all the 60 species. I designed two sets of primers as follows: Forward primers: 5'-TATGGCNCCCACCGTGATG-3' and 5'-TCCRTTCCAGGGSCTCAAGTCC-3'. Reverse primers: 5'-CGATGAAGATGATGCACTGC-3' and 5'-ACGGTGGCTTGTAGGCGATG-3'. I refer to these primers as A, B, C, and D, respectively. Referring to Phytozome Version 12, the sequences of the *rbcS* had two exons per species in Poaceae. Each primer was designed as is shown in Figure 1. The same region including intron was read twice by two sets of primers in order to increase the depth of sequencing coverage.

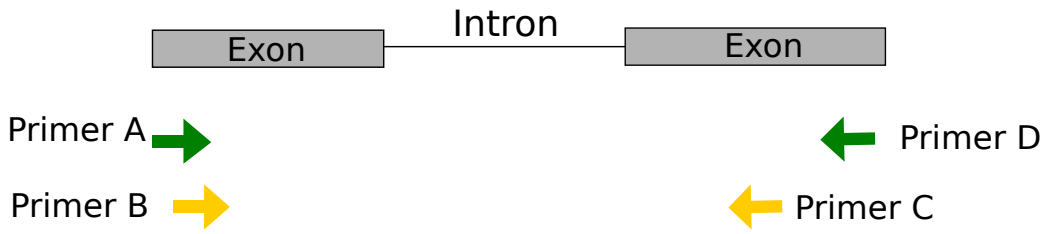



**Figure 1. Primer design for the *rbcS* gene in Poaceae**
Primers to amplify the *rbcS* gene were designed. According to previous studies, *rbcS* of Poaceae have two exons. Primer A was designed to amplify from the beginning of exon1. Primer B starts from 45 base pair inside of exon1. Primer D was designed to reach the end of exon2. Primer C was designed few base pairs inside from the end of exon2.

## DNA extraction

Thanks to the previous studies of *rbcL* and *ppc* evolution in Poaceae by Besnard et al. (2009) and Christin et al. (2011) the extracted DNA of most of the candidate species was available in our lab.

To make a new protocol for the *rbcS* primers, I collected fresh leaf tissues of some species of Poaceae at the campus of the University of Lausanne. Leaf tissues were homogenized by shaking with beads for one to two minutes in a homogenizer. The DNeasy Plant Mini Kit (Quiagen, USA) was used for extraction following the manufacturer's protocol. The quality and concentration of DNA were measured by NanoDrop and the integrity of the DNA was verified on 1.5% agarose gel.

## Preparation of aliquots for 454 sequencing

The 454 pyrosequencing is a technique using emulsion-based clonal amplification. Two steps of amplification are required for preparation of 454 sequencing: 1) amplification by standard PCR with standard primers, and 2) another amplification by the special primers called "fusion primers" including barcodes, so called Multiplex Identifiers (MID).

### 1) The 1$^{st}$ PCR conditions and purification

PCR was carried out using 10ng of DNA, 10µl of AccuPrime Buffer, 1µl of dNTPs, 1µl of each primer, 2.5µl of DMSO (dimethyl sulfoxide), 1µl of $MgCl_2$, and 0.2µl of Taq polymerase (AccuPrime DNA Polymerase, Invitrogen), and filled up to a final volume of 50µl with $H_2O$. The thermal cycler programme entailed one initial cycle of 94°C for 2 min, followed by 35 cycles of 94°C for 30 sec, 51°C for 30 sec, 72°C for 2 min, then extension at 72°C for 10 min. The concentrations of PCR products were measured using NanoDrop. The quality of the PCR products was verified by gel electrophoresis on 1.5% of agarose gel. Then, the PCR products were purified using the QIAquick PCR Purification Kit (Qiagen, USA) following the manufacturer's protocol.

## 2) Selection of MID and design of 454 plate

The fusion primers mainly consisted of two regions. One region was the same sequences as standard primers and another region was MID. MID is like a barcode and useful for distinguishing reads of each species in the same sequencing group (run). Lists of MID sequences which were usable for 454 sequencing were provided by Microsynth AG (Switzerland). The combinations of primers and MIDs were selected to prevent primer dimers or amplification of non-targeting regions. I selected 15 different MIDs.

**Table2. Combination of regions of primers and the design of sequencing plate for 454 sequencing**

| ID of MID (Barcodes) | Forward primer | Reverse primer | ID of plants | | | |
|---|---|---|---|---|---|---|
| | | | Lane 1 | Lane 2 | Lane 3 | Lane 4 |
| MID1 | A | D | 1 | 16 | 31 | 46 |
| MID2 | A | D | 2 | 17 | 32 | 47 |
| MID3 | A | D | 3 | 18 | 33 | 48 |
| MID4 | A | D | 4 | 19 | 34 | 49 |
| MID5 | A | D | 5 | 20 | 35 | 50 |
| MID6 | A | D | 6 | 21 | 36 | 51 |
| MID7 | A | D | 7 | 22 | 37 | 52 |
| MID8 | A | D | 8 | 23 | 38 | 53 |
| MID9 | A | D | 9 | 24 | 39 | 54 |
| MID10 | A | D | 10 | 25 | 40 | 55 |
| MID11 | A | D | 11 | 26 | 41 | 56 |
| MID12 | A | D | 12 | 27 | 42 | 57 |
| MID13 | A | D | 13 | 28 | 43 | 58 |
| MID14 | A | D | 14 | 29 | 44 | 59 |
| MID15 | A | D | 15 | 30 | 45 | 60 |
| MID1 | B | C | 1 | 16 | 31 | 46 |
| MID2 | B | C | 2 | 17 | 32 | 47 |
| MID3 | B | C | 3 | 18 | 33 | 48 |
| MID4 | B | C | 4 | 19 | 34 | 49 |
| MID5 | B | C | 5 | 20 | 35 | 50 |
| MID6 | B | C | 6 | 21 | 36 | 51 |
| MID7 | B | C | 7 | 22 | 37 | 52 |
| MID8 | B | C | 8 | 23 | 38 | 53 |
| MID9 | B | C | 9 | 24 | 39 | 54 |
| MID10 | B | C | 10 | 25 | 40 | 55 |
| MID11 | B | C | 11 | 26 | 41 | 56 |
| MID12 | B | C | 12 | 27 | 42 | 57 |
| MID13 | B | C | 13 | 28 | 43 | 58 |
| MID14 | B | C | 14 | 29 | 44 | 59 |
| MID15 | B | C | 15 | 30 | 45 | 60 |

The design of the 454 plate is shown in Table 2. The 60 species were divided into four groups of 15 species. I aimed to amplify each species using two different sets of primers, thus there were eight groups of 15 samples in total. In 454 technology, samples in one lane are sequenced at once, then sequenced reads are sorted to each species according to MIDs after sequencing. Thirty samples (i.e. two groups) were sequenced in one lane, thus four lanes were used in total. The combination of MID (15 different MIDs) and primers (two different sets of primers) enabled each fusion primer to be unique in each lane of the sequencing plate. I designed the experiment as explained above to reduce the cost of sequencing because using many fusion primers and lanes is costly.

### 3) The 2nd PCR conditions and purification

The conditions of the second amplification were as follows. An initial cycle at 94℃ was run for 2 min, followed by five cycles at 94°C for 30 sec, 51°C for 30 sec, and 72°C for 2 min. This was followed by a final extension step at 72°C for 10 min. The PCR products of the second amplification were purified using a magnetic beads purification kit, Agencourt AMPure XP (Beckman Coulter, USA), following the protocol of the manufacturer.

### 4) Pooling and purification

10ng of purified PCR products per sample was pooled into one aliquot, which corresponded to one lane of the 454 plate. Four aliquots were prepared in total, one per lane. Each aliquot was purified by the method of gel cutting purification of the QIAquick Gel Extraction Kit (Quiagen, USA).

### 5) Qualification and quantification

The final quality and quantity of aliquots were measured using Qubit and Bioanalyzer. The final concentrations of the four pooling aliquots were 9,1ng/μl, 14.4ng/μl, 15.1ng/μl, and 12.3ng/μl. The four aliquots were sent to Microsynth AG (Switzerland) to run the 454 sequencing.

## Sorting and clustering of the 454 reads

The obtained 454 reads of each lane were sorted into species according to the MIDs. The number of reads per species was around 3,000 and the average length of reads was around 500 base pairs.

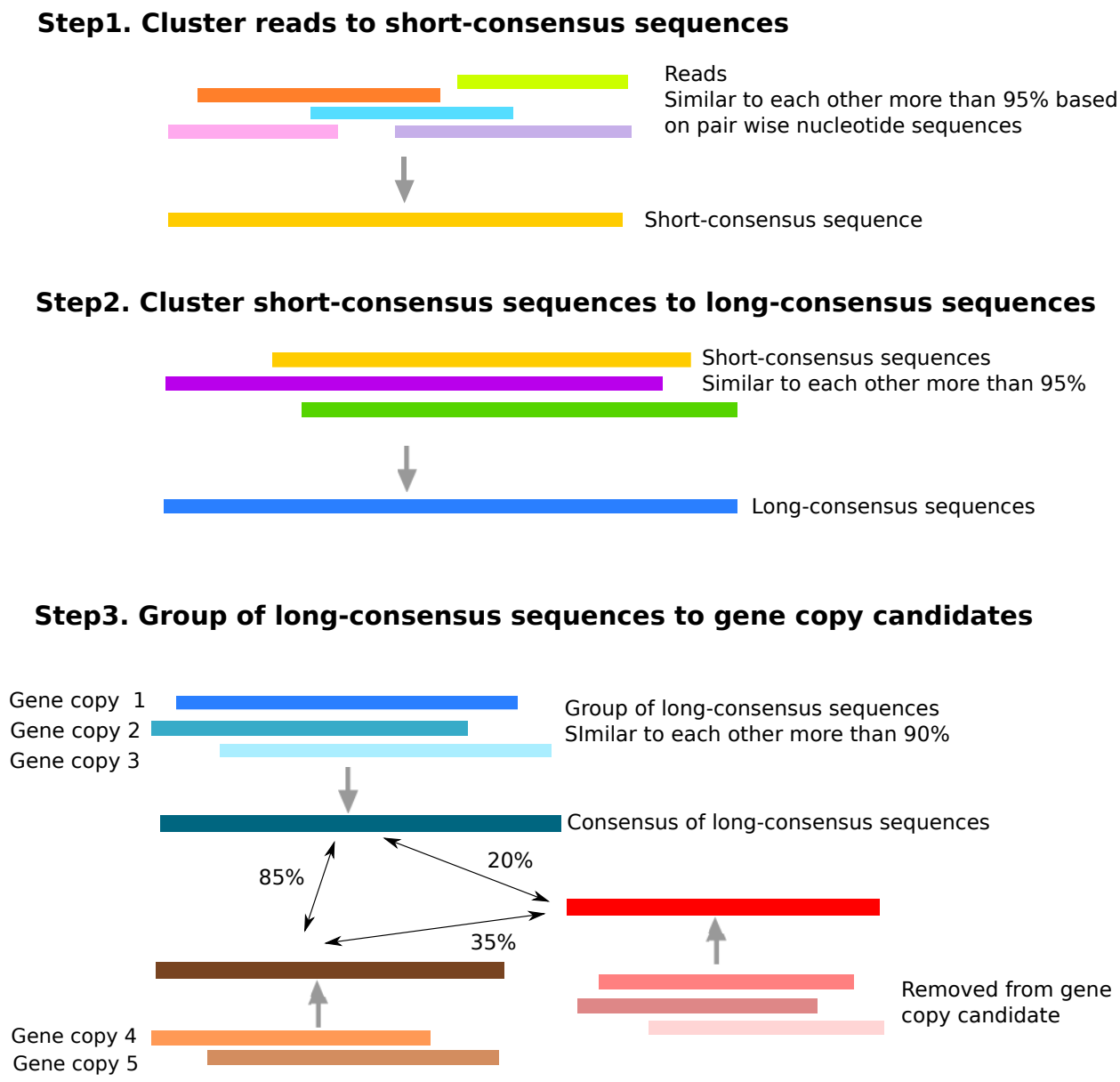



**Figure 2. Method of clustering reads of 454 sequencing**
The reads of 454 sequencing that are sorted per each species were aligned by MAFFT (Katoh & Standley, 2013) with default settings in Geneious software. Firstly, the reads that shared more than 95% similarity were merged, then the consensus sequences of them were extracted as short-consensus sequences. Secondly, the short-consensus sequences were aligned by MAFFT with default settings and extracted as long-consensus sequences. Thirdly, the long-consensus sequences were gathered into a group of potential gene copies. Then, if the consensus sequences of the group had similarities less than 70% with the consensus of other groups, I removed the group with low similarity.

I established a new pipeline to assemble reads (Figure 2). Firstly, I aligned reads of the same species using MAFFT (Katoh & Standley, 2013) with default settings in Geneious 10.2.3 (Kearse et al., 2012). After alignment, the pair-wise similarities between each pair of reads were automatically calculated based on nucleotide sequences. I extracted them as a pair-wise distance matrix. Then, I merged reads which were similar to each other as follows. Raw reads were merged using a 95% similarity threshold. This threshold was determined because the single-base error rate of 454 sequencing is estimated to be around 4.5% (Luo, Tsementzi, Kyrpides, Read, & Konstantinidis, 2012). If pair-wise reads have more than 95% similarity with each other, I assumed that 5% difference may have been caused by sequencing errors and these two reads belong to the same region of the same *rbcS* gene copy.

Consensus sequences of merged reads were then extracted from each alignment of pair-wise reads (hereafter referred to as short-consensus sequences). Repeating exactly the same method as above, all the short-consensus sequences were aligned using MAFFT (Katoh & Standley, 2013) with default settings and the pair-wise matrix was extracted. When a pair of short-consensus sequences had more than 95% similarity, I merged them together. Then, from the two short-consensus sequences I extracted the consensus alignment of them as the long-consensus sequences. The length of long-consensus sequences became almost the same length as of the *rbcS* gene in publicly available databases (i.e. around 600 to 1,000 base pairs).

The long-consensus sequences were then merged using a 90% similarity threshold. I used this threshold because similarities of *rbcS* gene copies within the same species available in the Phytozome Version 12 database were slightly higher than 90% (mean similarity: 91.8%; standard deviation: 4.7%). Thus, I assumed that long-consensus sequences which have more than 90% similarity can be considered as a group of potential gene copies, which resulted in each species having several groups of potential gene copies.

Finally, I aligned the potential gene copies of the same species using MAFFT (Katoh & Standley, 2013), and excluded potential gene copies which had less than 70% similarity with sequences from the same species because this represented the maximum level of similarity detected between the two lineages of *rbcS* within angiosperms from the Phytozome database (see Chapter 1).

## Alignment and reconstruction of the phylogenetic tree

Coding regions of *rbcS* gene copies of 10 species of Poaceae were downloaded from Phytozome Version 12. These sequences were aligned with the potential gene copies which passed the threshold of clustering using MAFFT (Katoh & Standley, 2013) with default settings. Intron regions of potential gene copies were identified and removed because of the large divergence in these regions. Nucleotide sequences were translated to amino acid using Geneious 10.2.3 (Kearse et al., 2012). The nucleotide and translated amino acid alignments were exported and aligned by codon using PAL2NAL (Suyama, Torrents, & Bork, 2006). The GTR + G model of evolution was used to estimate the phylogenetic tree of the *rbcS* sequences using PhyML3.0 (Guindon & Gascuel, 2003). The BEST option for the tree swapping was used during the tree reconstruction and the branch support was estimated using 1,000 bootstrap replicates (Collapsed tree is shown in Figure 3; Detailed tree is shown in Figure 4).

## Positive selection

Evidence for positive selection of C4 branches was tested using the branch-site model as implemented in Godon (Davydov, Robinson-Rechavi, & Salamin, 2017). The branch-site model implemented in Godon is the same as the CodeML of PAML (Yang, 2007); however, the Godon implementation has specific algorithms to ensure the convergence of the optimized parameters of the alternative model and is computationally faster than CodeML.

Positive selection was tested for two hypotheses: first, by taking into account all the C4 branches as foreground and all other branches as C3 (hypothesis of C4-specific positive selection); and second, by taking into account each branch regardless of photosynthetic type

as foreground and all other branches as background (hypothesis of constant positive selection). Secondly, tests were conducted to detect positive selection acting on single branches. To do so, each branch was successively set as the foreground branch, with all others as background branches. The C3–C4 intermediate species were considered as C4 species because they have the initial characteristics of the CCM. Firstly, I estimated the branch lengths by using the M0 model of codon evolution (option --m0-tree in Godon). Secondly, I tested for positive selection by branch-site model while keeping the branch lengths fixed to their M0 values (options of -no-branch-length, --procs=1, --seed=1, --json=output in Godon; Davydov et al., 2017). I determined the threshold of significant signal of positive selection using q-values. The q-values were estimated to control for false discovery rate by using the R package "qvalue" (Bass, Swcf, Dabney, & Robinson, 2015). The branches with qvalues < 0.1 were considered as showing evidence for positive selection. Branches under a significant signal of positive selection are shown in Figure 5. I used the BEB approach to estimate the probability (> 0.95) of sites to be under selection on the branches tested by the branch-site model (Table 3).

## Homology of neighbouring genes of *rbcS*

To further examine high similarities of *rbcS* gene copies, I initially aimed to discover orthologous relationships of *rbcS* gene copies among species. The orthology database (OMA) (Altenhoff et al., 2018) was used to estimate homologous relationships between genes among species based on sequence similarities. I tested the orthology of the *rbcS* gene family within *rbcS*-lineage1 that I suggested in Chapter1.

The OMA analysis was combined with the comparison of similarities based on nucleotide sequences of neighbouring genes of *rbcS* to help the identification of orthologous regions among species within *rbcS*-lineage1. I extracted nucleotide sequences of neighbouring genes of *rbcS*. I identified the gene that is at the upstream position next to *rbcS* as the "A-gene" and the gene that is at the downstream position next to *rbcS* as the "C-gene". The A-genes and C-genes of each of the *rbcS* gene copies were identified by examining available

genomes in Phytozome Version 12 using Jbrowse (Buels et al., 2016). Nucleotide sequences of A-genes and C-genes were downloaded from the Phytozome database. I aligned all the A-genes of all the *rbcS* gene copies of angiosperms using MAFFT (Katoh & Standley, 2013) with default setting in Geneious 10.2.3 (Kearse et al., 2012). I extracted automatically calculated pair-wise distance matrices of these sequences to obtain the information about the similarities between neighbouring genes of the *rbcS* gene. The process was repeated for C-genes.

# Results

## The phylogenetic tree of *rbcS* with newly sequenced species in Poaceae

After the clustering of reads and sorting of candidate gene copies, the alignment contained 576 base pairs for 111 candidate gene copies from 33 species. It includes 35 gene copies for the 10 species downloaded from Phytozome version 12.

The *rbcS* copies of the species belonging to the same subfamilies are clustered together (Figure 3). The two main groups of Poaceae, BEP and PACMAD, were grouped with 37.1% of branch support. The branches leading to each subfamily of BEP had high branch supports (>90%), while the ones leading to each subfamily of PACMAD had relatively low supports (<60%; except Aristidoideae with 96.1% and one of the Panicoideae lineage with 100%). In the BEP clade, Bambusoideae was placed outside of Ehrhatoideae and Pooideae. In the PACMAD clade, the tree diverged following the order from Arundinoideae to Microirodeae, Chloridoideae, Aristidoideae.

Gene copies belonging to Panicoideae were separated into two clades (Figure 3). Gene copies of the species of Panicoideae were mostly clustered with ones of the same tribes: Andropogoneae, Paspaleae, and Paniceae except for the few following exceptions. *Chrysopogon serrulatus* of Andoropogoneae clustered with species of Paspaleae. *Tristachya*

*betsileensis* (Arudinelleae) and *Lecomtella madagascariensis* (Lecomtelleae) were placed within the Paniceae clade (Figure 4).

Higher similarities of gene copies within the species than gene copies of different species were observed in the tree of Poaceae (Figure 4). The pattern was commonly observed all over the tree. However, as exceptions, duplications within genera were observed (e.g. *Setaria* and *Brachypodium* showed 90–100% and 70–100% of branch support, respectively). Apart from these, gene copies of *Sporobolus pyramidalis* and *Neyraudia arundinacea* did not cluster within the same species, but each copy was similar to different copies of *Eragrostis* species.

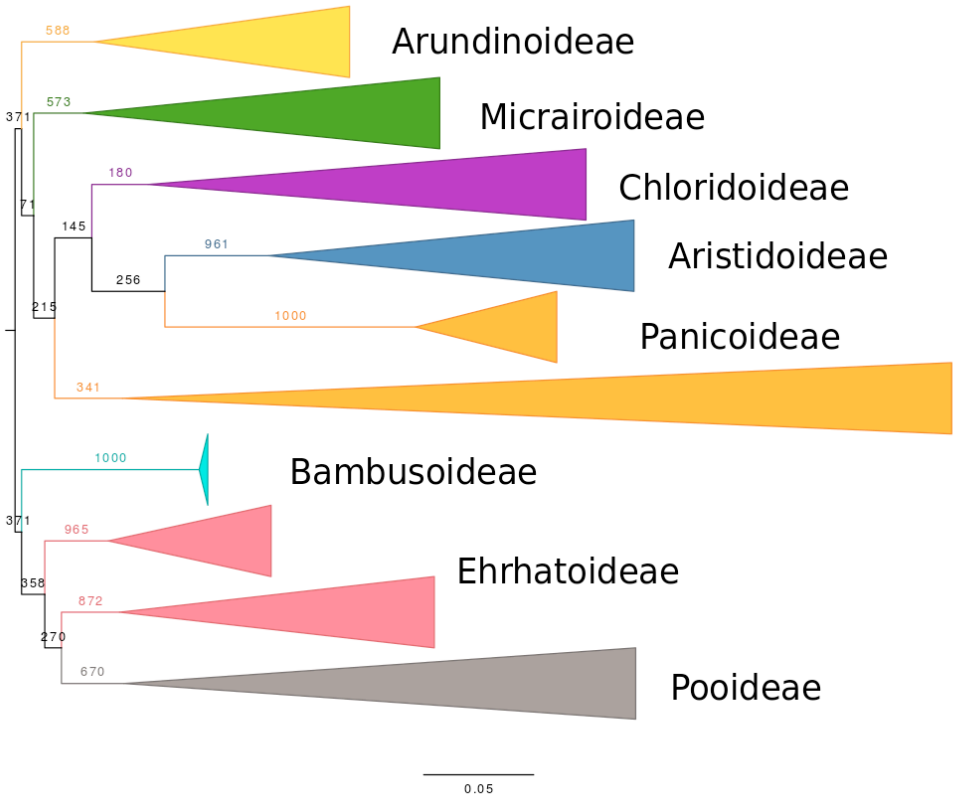



**Figure 3. Collapsed maximum likelihood tree of *rbcS* in Poaceae**
The maximum likelihood tree was reconstructed based on the alignment of *rbcS* gene copy candidates of 454 sequencing using PhyML3.0 with model GTR+G and 1,000 bootstrap replicates. The bootstrap values are shown next to the branches. The clades of subfamilies are collapsed. The colours of the clades represent each subfamily.


**Figure 4. The maximum likelihood tree of *rbcS* in Poaceae**

The maximum likelihood tree was reconstructed based on the alignment of rbcS gene copy candidates of 454 sequencing using PhyML3.0 with model GTR+ G and 1,000 bootstrap replicates. The bootstrap values are shown next to the branches. The colours of tips represent the different types of photosynthesis: green, orange, red for C3, C3-C4 intermediate, and C4, respectively.

## Positive selection

The signal of positive selection on all C4 branches was not significant, so the hypothesis of the C4-specific positive selection was denied. The signal of positive selection was observed on 45 branches that were spread all over the tree regardless of photosynthetic type (Figure 5). These results suggest that positive selection acting on *rbcS* is caused by other reasons besides the transition of photosynthetic types. The sites detected as evolving under positive selection along multiple branches tested with more than 0.95 posterior probabilities are shown in Table 3.

**Table 3. Positions under positive selection in multiple branches**

| Positions under positive selection<br><br>(Based on nucleotide alignment used for this analysis) | Number of branches where positive selection of each position was detected |
|---|---|
| 7 | 2 |
| 10 | 2 |
| 16 | 2 |
| 18 | 4 |
| 33 | 2 |
| 36 | 2 |
| 38 | 2 |
| 43 | 2 |
| 58 | 2 |
| 60 | 2 |
| 74 | 3 |
| 75 | 5 |
| 77 | 3 |
| 99 | 2 |
| 107 | 2 |
| 132 | 5 |
| 133 | 7 |
| 134 | 4 |
| 135 | 2 |
| 137 | 5 |
| 139 | 3 |
| 140 | 2 |
| 144 | 2 |
| 148 | 2 |
| 150 | 4 |
| 158 | 2 |
| 160 | 3 |
| 162 | 3 |

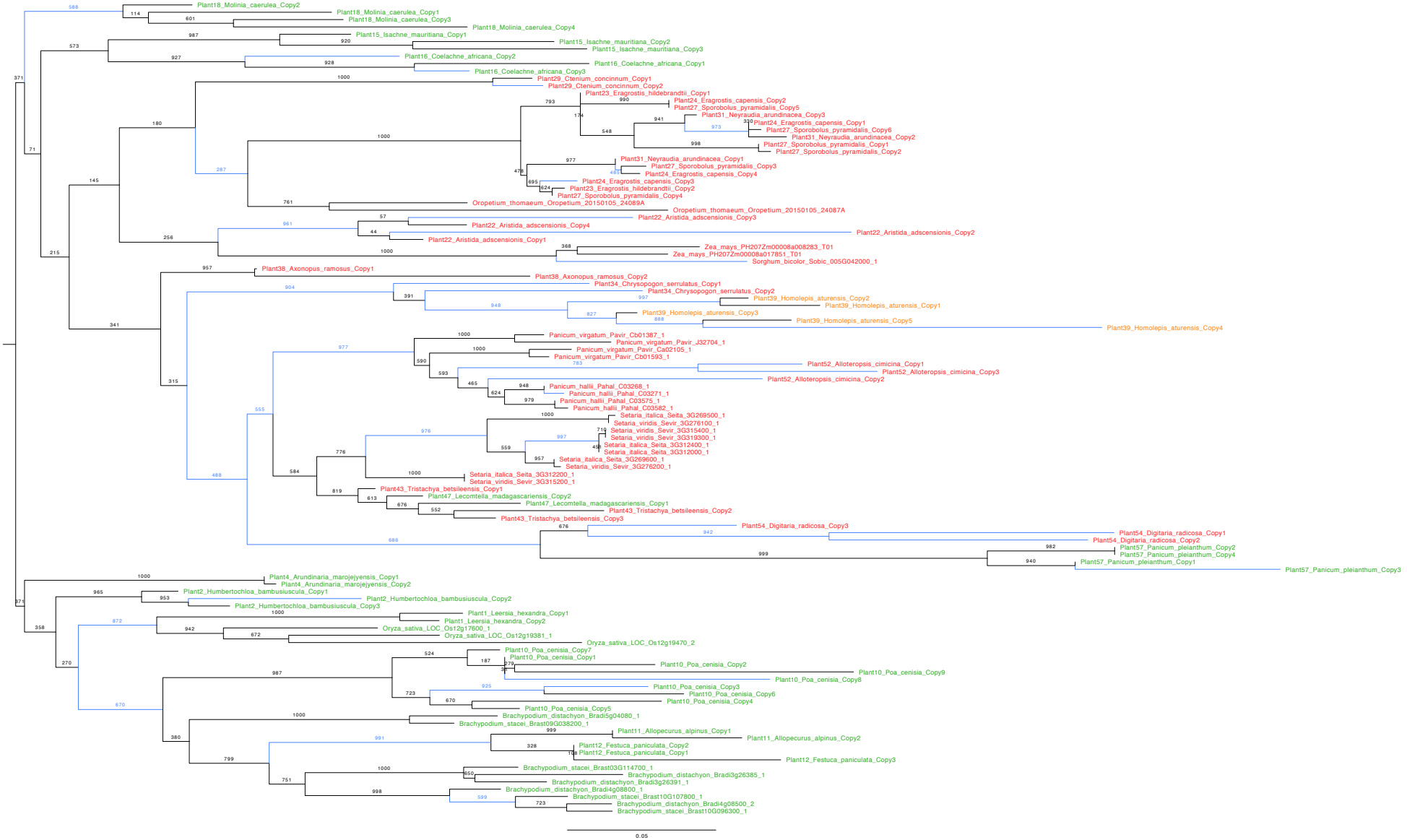



**Figure 5. Maximum likelihood tree of *rbcS* in Poaceae and branches under positive selection**

Positive selection on C4 branches was tested by Godon (Davydov et al., 2017), which implements the branch-site model of Zhang et al. (2015). Each branch was taken as foreground and all other branches were taken as background. Firstly, branch lengths were estimated using the codon substitution model. The estimated branch length was used to run the null model (H0 model) and the alternative positive selection model (H1 model). The options –-m0-tree and –-no-branch-length were used. The branches under positive selection were coloured in blue. The colours of tips represent the different types of photosynthesis: green, orange, red for C3, C3-C4 intermediate, and C4, respectively.

## Orthologous relationships of the *rbcS* gene and its neighbouring genes

I tested orthologous relationships of *rbcS* gene copies using the approach developed in OMA (Altenhoff et al., 2018). However, the high similarities between gene copies within species did not allow for the correct detection of orthologous relationships between the gene copies of *rbcS*-lineage1. Therefore, the similarities of the neighbouring genes of *rbcS* were examined because highly similar fragments between species can suggest a history of duplication events that have occurred before and after speciation. High similarities of neighbouring genes were observed between gene copies of species from the same genera. For example, tandem duplicates of *rbcS* were observed both in *S. italica* (Seita3G.269500.1 and Seita3G.269600.1) and *S. viridis* (Sevir3G.276100.1 and Sevir3G276200.1). A-genes (Seita3G.269400.1 and Sevir3G276000.1; orange in Figure 6) and C-genes (Seita3G.269700.1 and Sevir3G276300.1; dark green in Figure 6) of tandem copies were found to be 98% and 100% similar to each other, respectively. The pair-wise similarities of neighbouring genes based on nucleotide sequences are shown in Table 4. The colours indicated in Table 4 correspond to the colours of the genes in Figure 6. To investigate the structure of the *rbcS* gene and its neighbouring genes on each chromosome, I drew a schematic representation showing the locations of genes on each chromosome (Figure 7). The sequences of neighbouring genes, and the positions of the *rbcS* gene copies and their neighbouring genes on chromosomes were similar between species of the same genera (e.g. between *S. italica* and *S.viridis*, between *B.distachyon* and *B.stacei*). Gene copies of *rbcS*-lineage2 of *O. sativa* and *Panicum hallii* had high similarities of neighbouring genes. This result supports my proposition in Chapter 1 that these copies (of lineage2) are divergent from gene copies of *rbcS*-lineage1.

**Table4. Similarities of neighbouring genes of *rbcS***

| | Corresponding colour in Figure 6 | Species name that neighbouring gene belongs to | | Pair-wise similarities |
|---|---|---|---|---|
| A-gene | Purple | *O.sativa* | *P.hallii* | 77 |
| | | *O.sativa* | *S.italica* | 78 |
| | | *P.hallii* | *S.italica* | 97 |
| | Orange | *S.italica* | *S.viridis* | 100 |
| | Yellow | *O.sativa* | *P.hallii* | 65 |
| | Dark blue | *S.italica* | *S.viridis* | 100 |
| C-gene | Pink | *O.sativa* | *P.hallii* | 76 |
| | Dark green | *P.hallii* | *S.italica* | 61 |
| | | *P.hallii* | *S.viridis* | 61 |
| | | *S.italica* | *S.viridis* | 98 |
| | Ice green | *O.sativa* | *P.hallii* | 79 |
| | Brown | *B.distachyon* | *B.stacei* | 64 |
| | Ice blue | *B.distachyon* | *B.stacei* | 89 |
| | Dark blue | *S.italica* | *S.viridis* | 100 |

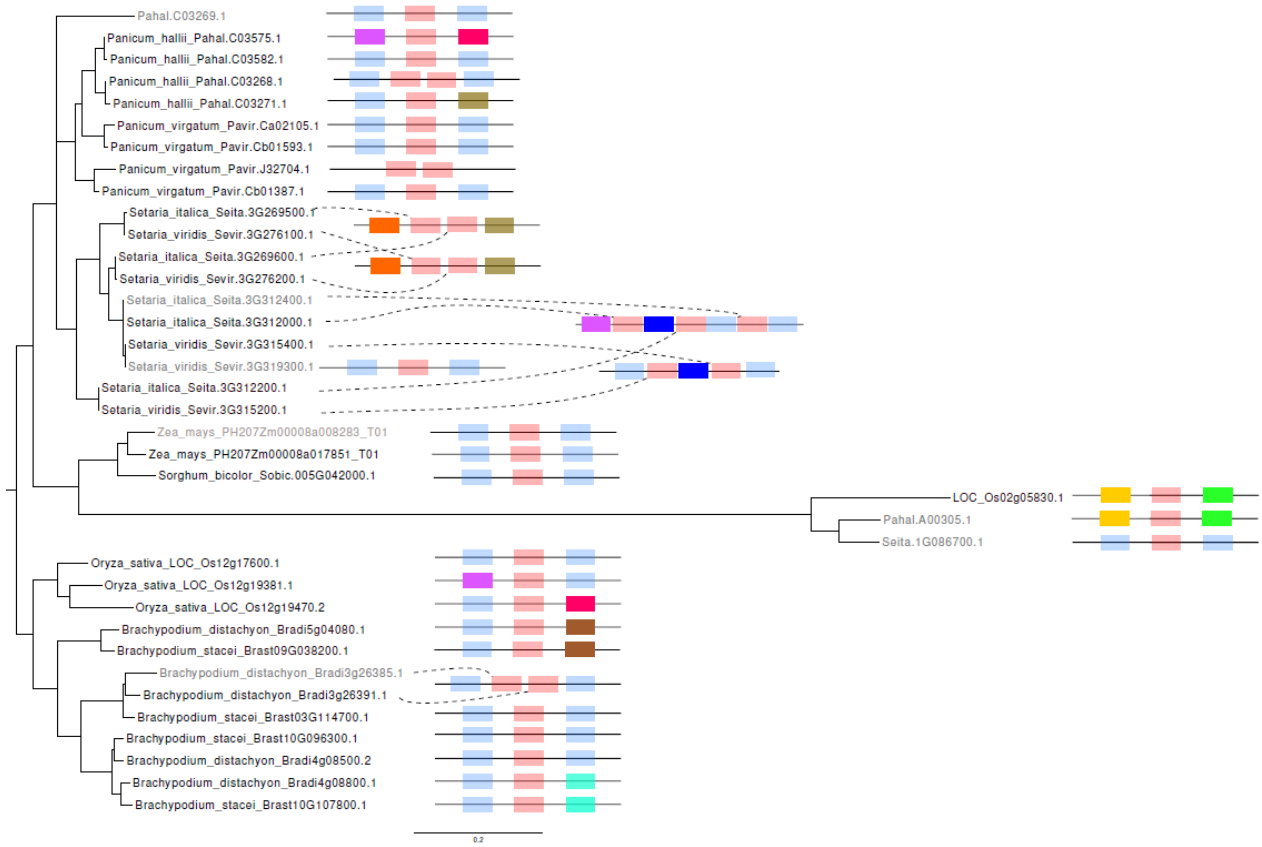



**Figure 6. Phylogenetic relationships of *rbcS* and similarities of neighbouring genes of each *rbcS* copy in Poaceae**

The neighbouring genes of each *rbcS* copy were identified in the genome data of Phytozome version 12 using Jbrowse (Buels et al., 2016). Nucleotide sequences of neighbouring genes, which locate just next to *rbcS* – up-stream (A-gene) and down-stream (C-gene) – were downloaded from Phytozome version 12. Pair-wise similarities based on nucleotide sequences between A-genes of different *rbcS* copies were calculated in Geneious 10. The same calculation was done for their C-genes. The phylogenetic relationships of *rbcS* in Poaceae were extracted from the results shown in Figure 4. Then, the structure of *rbcS* genes and neighbouring genes were drawn in cartoon style next to the phylogenetic tree of *rbcS*. Boxes in light pink are *rbcS* copies. When there was more than 60% similarity between neighbouring genes, these sets of genes were drawn as boxes of the same colour. The similarities of neighbouring genes and corresponding colours are shown in Table 4. Other neighbouring genes are drawn in light blue. The gene copies written in grey are non-expressed ones according to the expression analysis performed in Chapter 3.

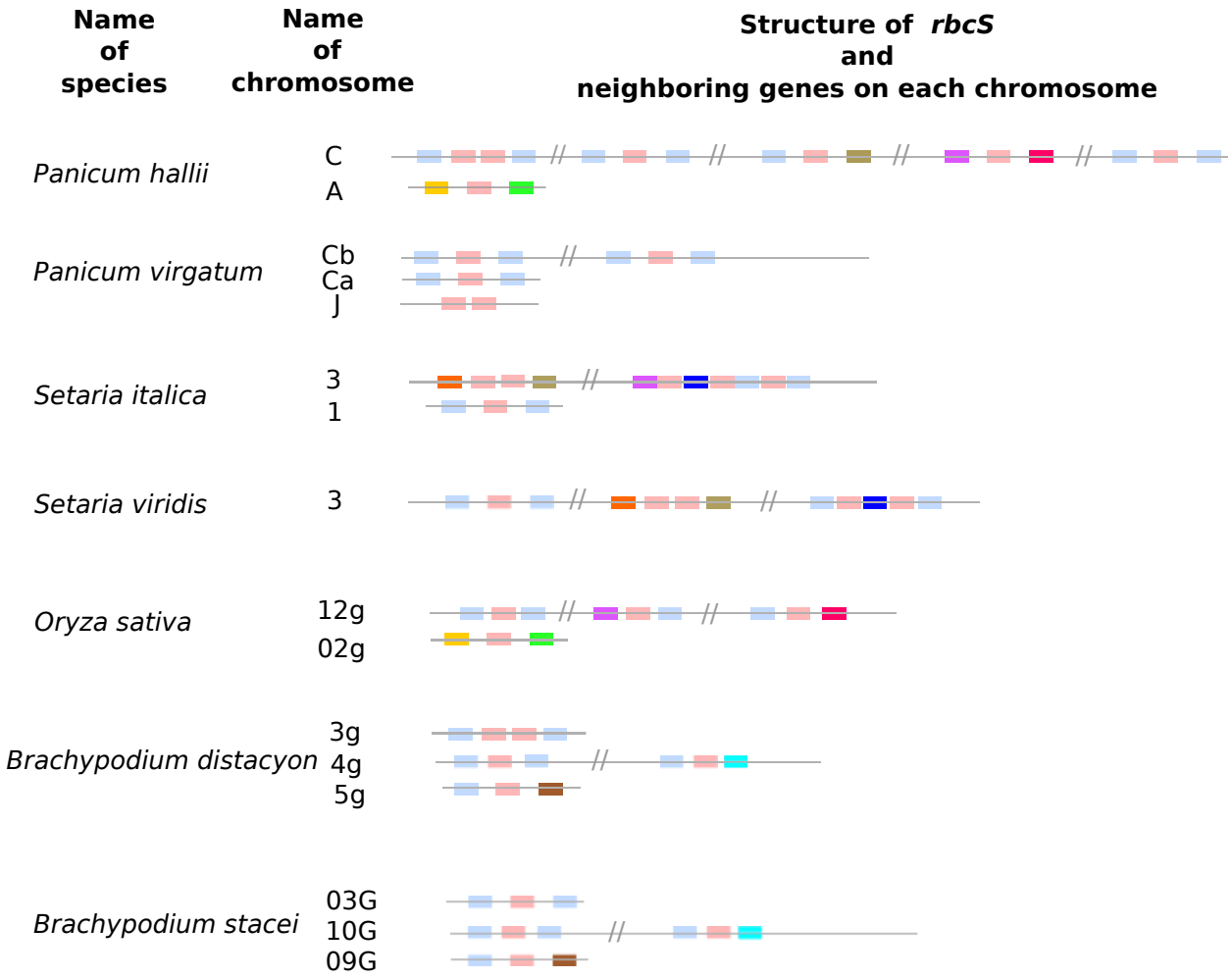



**Figure 7. The location of the *rbcS* gene and its neighbouring genes on each chromosome in Poaceae**

The locations of each *rbcS* copy and its neighbouring genes on each chromosome are drawn in cartoon style. Light pink boxes indicate *rbcS* genes. Boxes of the same colours (except light blue) indicate neighbouring genes that are more than 60% similar to each other based on nucleotide sequences. Other neighbouring genes were coloured in light blue. The genes and box colours correspond to Table 4 and Figure 6. The "//" symbol indicates that there are other genes between them. The length of each chromosome is not reflected in the length of the line in the cartoon.

# Discussion

## *RbcS*-lineage1 forms a single-gene lineage with copy number variation

The phylogenetic tree based on *rbcS* genes implies the same relationships among grass subfamilies as other datasets assumed to represent the species tree (Edwards, 2012; Giussani, Cota-Sánchez, Zuloaga, & Kellogg, 2001; Sánchen-Ken, Gabriel Sánchen-Ken, Clark, Kellogg, & Kay, 2007). Within subfamilies, there are some discrepancies between the relationships based on *rbcS* and those inferred previously on chloroplast or nuclear markers (Christin et al., 2008a; Vicentini et al., 2008). It is, however, not surprising that the relationships differ between chloroplast and nuclear markers, because this has been recurrently reported (Christin, Salamin, Kellogg, Vicentini, & Besnard, 2009; Washburn et al., 2017). In addition, individual gene trees are expected to differ from the species trees because gene trees can be affected by gene duplication, gene loss, gene flow after speciation, and recombination (Mitchell et al., 2013). Also, the differences between gene trees can be caused by a lack of phylogenetic support, phylogenetic errors, and incomplete lineage sorting (Szöllősi, Davín, Tannier, Daubin, & Boussau, 2015; Wiens & Morrill, 2011). Accordingly, previous analyses of individual nuclear genes have recovered phylogenetic trees that presented small differences (Christin, Salamin, Savolainen, Duvall, & Besnard, 2007).

While multiple *rbcS* sequences were detected in a number of species, they grouped most commonly in each species. This indicates that the multiple copies are recent duplicates that emerged after the split of the species sampled here. In cases where closely related species have been sequenced, duplicates predating their split can be observed (e.g. *S. italica* and *Setaria viridis*, *B. distachyon*, and *Brachypodium stacei*). There is, however, only minimal evidence of more ancient duplications, besides the one reported in Chapter 1 that occurred before the origin of land plants, and the ones that occurred before the divergence of Brassicaceae and Rosaceae. I conclude that duplicates, which must have originated at least during the multiple whole genome duplications in the ancestors of grasses (Paterson, Bowers, & Chapman, 2004; Paterson, Bowers, Van de Peer, & Vandepoele, 2005; Paterson

et al., 2006; Wang, Shi, Hao, Ge, & Luo, 2005; Yu et al., 2005), have not been retained in the long term. It is likely that the function in a protein complex composed of the large subunit encoded by the single chloroplast encoded gene *rbcL* prevents functional diversification of duplicates (e.g. neofunctionalization or subfunctionalization; Zhang, 2003) of *rbcS*. Importantly, the emergence of C4 photosynthesis was not linked to an increased retention of duplicates, so that *rbcS* genes in C3 and C4 can be considered as orthologous.

Because the multiple copies existing in numerous species are highly similar and form monophyletic clades (Figure 4), they represent copy number variants rather than duplicates that evolved independently for a consequent amount of time. High levels of gene copy number variation are not uncommon among species or even individuals within the same species (Bianconi, Dunning, Moreno-Villena, Osborne, & Christin, 2018; Cheeseman et al., 2016; Zhang, 2003; Zmieńko, Samelak, Kozłowski, & Figlerowicz, 2014). The genomic context of the duplications was evaluated by comparing neighbouring genes of *rbcS* genes from specific grass species whose genomic information is available. High similarities between neighbouring genes of *rbcS* were found in different species of the same genera (e.g. *Setaria, Brachypodium*). In particular, tandem copies of Seita.3G269500.1 and Seita.3G269600.1 of *S. italica* and tandem copies of Sevir.3G276100.1 and Sevir.3G276200.1 of *S. viridis* have neighbouring genes with more than 97% similarity. The locations of genes on each chromosome of each species suggest that these regions including *rbcS* may have diverged from the same regions of ancestral species of *S. italica* and *S. viridis* (Figure 7). However, it should be noted that significant similarities between neighbouring genes are not in themselves strong evidence to indicate orthologous relationships of genes. There are other pairs of neighbouring genes with high similarities in other species. However, except these few cases described above, the neighbouring genes of *rbcS* were not similar (less than 60% similarity). This suggests that the regions that include *rbcS* have been shuffled by gene conversion, regional duplications, recombination, inversion, and crossing-over. This, in turn, suggests that the *rbcS* evolution is dynamic. Therefore, it is not easy to track the evolutionary histories of *rbcS* gene copies because the

evidence of orthology may disappear quickly by the involvement of complicated evolutionary processes.

### Evidence of positive selection, but not specific to C4 plants

Using tests that considered successively each branch as being under positive selection, I found evidence of positive selection on most branches, independently of the photosynthetic type (Figure 5). This indicates that selective pressures other than the evolution of C4 photosynthesis have lead to increased rates of non-synonymous substitutions on *rbcS.* RuBisCO function is critical to all photosynthetic organisms, and different properties are selected in contrasted environments. While C4 photosynthesis presumably selects for RuBisCO enzymes with faster catalytic rates, which happens at the expense of $CO_2/O_2$ specificity (Tcherkez et al., 2006), arid, saline, and warm habitats selected for higher specificity, thereby reducing catalytic efficiency (Cavanagh & Kubien, 2014; Galmés, Hermida-Carrera, Laanisto, & Niinemets, 2016; Sage, 2002). RuBisCO is therefore expected to be under positive selection both after a switch to C4 photosynthesis and after migration to environments requiring different catalytic properties. Analyses of *rbcL* have accordingly found evidence for positive selection both related to and independent of photosynthetic transitions (Christin et al., 2008a; Kapralov & Filatov, 2007). Because RBCS encoded by *rbcS* has been shown to influence the catalytic efficiency and $CO_2/O_2$ specificity of the enzyme as well as its quantity and activity (Andrews & Ballment, 1983; Bracher et al., 2011; Furbank et al., 2000; Genkov & Spreitzer, 2009; Genkov et al., 2010; Quick et al., 1991; Spreitzer, 2003; Stitt et al., 1991), it is not surprising that *rbcS* genes are also positively selected for a variety of reasons. In the case of *rbcL*, analyses based on a similar sample size were able to distinguish positive selection specific to C4 species from that occurring in all taxa independently of their photosynthetic type (Christin et al., 2008a). In my study, models assuming positive selection in large groups of branches were not significantly better than models without positive selection.

# Conclusions

Studying the *rbcS* tree in Poaceae confirmed the results of Chapter 1 that most of the ancient duplications are removed in the long term in *rbcS* evolution, except the duplication event before the divergence of land plants and duplication events before the divergence of Brassicaceae and Rosaceae. It is easy to identify the orthology in a long-term scale. However, when I look at the recent time scale (e.g. closely related species of the same genera), I see a lot of duplications that reveal the dynamic process of gene evolution. Positive selection acting on the *rbcS* gene seems not be lead by the switch from C3 type to C4 type. The result suggests that RBCS may have been involved in the optimization of RuBisCO after the C4 type was established or after the migration to environments requiring different catalytic properties.

# Chapter 3. Comparison of expression levels of *rbcS* gene copies within and between species

## Introduction

The multigene families are often organized by gene duplication, gene conversion, and crossing over. One of the duplicated genes can obtain a new function or sub-function (Force et al., 1999). Obtaining varieties of functions is evolutionarily advantageous in the long term (Ohta, 1991). On the other hand, when the abundance of protein is required, members of a gene family keep the unique function (Ohta, 1991). The high rate of gene interaction such as gene conversion and/or unequal crossing over can homogenize sequences and subsequently help to maintain a unique function among a multigene family (Ohta, 1980, 1983).

In some cases, the function of each member of a multigene family can diverge only by alternation of gene expression levels and post-translational modification (Ohta, 1991). For example, the specific members of gene families encoding enzymes such as NADP-ME (NADP-malic enzyme) and PCK (phosphoenolpyruvate carboxykinase) in photosynthetic pathway are recruited in C4 type (Christin et al., 2013). The member of the multigene family may differentiate by gene expression rather than novel gene duplication for the development of C4 biochemistry (Brautigam & Mullick, 2011; Külahoglu et al., 2014).

Therefore, together with the result of Chapter 1, I build two hypotheses about gene expression levels of the *rbcS* copies: i) all the gene copies of *rbcS* have a unique function but they differ in their expression levels, and ii) the multi-copy *rbcS* exist to maintain the amount of gene product at the same level as that of single-copy *rbcS* (the dosage effect). In previous studies, it has been reported that the expression of *rbcS* is altered by temperature, $CO_2$ concentration, water deficit, light regulation, tissue, different developmental stages, and cell-specific localization (Cavanagh & Kubien, 2014; Dean et al., 1989; Hudsona et al., 1992;

Manzara et al., 1991; Morita et al., 2014; Thomas-Hall et al., 2007; Wanner & Gruissem, 1991; Zhang et al., 2013). This information is useful, but the focus of these studies was not on the divergence of the gene copies. Thus, they only focused on the total expression of all the *rbcS* gene copies but not the expression level of each gene copy. Few studies discuss the gene expression levels of each *rbcS* gene copy. Cavanagh and Kubien (2014) have shown that highly expressed gene copies are altered by $CO_2$ pressure or growth temperature in *Arabidopsis thaliana*, referring the previous studies (Cheng et al., 1998; Yoon et al., 2001). In *O. sativa*, it has been shown (Morita et al., 2016) that expression was rather located in organs related to metabolic pathway than to the photosynthetic pathway.

This chapter aims to expand scholarly knowledge about the gene expression of each *rbcS* gene copy. The gene expression levels of gene copies were estimated using publicly available RNA-seq data. Poaceae was selected because more was known about the divergence of the *rbcS* family than other families (see Chapter 2). Seven species of Poaceae were selected because of the availability of their genomic data, annotation data, and expression data by RNA-seq. First, the expression levels of gene copies in control conditions were compared to find out whether all the gene copies were equally expressed or whether the expressions of each gene copy were different. Second, I tested the expressions of each gene copy in different tissues to understand if the gene expression levels altered depending on tissues rather than on a change of conditions. Third, I tested the differential expression of gene copies in different conditions to understand if all the copies changed their expression levels in severe conditions or if specific copies responded more sensitively to the change of conditions. Finally, I compared the *rbcS* gene expression levels between species to test whether the presence of a single copy led to higher expression levels than when multiple copies were present.

# Materials and Methods

## Collection of available genome, annotation, and expression data

For the estimation of gene expression, assembled genome, gene annotation, and RNA-seq data were required. To identify all the publicly available RNA-seq data, the Sequence Read Archive (SRA) was queried on 5th October 2017 by searching keywords "RNA-seq" and the species name of Poaceae (e.g. "*Brachypodium distachyon*"). The RNA-seq data from leaf tissue of all the species of Poaceae that had annotated *rbcS* genes on the Phytozome version 12 database (Goodstein et al., 2012) were downloaded. The archive RNA-seq experiments that included control conditions were selected. If the contrast conditions such as drought-stress, cold-stress, or salt soil-stress existed in the same experiments as the control condition, they were also downloaded for the differential expression analyses. Two species of *Panicum* had complete information, as described above; however, they were excluded from the analyses because some copies were not variant enough (less than 1% of sequence divergence between some gene copies within species) to assign reads correctly to different copies. After the search, seven species (*B. distachyon, B. stacei, O. sativa, S. italica, S. viridis, Sorghum bicolor*, and *Z. mays*) were selected to perform the analyses of gene expression. Unfortunately, the experiments of the same contrast condition for all the seven species were not found on the SRA database (queried on 5th October 2017). Thus, the test of differential expression was limited to few species in different conditions: *B. distachyon* in drought conditions, and *O. sativa*, and *S. italica* in cold conditions. Each RNA-seq experiments included multiple biological replicates and all replicates were included in my analyses.

For the expression analyses on different tissues, SRA was queried with the keywords "RNA-seq", "tissue", "different", and the species names of the selected nine species. Then, the experiments performed on different tissues under control conditions were found for *O. sativa, S. italica*, and *S. bicolor*.

Expression data were downloaded from the SRA repository of NCBI using the “fastq-dump” function of SRAtoolkit 2.8.0 (Leinonen, Sugawara, Shumway, & International Nucleotide Sequence Database Collaboration, 2011). The selected IDs of SRA are shown in Table 1. The annotation file in compressed gtf3 format and assembled genome file in fasta format were obtained from the Phytozome database version 12 (Goodstein et al., 2012).

**Table1. List of SRA runs downloaded for expression analyses**

| Species name | SRA run code | Condition or used tissues | Layout |
|---|---|---|---|
| *Brachypodium distachyon* | SRR522511 | Control | Pair |
| | SRR522512 | Control | Pair |
| | SRR522515 | Control | Pair |
| | SRR522516 | Control | Pair |
| | SRR522513 | Dry | Pair |
| | SRR522514 | Dry | Pair |
| | SRR522517 | Dry | Pair |
| | SRR522518 | Dry | Pair |
| *Brachypodium stacei* | DRR090117 | Control | Pair |
| | DRR090118 | Control | Pair |
| | DRR090119 | Control | Pair |
| *Oryza sativa* | SRR3647326 | Control | Pair |
| | SRR3647328 | Control | Pair |
| | SRR3647329 | Control | Pair |
| | SRR3647330 | Control | Pair |
| | SRR1213691 | Flower | Pair |
| | SRR1213692 | Leaf sampled before flowering | Pair |
| | SRR1213694 | Root sampled before flowering | Pair |
| | SRR1213697 | Mature seed | Pair |
| | SRR3647326 | Salt | Pair |
| | SRR3647327 | Salt | Pair |
| | SRR3647328 | Salt | Pair |
| *Setaria italica* | SRR4280407 | Control | Single |
| | SRR4280418 | Control | Single |
| | SRR4280429 | Control | Single |
| | SRR4280406 | Cold | Single |
| | SRR4280417 | Cold | Single |
| | SRR4280428 | Cold | Single |
| | SRR442161 | Root | Pair |
| | SRR442162 | Leaf | Pair |
| | SRR442163 | Stem | Pair |
| | SRR442164 | Tassel | Pair |
| *Setaria viridis* | SRR2319666 | Control | Single |

| | | | |
|---|---|---|---|
| | SRR2320708 | Control | Single |
| | SRR2320953 | Control | Single |
| *Sorghum bicolor* | SRR4280410 | Control | Pair |
| | SRR4280412 | Control | Pair |
| | SRR4280421 | Control | Pair |
| | SRR4280400 | Cold | Pair |
| | SRR4280404 | Cold | Pair |
| | SRR4280415 | Cold | Pair |
| | DRR059875 | Leaf | Pair |
| | DRR059876 | Stem | Pair |
| | DRR059877 | Panicle | Pair |
| | DRR059878 | Leaf | Pair |
| | DRR059879 | Stem | Pair |
| | DRR059880 | Panicle | Pair |
| | DRR059881 | Leaf | Pair |
| | DRR059882 | Stem | Pair |
| | DRR059883 | Panicle | Pair |
| *Zea mays* | SRR4280408 | Control | Pair |
| | SRR4280419 | Control | Pair |
| | SRR4280425 | Control | Pair |
| | SRR4280402 | Cold | Pair |
| | SRR4280413 | Cold | Pair |
| | SRR4280424 | Cold | Pair |

## Preparation of reference genes

Gtf3 files were converted into gtf files using the “gffread” function of Cufflinks 2.2.1 1 (Roberts, Pimentel, Trapnell, & Pachter, 2011; Roberts, Trapnell, Donaghey, Rinn, & Pachter, 2011; Trapnell et al., 2010, 2012). Gene annotation data and genome data were given as inputs and the reference sequences were extracted using the “rsem-prepare-reference” function of RSEM (Li & Dewey, 2011) with the “--bowtie2” alignment option (Langmead & Salzberg, 2012).

## Calculation of expression

The gene expression was estimated using the “rsem-calculate-expression” function of RSEM using the “EM” method with the option of “--bowtie2”. When the paired-end file was used as input, the option “--paired-end” was added.

## Normalization

### 1. Control conditions and tissue-specific comparisons

The estimated expression levels of each *rbcS* gene copy were extracted. The transcript per million ("TPM") considering the gene length was used to compare the relative abundance of transcripts. The TPM values between gene copies of the same species were compared. The statistical differences of expressions were calculated by one-way ANOVA using the "aov" and "TukeyHSD" functions in the R package "stats v.3.4.3" (Tierney, 2012). The proportion of expression levels of each gene copy out of the total expression of all the gene copies were calculated (Figures 1-a, 1-b, and 1-c).

### 2. Differential expression

The differential expression of *rbcS* gene copies in control conditions versus contrast conditions was tested using the R package "limma voom" (Law, Chen, Shi, & Smyth, 2014). According to Robinson et al. (2010) read count is the preferred method to be used in normalization between different conditions. The values of "expected read counts" for each run were concatenated and normalized using the function "calcNormFactors" from the R package edgeR (Robinson et al., 2010). The fitting to the linear model was tested using the "lmFit" function to normalize the data considering the library size.

### 3. The comparison of *rbcS* gene expression between gene copies of different species

The comparison of gene expression between gene copies of different species was conducted by taking each species as different batches. I used the function "Combat" of the package "sva" (Leek, Johnson, Parker, Jaffe, & Storey, 2012) in R. Combat normalizes the gene expression level of the target gene by referring to the gene expression levels of orthologous genes. I used Orthologous Matrix (OMA, Altenhoff et al., 2018) to identify orthologs. Among the seven species of Poaceae that are used in this chapter, five species (*B. distachyon, O. sativa, S. italica, S. bicolor*, and *Z. mays*) already have pre-computed

information about the orthologous relationships in the OMA. By definition, the orthologs change depending on included species and considered diversification events. Thus, the orthology within these five species was estimated by first computing all-against-all Smith-Waterman alignments of five species from OMA. Then, I ran the function “oma” to obtain the orthologous relationships of genes among species and name of encoding protein. When the names of proteins were “Uncharacterized” or included the word “uncharacterized”, I removed them to prevent using unknown genes as references of normalization. To match the IDs of genes between different genomic data, I used the “translation tool” of MaizeGDB (https://www.maizegdb.org/) for *Z. mays* and the “ID converter” of The Rice Annotation Project Database (http://rapdb.dna.affrc.go.jp/tools/converter) for *O. sativa*. For other species, I manually converted ID of OMA to ID of Phytozome. I removed the genes for which I could not find corresponding ID. In the end, I obtained 49 reference genes to be used as target genes.

The TPM values of each *rbcS* gene copy and reference gene were estimated using the same method as previously explained in the section on control conditions. Combat requires two types of files. One file includes expression data of the target gene and reference genes. The other file defines which columns of the first file need to be taken as the same batch (the same species in this case). In the first file, TPM values of each *rbcS* gene copy of all the species were entered in the first row. From the second row for 49 rows, I entered the TPM value of each reference gene. The same column included TPM values calculated from exactly the same run of RNA-seq (from the same species, the specific run among biological replicates). Then, I normalized the TPM value of each *rbcS* copy of each species by running the command “Combat”. I plotted the log2 value of normalized expression of each *rbcS* copy of different species using the boxplot function of the “graphic” package of R.

**Table 2. Example input file for Combat including TPM values of *rbcS* and reference genes**

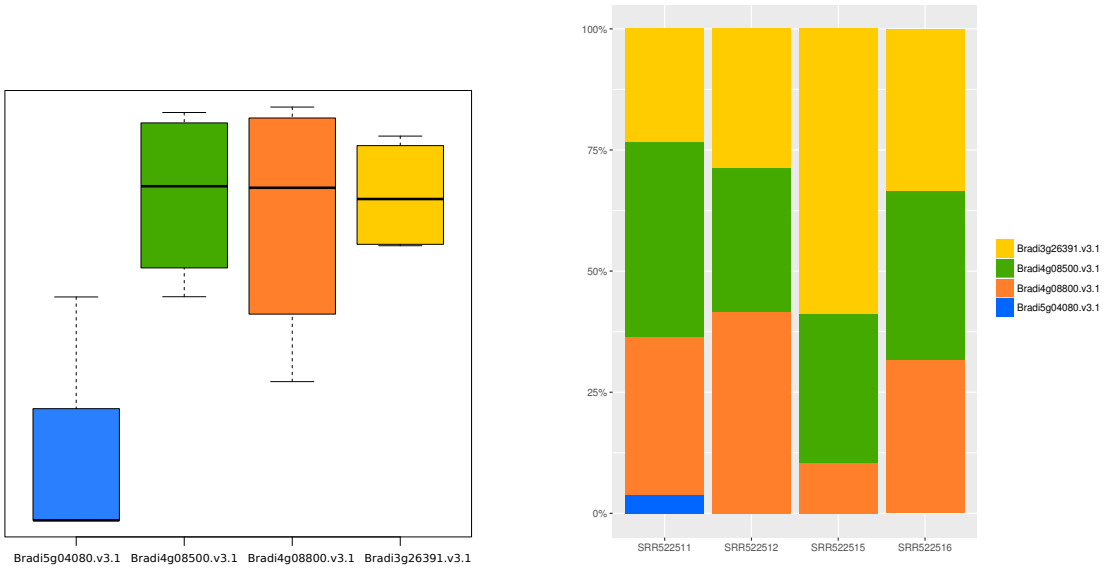
| *Osa* Bio.rep.1 | | *Osa* Bio.rep.2 | | *Sit* Bio.rep.1 | | *Sit* Bio.rep.2 | |
|---|---|---|---|---|---|---|---|
| *Osa* *rbcS*copy1 | *Osa* *rbcS*copy2 | *Osa* *rbcS*copy1 | *Osa* *rbcS*copy2 | *Sit* *rbcS*copy1 | *Sit* *rbcS*copy2 | *Sit* *rbcS*copy1 | *Sit* *rbcS*copy2 |
| Ref.gene1 | Ref.gene1 | Ref.gene1 | Ref.gene1 | Ref.gene1 | Ref.gene1 | Ref.gene1 | Ref.gene1 |
| Ref.gene2 | Ref.gene2 | Ref.gene2 | Ref.gene2 | Ref.gene2 | Ref.gene2 | Ref.gene2 | Ref.gene2 |
| Ref.gene3 | Ref.gene3 | Ref.gene3 | Ref.gene3 | Ref.gene3 | Ref.gene3 | Ref.gene3 | Ref.gene3 |
| Ref.gene4 | Ref.gene4 | Ref.gene4 | Ref.gene4 | Ref.gene4 | Ref.gene4 | Ref.gene4 | Ref.gene4 |

# Results and Discussion

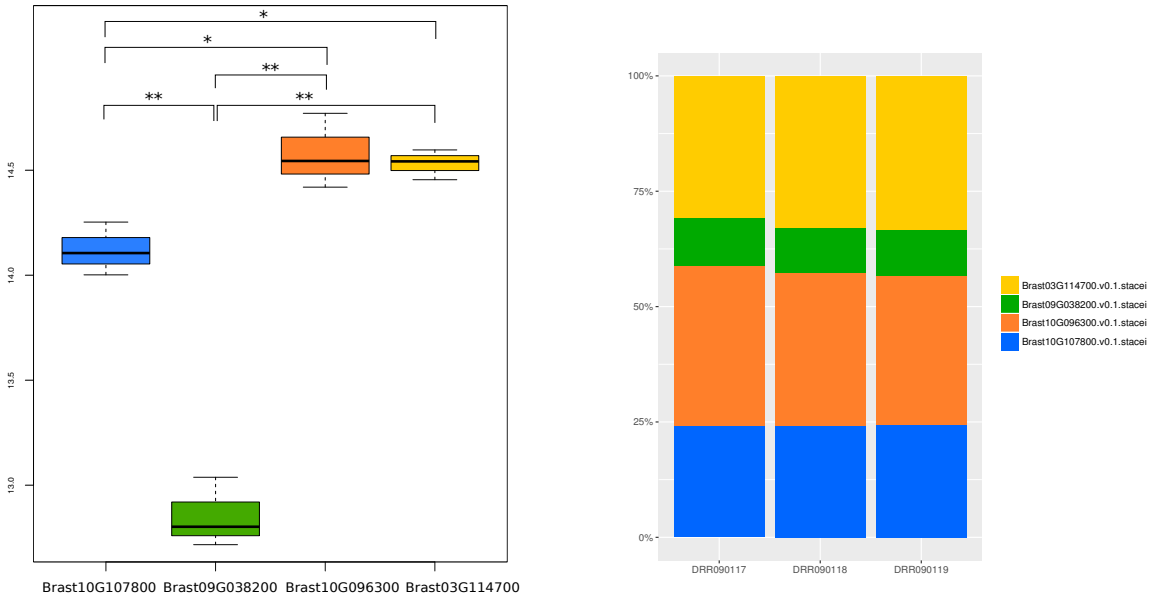
In this chapter, I estimated the gene expression levels of each *rbcS* copy by using transcriptome data. The results of Chapter 1 and 2 have shown that coding regions of *rbcS* are highly conserved between gene copies within species. However, the promoter region of *rbcS* can be divergent among gene copies and gene expression levels can differ between copies (shown in *S. lycopersicum*; Manzara et al., 1991). In this chapter, I first tested the hypothesis that each *rbcS* copy of the same species may have different expression levels by comparing the gene expression levels (TPM value) of each gene copy in leaf tissues in control conditions. I found that the gene expression levels were different between gene copies of the same species and one gene copy was more lowly expressed, either significantly or relatively, than the other copies. Then, I tested the second hypothesis that the multiple copies of one species may exist to maintain the gene products at the same level as that of single species carrying species (the dosage effect hypothesis: Birchler & Veitia, 2012; Papp et al., 2003). I tested this hypothesis by estimating gene expression levels (TPM) of gene copies in single or multiple carrying species under control conditions. I found that the expression levels of single-*rbcS* and multiple-*rbcS* were similar. To better understand the differences of gene expression levels between copies, I compared them in different tissues (e.g. leaf, root, tassel) and different conditions (e.g. drought, salty soil, and cold).

There were significant differences in the expression levels within species, which shows that copies are not equally expressed in control conditions in leaf tissues (Figure 1-a, 1-b, 1-c). Among the seven species tested, *B.distacyon, B. stacei, O. sativa*, *S. italica*, and *S. viridis* carried multiple-*rbcS* and more than two expressed copies. One copy was significantly (in *B. stacei, O. sativa, S. italica*) or relatively (in *B. distachyon, S. viridis*) lowly expressed in comparison with other copies within the species. Although the trend of gene expressions was similar in *B. distachyon* and *S. italica* as in the other three species, the significance was not detected, probably due to the larger variation of gene expression levels between

biological replicates (Figure 1-a and 1-c). The lowest expressed copies of four species, except for *O. sativa*, belong to *rbcS*-lineage 1; however, phylogenetic analysis suggests that, within each species, these lowly expressed copies are more divergent than copies of *rbcS*-lineage1 with high levels of expression (Figure 2). The lowest expressed copy of *O. sativa*, LOC_Os02g05830, belongs to *rbcS*-lineage2. To better understand the expression profile of the gene copies, the expression levels of each gene copy in control conditions extracted from different tissues were tested in *O. sativa*, *S. italica*, and *S. bicolor* (Figures 3-a and 3-b).



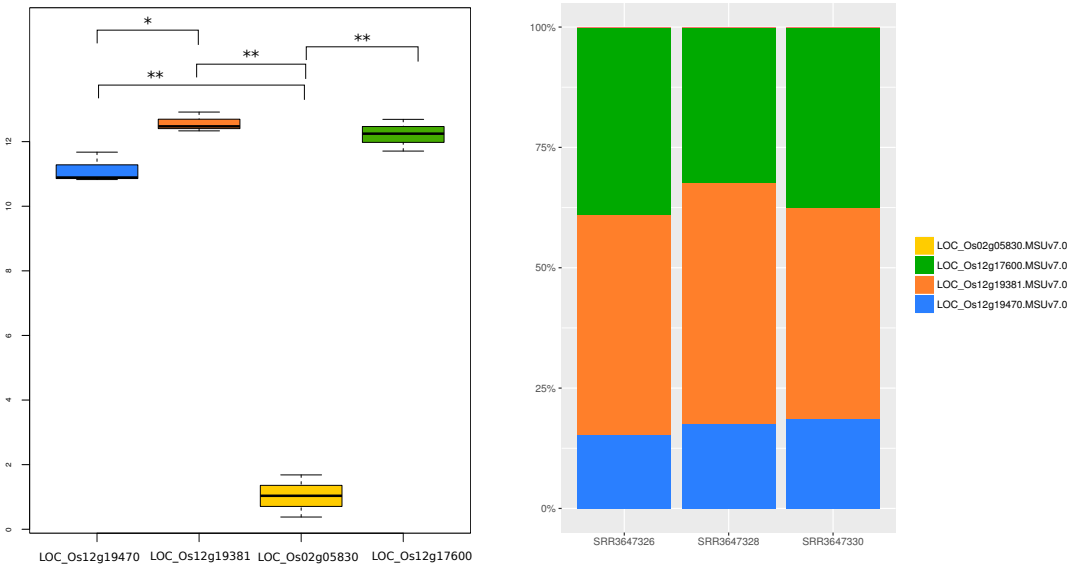

*Brachypodium distachyon*




*Brachypodium stacei*

**Figure 1-a. Comparison of expression between *rbcS* gene copies of two species of *Brachypodium* under control conditions in leaf tissues**

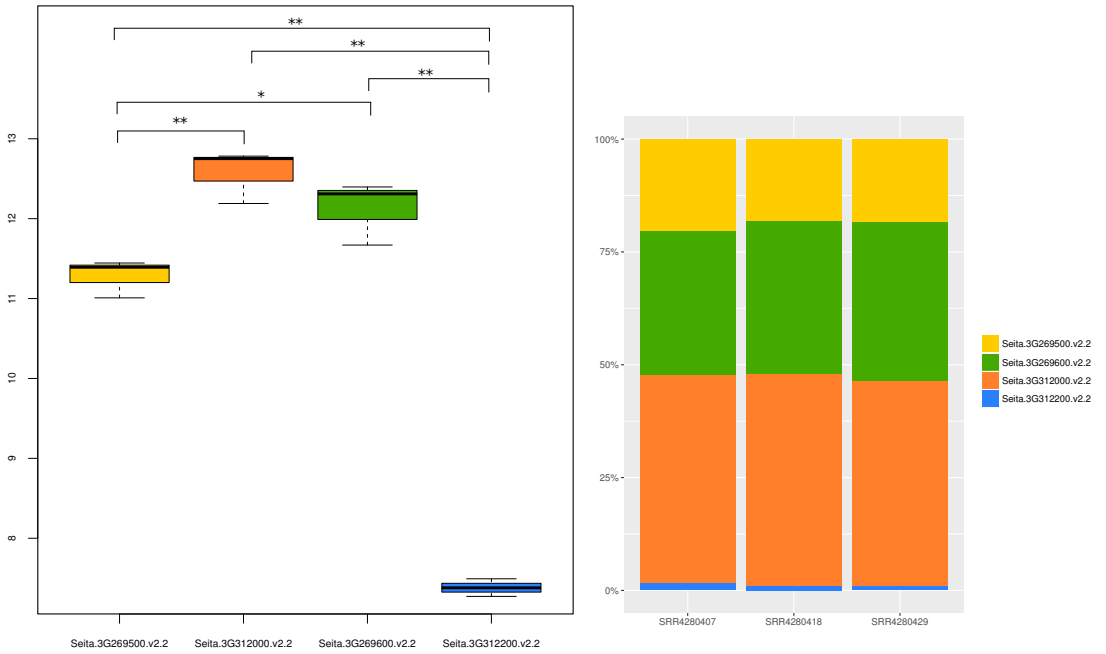
The gene expression levels (TPM) of each *rbcS* gene copy were estimated using the "rsem-calculate-expression" function of RSEM (Li et al., 2011). Biological replicates of the same RNA-seq experiments were conducted four and three times for *Brachypodium distacyon* and *Brachypodium stacei*, respectively. The log2 of TPM were calculated and plotted in a boxplot (shown on the left side). The proportion of expression levels of each gene copy among total expression was calculated and plotted using the ggplot2 package (Wickham et al., 2009) of R (shown on the right side). Each bar plot corresponds to each run (biological replicates) of RNA-Seq. The significant differences were tested using Turkey's HSD test in R and the significant differences are shown by **(p-value<0.01) and *(p-value<0.05).



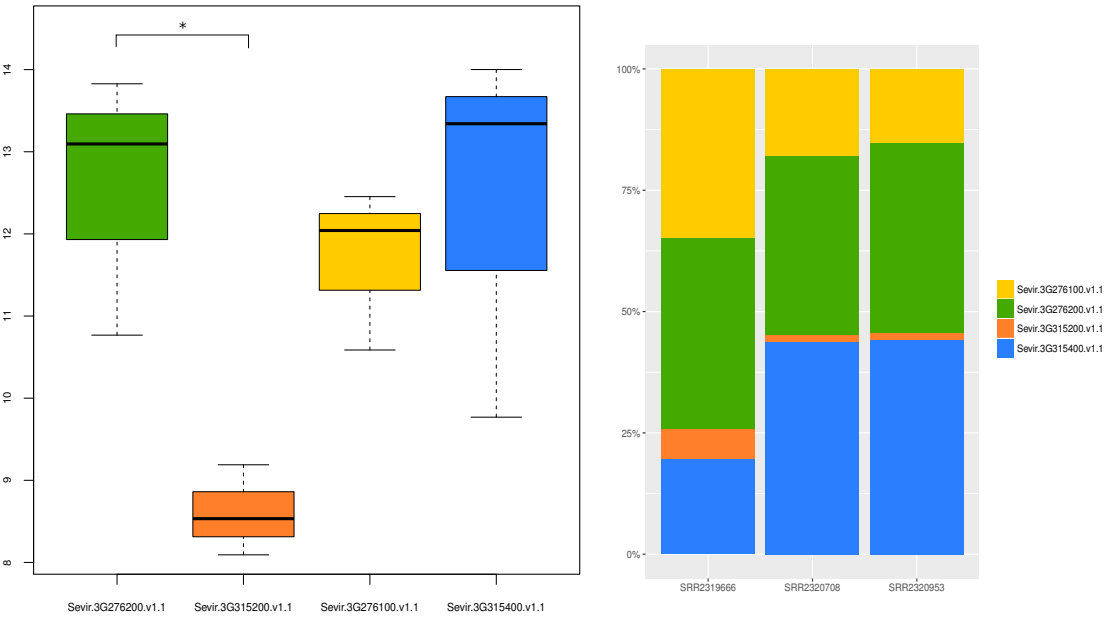

*Oryza sativa*

**Figure 1-b. Comparison of expression between *rbcS* gene copies of *Oryza sativa* under control conditions in leaf tissues**

The gene expression levels (TPM) of each *rbcS* gene copy were estimated using the "rsem-calculate-expression" function of RSEM (Li et al., 2011). Biological replicates of the same RNA-seq experiments were conducted three times. The log2 of TPM were calculated and plotted in a boxplot (shown on the left side). The proportion of expression levels of each gene copy among total expression was calculated and plotted using the ggplot2 package (Wickham et al., 2009) of R (shown on the right side). Each bar plot corresponds to each run (biological replicates) of RNA-Seq. The significant differences were tested by Turkey's HSD test in R and the significant differences are shown by **(p-value<0.01) and *(p-value<0.05).




*Setaria italica*




*Setaria viridis*

**Figure 1-c. Comparison of expression between *rbcS* gene copies of two species of *Setaria* under control conditions in leaf tissues**

The gene expression levels (TPM) of each *rbcS* gene copy were estimated using the "rsem-calculate-expression" function of RSEM (Li et al., 2011). Biological replicates of the same RNA-seq experiments were conducted three times for each species. The log2 of TPM were calculated and plotted in a boxplot (shown on the left side). The proportion of expression levels of each gene copy among total expression was calculated and plotted using the ggplot2 package (Wickham et al., 2009) of R (shown on the right side). Each bar plot corresponds to each run (biological replicates) of RNA-Seq. The significant differences were tested by Turkey's HSD test in R and the significant differences are shown by **(p-value<0.01) and *(p-value<0.05)

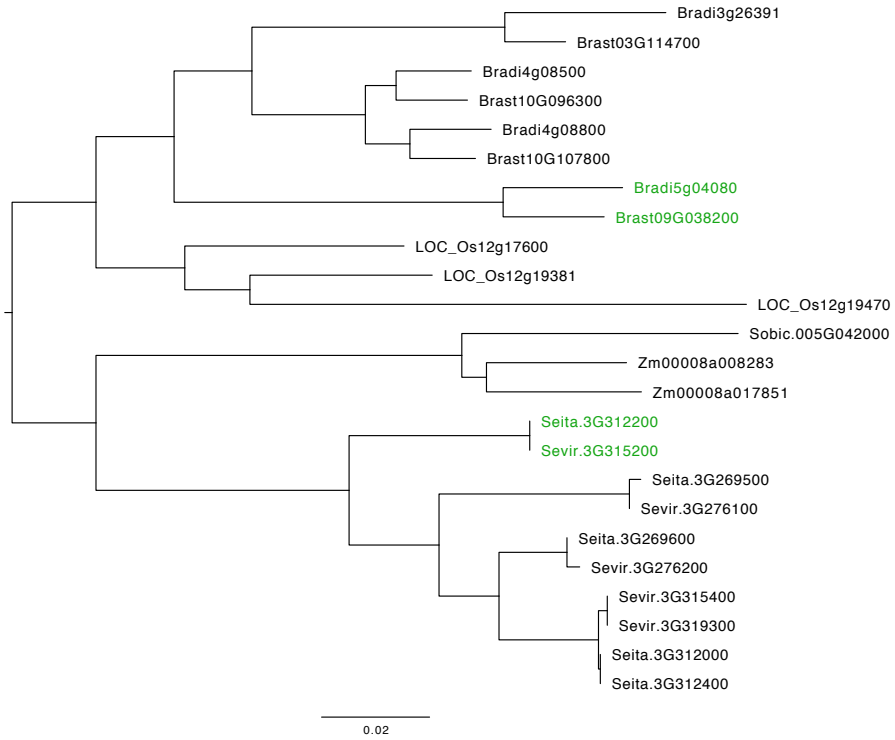



**Figure 2. The *rbcS* gene tree of species used for gene expression analysis in control conditions**
The sequences of *rbcS* of species that I used for comparing gene expression levels under control conditions were extracted from the alignment that I used to reconstract the phylogenetic tree of *rbcS* among angiosperms in Chapter 1. The tree was reconstructed using the Maximum Likelihood method of PhyML3.0 with the GTR model. The gene copies that are most lowly expressed within each species are shown in green.

Unexpectedly, the lowest expressed copy of *O. sativa* (LOC_Os02g05830, *rbcS*-lineage2) in leaf was the highest expressed copy in root (Figure 3-a). However, the lowest expressed copy of *S. italica* (Seita.3G312200.v2.2, *rbcS*-lineage1) in leaf was also the lowest expressed copy in root. This copy of *S. italica* and the lowest expressed copy of *B. distacyon, B. stacei*, and *S. viridis* were divergent from other copies of the same species. A ;ower expression level could indicate a change of function or a potential loss of function. The single *rbcS* of *S. bicolor* was highly expressed in leaf and lowly expressed in root. This suggests that the copy expressed in leaf tissue may play a more important role of RBCS and the predominantly expressed copy in root may not be necessary to survive. Further tests of the expression of single copy carrying species are required to confirm this proposition.

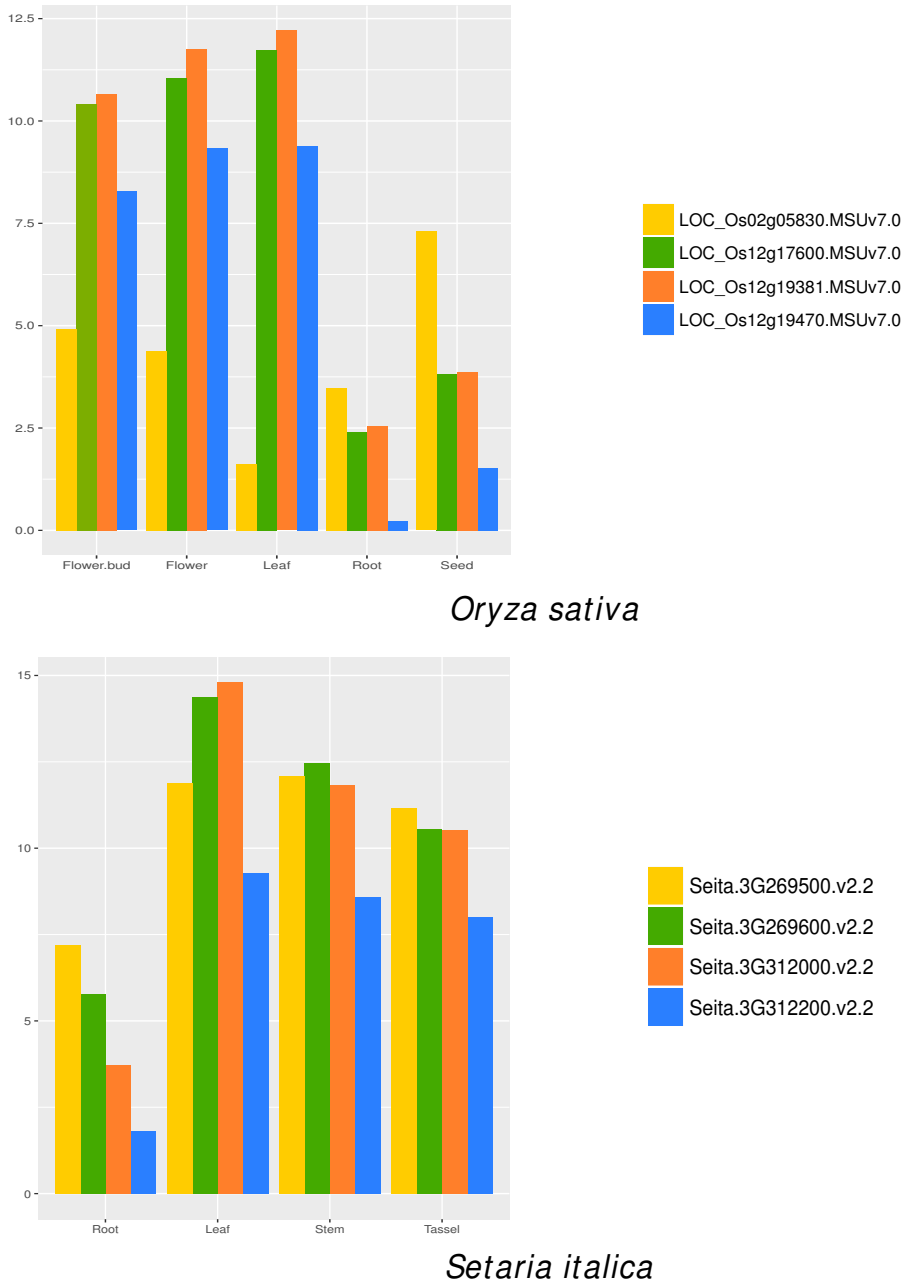



**Figure 3-a. Expression levels of each *rbcS* gene copy in different tissues of *Oryza sativa* and *Setaria italica* under control conditions**
The gene expression levels (TPM) of each *rbcS* gene copy in different tissues under control conditions were estimated using the “rsem-calculate-expression” function of RSEM (Li et al., 2011). There were no biological replicates in both experiments. The log2 of TPM were calculated and plotted as a bar plot using the ggplot2 package (Wickham et al., 2009) of R. Flower bud, flower, leaf root, and seed were used for *O. sativa* and root, leaf, stem, and tassel were tested in *S. italica*.

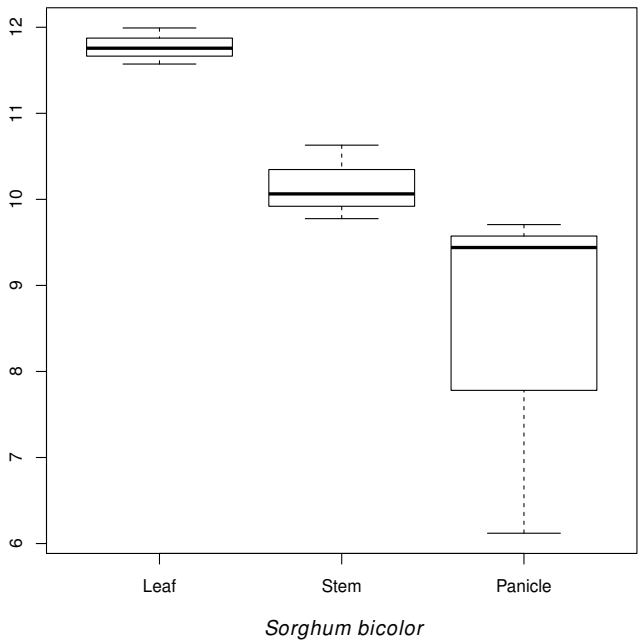



**Figure 3-b. Expression levels of each *rbcS* gene copy in different tissues of *Sorghum bicolor* in control conditions**
The gene expression levels (TPM) of each *rbcS* gene copy in different tissues under control conditions were estimated using the "rsem-calculate-expression" function of RSEM (Li et al., 2011). Three were biological replicates for each tissue. The log2 of TPM were calculated and plotted as a bar plot using the ggplot2 package (Wickham et al., 2009) of R. The expression was tested in leaf, stem, and panicle.

Only *O. sativa* and *S. italica* had gene copies that belonged to *rbcS*-lineage2. The copy of *S. italica* (Seita.1G086700.1) was not found in the annotation file of Phytozome. In the previous study, *rbcS*-lineage2-like genes were identified based on the similarities of amino acid sequences to LOC_Os02g05830 (*rbcS*-lineage2 of *O. sativa*) (Morita et al., 2016). Then, they investigated the expression of *rbcS*-lineage2-like copies of five species of angiosperms in different organs (E.g. seed of *S. italica*, stamen, pistil, and green fruit of *S. lycopersicum*, root, nodule, and seed of in *Lotus japonicus*, mature leaves and green berry of *Vitis. vinifera*, rhizome, and root of *Selaginella moellendorffii*). These results suggest that *rbcS*-lineage2 is expressed mainly in non-photosynthetic organs, but also some in photosynthetic organs (e.g. mature leaf of *V. vinifera*). *RbcS*-lineage2 has been maintained for millions of years since the divergence of land plants and this lineage might have acquired a different function from the *rbcS*-lineage1. To test the role played by *rbcS*-lineage2 in *O. sativa*, Morita and his colleagues overexpressed the gene of the root-predominant copy (LOC_Os02g05830, the same as the copy belong to *rbcS*-lineage2 in my study) then incorporated it into RuBisCO.

The engineered RuBisCO had improved catalytic efficiency and decreased $CO_2$ specificity (Morita et al., 2014). This result supports my proposition that lineage2 is functional and it may have a function that is different from *rbcS*-lineage1. However, explaining why the lineage2 of *rbcS* can still carry some important function through the increased catalytic efficiency, while being maintained in only very few species, remains a challenge. It may be because the function may no longer be necessary for some species. Alternatively, it is possible that the modified RuBisCO of Morita's study (Morita et al., 2014) increased catalytic efficiency, but this may have come with some negative "side effect" such as an increased rate of photorespiration that may result in the loss of energy or a non-preferable influence on the function carried out by photorespiration (Peterhansel et al., 2010). Besides that, it is interesting that the *rbcS* copy is expressed in root. The evidence of low expression of *rbcL* was detected in root (Isono, Fukushima, Kawakatsu, & Nakajima, 1997). This suggests that little amount of RuBIsCO is synthesized in the root. However, the question is: what is the function of RuBisCO in root? In previous studies, RBCS have been suggested to have functions related to the catalytic properties, $CO_2$ specificity, quantity, activity, structural stability, regulation of mRNA of *rbcL* (Andrews & Ballment, 1983; Bracher et al., 2011; Furbank et al., 2000; Genkov & Spreitzer, 2009; Genkov et al., 2010; Quick et al., 1991; Spreitzer, 2003; Stitt et al., 1991; Suzuki & Makino, 2012). Also, the specific regions on the surface of RBCS have reported to have higher affinity to $CO_2$ and captured $CO_2$ on these regions may migrate to the catalytic sites of closest RBCS (van Lun et al., 2014). However, none of the suggested functions seem likely to be necessary in root. According to our current knowledge, it is not possible to indicate what the function of RuBIsCO is in root. Morita et al. (2014) have proposed that RuBisCO may have some function in the metabolic pathway; however, their proposition is simply because it is the alternative idea of function in the photosynthetic pathway and clear evidence was not provided.

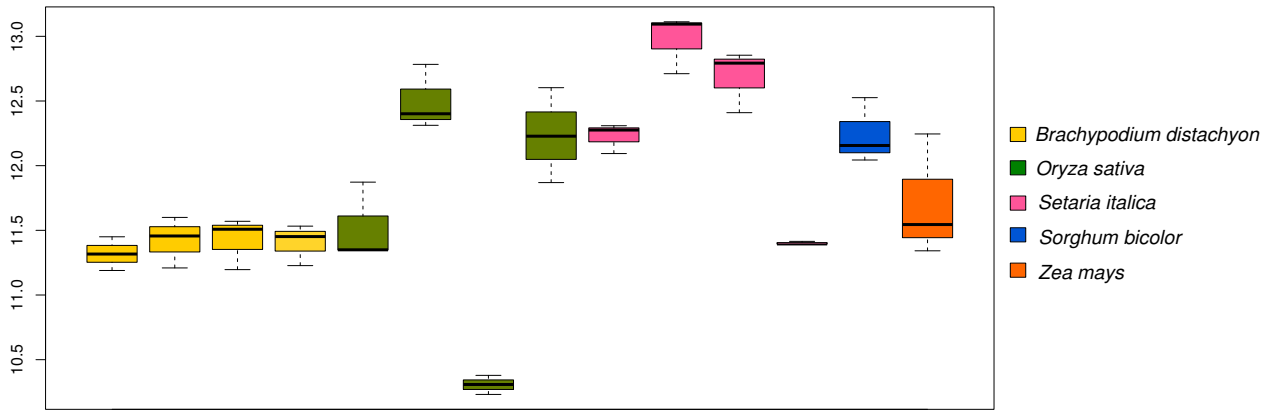



**Figure 4. Comparison of gene expression levels of *rbcS* gene copies between species**
The gene expression levels (TPM) of each *rbcS* gene copy in leaf tissues in control conditions were estimated using the "rsem-calculate-expression" function of RSEM (Li et al., 2011). The gene expression levels between gene copies were normalized using the Combat method (Leek et al., 2017) that takes into account species as batches and makes the TPM values comparable between species. The expression levels of 49 orthologous genes were used to normalize the TPM values among species. Then, TPM values were plotted using the boxplot function of ggplot2 (Wickham et al., 2009) in R.

The second hypothesis – that each *rbcS* copy between species may have different expression levels – was rejected. The comparisons of gene expression levels between species showed that the single-*rbcS* and multiple-*rbcS* are expressed at similar levels (Figure 4). Therefore, I suppose that the species carrying higher number of gene copies may have larger quantity of RBCS. The species carrying higher numbers of gene copies may have larger quantities of RBCS. This proposition is congruent with previous studies showing that the amount of RBCS differs between species (e.g. the comparison between *Eucalyptus globulus* and *O. sativa* in Suzuki, Kihara-Doi, Kawazu, Miyake, & Makino, 2010). These authors have suggested that the species-dependent difference was related to post-transcriptional processes of the expression of the *rbcS* gene (Suzuki et al., 2010). Then, the question is: why do some species need more RBCS and some do not? Previous studies have shown that the amount of RuBisCO decreases with drought stress (Bartholomew, Bartley, & Scolnik, 1991; Bota, Medrano, & Flexas, 2004; Carmo-Silva et al., 2007; da Silva & Arrabaça, 1995; Lal, Ku, & Edwards, 1996; Parry, 2002; Tezara, Mitchell, Driscoll, & Lawlor, 2002; Vu, Gesch, Hartwell Allen, Boote, & Bowes, 1999). If the amount of RuBisCO decreases because of degradation or down-regulation of RuBisCO by stress, plants that live

in drought conditions may synthesize more RuBisCO to prevent the shortage of RuBisCO to survive. To test if the amount of transcripts increases in specific environmental conditions, I compared the gene expression levels of each copy in different environmental conditions (Figure 5; Table 3). All copies of *O. sativa* (both lineage1 and lineage2) were significantly more lowly expressed in salt conditions than in control conditions. This may be because transcripts may be down-regulated or degraded under salty conditions. This result suggests that, similarly to my proposition that all the gene copies are down-regulated in stress conditions, a higher number of copies are required to have a sufficient amount of RBCS. However, in cold conditions, only one copy of *S. italica* (of *rbcS*-lineage1) was significantly highly expressed relative to the other gene copies. I suppose that only specific conditions such as drought and salt stress may degrade or down-regulate RuBisCO and, for prevention of insufficient RuBisCO, a greater number of gene copies may be required. However, in other stress conditions such as cold conditions, the higher expression of specific copies may be able to increase the amount of the most suitable isoform to fold with RBCL in this kind of stress. This proposition supports my conclusion in Chapter 2 that RBCS is probably involved in the optimization of RuBisCO depending on the environmental habitat.

**Table 3. Differential expression of the *rbcS* gene in different environmental conditions**

| Environmental Condition | Name of species | Name of *rbcS* gene copy | F.p.value |
|---|---|---|---|
| Cold | *Setaria italica* | Seita.3G269500.v2.2 | 0.0240794647701591 * |
| | | Seita.3G312000.v2.2 | 0.104768074741861 |
| | | Seita.3G269600.v2.2 | 0.1380309653 |
| | | Seita.3G312200.v2.2 | 0.0950052899444647 |
| Dry | *Brachypodium distachyon* | Bradi5g04080.v3.1 | 0.0664417418734997 |
| | | Bradi4g08500.v3.1 | 0.471101005576619 |
| | | Bradi4g08800.v3.1 | 0.538114812079926 |
| | | Bradi3g26391.v3.1 | 0.598838004950032 |
| Salt soil stress | *Oryza sativa* | LOC_Os12g19470.MSUv7.0 | 0.00363236939283557 ** |
| | | LOC_Os12g19381.MSUv7.0 | 0.00283809597499853 ** |
| | | LOC_Os02g05830.MSUv7.0 | 0.00015797204316198 ** |
| | | LOC_Os12g17600.MSUv7.0 | 0.00211061757656644 ** |

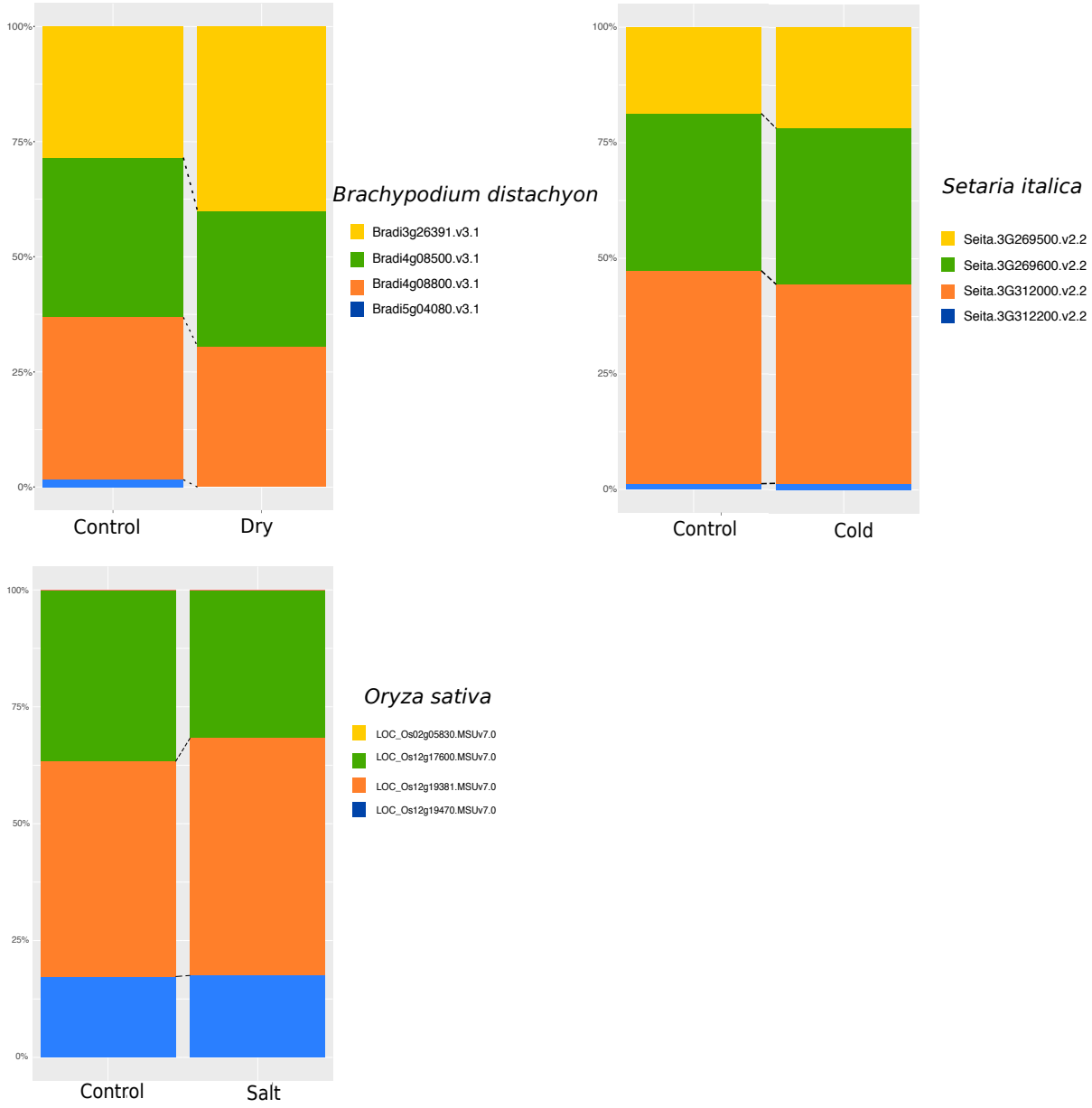



**Figure 5. Differential expression of *rbcS* in different environmental conditions in leaf tissues**

The gene expression levels (expected read counts) of each *rbcS* gene copy at different condition in leaf were estimated using the "rsem-calculate-expression" function of RSEM (Li et al., 2011). The expression levels were tested in *B. distachyon* in dry conditions, *S. italica* in cold conditions, and *O. sativa* in salt soil conditions. Three biological replicates existed for each experiment. The differential expression level of each gene copy was estimated using the package edgeR (Robinson et al., 2010) of R. The proportion of expression of each gene copy among total expression of all the gene copies within the same species within each condition was plotted based on the calculated TPM values. The averages of proportions of each copy were calculated among three biological replicates and proportions were plotted using the bar plot function of "ggplot2" package (Wickham et al., 2009) in R.

In this chapter, I have studied the gene expression levels of *rbcS* copies within and between species. The number of species and conditions was limited. However, among all the tested multi-*rbcS* species, one of the gene copies was more lowly expressed than others within the species. I suppose that the species carrying higher number of *rbcS* may have a larger quantity of RBCS. The required amount of RBCS is probably species-dependent and it may differ according to the environmental stresses of the habitat of each species. The relationships between the number of gene copies (total expression of *rbcS*) and ecological habitats (stress) need to be further examined to better understand the role played by *rbcS*.

# General Discussion and Future Perspectives

In this thesis, the evolution of the *rbcS* multigene family has been studied to gain a better understanding of the evolution of RuBisCO. It has addressed three main questions: 1) how has the *rbcS* multigene family evolved? 2) was the *rbcS* gene involved in the evolution of C4 photosynthesis? 3) how do the *rbcS* gene copies differ? Firstly, the phylogenetic tree of *rbcS* has revealed two lineages that may have originated from a duplication event before the divergence of land plants. The observed pattern of higher similarities between gene copies of the same species may be the result of recent species-specific duplications. Ancient duplicates seem not to be retained for the long term in *rbcS* evolution. Secondly, I found that positive selection acting on the *rbcS* gene is not C4-specific. This suggests that RBCS have not been involved in the transition from C3 to C4 type photosynthesis and that positive selection of *rbcS* may have been driven by other reasons besides the evolution of C4 photosynthesis. Thirdly, I found that the coevolution pattern with *rbcL* and the stabilities of the encoding subunit in the RuBisCO structure were the same among different gene copies of *rbcS.* The gene expression levels of each gene copy were similar, but a few gene copies that are expressed at low level in leaf and dominantly expressed in root were found. The gene expression levels individual genes among species were almost the same as between single-copy and multiple-copies carrying species. Therefore, the amount of RBCS may be species-dependent and the required amount of RBCS may change depending on environmental habitat. Together with the result of positive selection, I propose that RBCS may optimize the catalytic properties, structure, and stability of RuBisCO after the transition from C3 and C4 type, or after the migration of plants to new environmental conditions. Initially, I hypothesized that the *rbcS* multigene family may have maintained many lineages as it had been already reported in other multigene families of the photosynthetic pathway (e.g. four lineages for NADP-ME) (Badger & Price, 1994; Christin et al., 2013; Ku et al., 1996; Schaffner & Sheen, 1992). The phylogenetic relationships of *rbcS* revealed that there

are two lineages that seem to have originated from a duplication event before the divergence of land plants. Interestingly, all the species carry lineage1, but only a limited number of species carry lineage2. Most of the duplicates that may have been created during whole genome duplications seem to have been removed. On the other hand, recent duplications (e.g. species-specific duplications) seem to be retained. Because the retention rate of duplicated genes seems to be low in the long term, it is easy to track the evolution of *rbcS* in ancient times; however, the dynamics of *rbcS* evolution make it difficult to track recent evolutionary processes acting on *rbcS.*

The recent duplications and low retention rate of duplicates are probably the reasons for the variations in gene copy numbers among species. The comparisons of gene expression levels of gene copies among species have shown that *rbcS* gene copies are similarly expressed in all the species in spite of gene copy numbers per species. Previous studies have shown that the amount of RuBisCO is probably species-dependent (Suzuki et al., 2010) and RBCS has influences on the quantity of RuBisCO and RBCL (Suzuki & Makino, 2012). My results can propose additional explications to previous studies that the number of gene copies of *rbcS* may determine the amount of RBCS per species and the amount of RBCL and RuBisCO may have changed depending on the amount of RBCS. Then, the new question is why some species need more RuBisCO than other species. Degradation of RuBisCO has been reported in various biotic and abiotic stresses (Cheng et al., 1998; Esquível, Ferreira, & Teixeira, 2000; Makino, Mae, & Ohira, 1985). Therefore, I propose that more RBCS may have been synthesized in plants that live in stress conditions to ensure sufficient enzymes for survival. I estimated the differential expression of gene copies in few species in different environmental stresses. I should note that my results are not sufficient to indicate the response of the expression levels to the different environmental conditions because the tested species and conditions were limited. In my results, all the gene copies of *O. sativa* were down-regulated in salt soil conditions, but only one copy of *S. italica* was up-regulated in cold conditions. I propose that the environmental stresses can be sorted into two types depending on the influence on RuBisCO: 1) the type of stress that degrades or

down-regulates RuBisCO and 2) the type of stress that does not lead to the degradation of RuBisCO but for which optimization is required to recover the catalytic properties. In the first type of stress, all the gene copies of the same species may be down-regulated. On the other hand, in the second type of stress conditions, only specific copies may react to the environmental change. For further understanding, first it is necessary to test the differential expressions of *rbcS* copies within the same species in several stress conditions to reveal if there are two types of stress conditions for RuBisCO as I propose. Second, I want to test the specific copies reacting to the second type of environmental stress, and verify if the specific copies can encode RBCS that are more suitable for survival in specific environmental stress. Additional to the comparison of gene expressions, the investigation of sequences of promoter regions of *rbcS* may help to understand the regulation of gene expression levels of each gene copy in different environmental conditions.

The environmental conditions may have an important influence on the expression of *rbcS*. However, the result of positive selection has suggested that the evolution of C4 photosynthesis did not alter the selective pressures in *rbcS.* On the other hand, some studies have suggested that RBCS may improve the catalytic efficiency of RuBisCO (Karkehabadi, Peddi, & Anwaruzzaman, 2005; Spreitzer, 2003). I suppose this can be explained by the general knowledge about protein folding and stabilities. Most enzymes are composed of two domains (subunits): catalytic domains and regulatory domains. In the case of RuBisCO, RBCLs are the catalytic domains and RBCSs are the regulatory domains. To have better catalytic efficiency, folding of catalytic sites need to be modified. Other motifs on the catalytic domain (e.g. RBCL) will change corresponding to the modification on catalytic sites. Then, the motifs of the regulatory domain (e.g. RBCS) will also change corresponding with the modification of the catalytic domain. These interactions may also occur in the oposite direction. Therefore, I suppose that RBCS itself does not have a function to improve catalytic efficiency because it does not include catalytic sites (Andersson, 2008). However, the modification of the motifs of RBCS may influence the motifs of the catalytic sites of RBCL. It has already been shown that there is a trade-off between the catalytic efficiency

and $CO_2$ specificity of RuBisCO (Tcherkez et al., 2006). Therefore, I suppose that the improvement of the $CO_2$ specificity is also not the direct function of RBCS and it may have been driven by the modification of the motifs of subunits.

From the results of coevolution and protein stability, I found that each isoform of RBCS interacts with RBCL in the same manner and that carrying different isoforms in a RuBisCO structure does not change the stability of the enzyme. The composition of isoforms of RBCS in a RuBisCO structure has not been understood. I modelled the different combinations of isoforms of RBCS encoded by different *rbcS* genes to compare the stability between them in *Arabidopsis thaliana* (results not shown). There are no combinations that result in extreme loss or gain of stability. Thus, I suppose that keeping the same level of interaction with RBCL and the similar levels of protein stability are important for RBCS because it allows RuBisCO to be composed of any combination of RBCS isoforms. This system is probably advantageous because the interaction with RBCL and the overall stability of RuBisCO will stay stable when the composition of RBCS isoforms changes depending on the environmental conditions.

Surprisingly, one copy of *O. sativa* and one copy of *S. italica* were the lowest expressed copies in leaf but most predominantly expressed in root. Previously, the *rbcS* gene copies that are expressed in non-photosynthetic organs have been reported in fives species of angiosperms (Morita et al., 2016). This suggests that a small amount of RuBisCO exists in non-photosynthetic organs. I suppose that the *rbcS* gene copies expressed in photosynthetic organs and non-photosynthetic organs may have different origins. The existence of the different types of RuBisCO (types I and II) in the same organisms has already been shown in some prokaryotes (Tabita, 1999; Watson & Tabita, 2006). Having two types of RuBisCO is beneficial because different types of enzymes are regulated differently in different environmental conditions (Spreitzer & Salvucci, 2002). The *rbcS* gene from red algae is more closely related to the *rbcS* of proteobacteria than cyanobacteria (Delwiche et al., 1996). Considering the endosymbiosis theory, I propose that photosynthetic *rbcS* and non-

photosynthetic *rbcS* may have originated from RuBisCO of cyanobacteria (green-type) and RuBisCO of proteobacteria (red-type), respectively. Joshi and his colleagues (Joshi, Mueller-Cajar, Tsai, Hartl, & Hayer-Hartl, 2015) have shown that red-type RuBisCO has better catalytic efficiency than green-type RuBisCO and red-type has the extension in C-terminal Beta-hairpin of RBCS. The overexpression of the root-predominant copy of *O. sativa* has improved the catalytic efficiency of RuBisCO (Morita et al., 2014) and also the C-terminal Beta-hairpin of RBCS encoded by this copy was longer than that of other gene copies (from sequences downloaded from Phytozome in Chapter 1). These few characteristics of the root-predominant copy seem to match with the characteristics of *rbcS* of red-type RuBisCO. However, I proposed this hypothesis based only on the observation of *O. sativa*, so further experiments are required to test it. I propose first to reconstruct the phylogenetic tree of *rbcS* including different types of RuBisCO of bacteria and higher plants. Then I will verify if the root-predominant *rbcS* of higher plants and *rbcS* from red-type RuBisCO cluster together. Also, overexpression of *rbcS* gene copies highly similar to the root-predominant *rbcS* of *O. sativa* (*rbcS* gene copies of other species that are highly similar to the root-predominant copy of *O. sativa* have suggested by Morita et al., 2016) and testing the catalytic efficiency of chimeric RuBisCO may also help to identify if the red-RuBisCO-like characteristics can be commonly observed in root-predominant *rbcS*.

The evolution of the *rbcS* gene family has been studied in this thesis. The results were not congruent with my prior expectations. Initially, I supposed that RBCS was involved in the transition from C3 to C4. However, my results have shown that RBCS is probably more involved in the optimization of RuBisCO after the C4 evolution or migration to the new environment. My results have helped to discuss more profoundly about previously suggested functions of RBCS. The improvement of the catalytic efficiency and the influence on the $CO_2$ specificity are probably not the direct functions of RBCS. However, the amount of RBCS, RBCL, and RuBisCO are probably determined depending on the gene copy number of *rbcS.* I suggest that the amount of RuBisCO and differential expression of each *rbcS* copy may

play a role in the adaptation of RuBisCO to the new environmental habitat. This needs to be tested in future experiments.

# References


Altenhoff, A. M., Glover, N. M., Train, C.-M., Kaleb, K., Warwick Vesztrocy, A., Dylus, D., de Farias, T. M., Zile, K., Stevenson, C., Long, J. Redestig, H., Gonnet, G.H., Dessimoz, C. (2018). The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. *Nucleic Acids Research*, *46*(D1), D477–D485.

Altenhoff, A. M., Schneider, A., Gonnet, G. H., & Dessimoz, C. (2011). OMA 2011: orthology inference among 1000 complete genomes. *Nucleic Acids Research*, *39*(Database issue), D289–D294.

Altenhoff, A. M., Škunca, N., Glover, N., Train, C.-M., Sueki, A., Piližota, I., Gori, K., Tomiczek, B., Müller, S., Redestig, H., Gonnet, G. H., Dessimoz, C. (2015). The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements. *Nucleic Acids Research*, *43*(D1), D240–D249.

Andersson, I. (2008). Catalysis and regulation in Rubisco. *Journal of Experimental Botany*. *59*(7), 1555-1568

Andrews, T. J., & Ballment, B. (1983). The function of the small subunits of ribulose bisphosphate carboxylase-oxygenase. *Journal of Biological Chemistry*, *258*(12), 7514-7518.

Badger, M. R., & Andrews, T. J. (1987). CO-Evolution of Rubisco and CO2 Concentrating Mechanisms. *Progress in Photosynthesis Research,* 601–609.

Badger, M. R., & Price, G. D. (1994). The Role of Carbonic Anhydrase in Photosynthesis. *Annual Review of Plant Physiology and Plant Molecular Biology*, *45*(1), 369–392.

Bartholomew, D. M., Bartley, G. E., & Scolnik, P. A. (1991). Abscisic Acid Control of *rbcS* and cab Transcription in Tomato Leaves. *Plant Physiology*, *96*, 291–296.

Bass, J. D., Swcf, A. J., Dabney, A., & Robinson, D. (2015). qvalue: Q-value estimation for false discovery rate control. R package version 2.12.0, http://github.com/jdstorey/qvalue.

Bauwe, H., Hagemann, M., Kern, R., & Timm, S. (2012). Photorespiration has a dual origin and manifold links to central metabolism. *Current Opinion in Plant Biology*, *15*(3), 269–275.

Beerling, D. J., & Royer, D. L. (2011). Convergent Cenozoic CO2 history. *Nature Geoscience*, *4*(7), 418–420.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57*(1), 289-300.

Benton, R. (2015). Multigene Family Evolution: Perspectives from Insect Chemoreceptors. *Trends in Ecology & Evolution*, *30*(10), 590–600.

Besnard, G., Muasya, A. M., Russier, F., Roalson, E. H., Salamin, N., & Christin, P. A. (2009). Phylogenomics of C4 photosynthesis in sedges (Cyperaceae): Multiple appearances and genetic convergence. *Molecular Biology and Evolution, 26*(8), 1909-1919.

Bianconi, M. E., Dunning, L. T., Moreno-Villena, J. J., Osborne, C. P., & Christin, P.-A. (2018). Gene duplication and dosage effects during the early emergence of C4 photosynthesis in the grass genus Alloteropsis. *Journal of Experimental Botany*, *69*(8), 1967–1980.

Birchler, J. A., & Veitia, R. A. (2012). Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(37), 14746–14753.

Bota, J., Medrano, H., & Flexas, J. (2004). Is photosynthesis limited by decreased Rubisco activity and RuBP content under progressive water stress? *The New Phytologist*, *162*(3), 671–681.

Bracher, A., Starling-Windhof, A., Ulrich Hartl, F., & Hayer-Hartl, M. (2011). Crystal structure of a chaperone-bound assembly intermediate of form I Rubisco. *Nature Publishing Group*, *18*(8), 875-880.

Bräutigam, A., & Mullick, T. (2011). Critical assessment of assembly strategies for non-model species mRNA-Seq data and application of next-generation sequencing to the comparison of C3 and C4 species. *Journal of Experimental Botany*. *62*(9), 3093-3102.

Buels, R., Yao, E., Diesh, C. M., Hayes, R. D., Munoz-Torres, M., Helt, G., Goodstein, D. M., Elsik, C. G., Lewis, S. E., Stein, L., Holmes, I. H. (2016). JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biology*, *17*, 66.

Carmo-Silva, A. E., Soares, A. S., da Silva, J. M., da Silva, A. B., Keys, A. J., & Arrabaça, M. C. (2007). Photosynthetic responses of three C4grasses of different metabolic subtypes to water deficit. *Functional Plant Biology: FPB*, *34*(3), 204-213.

Cavanagh, A. P., & Kubien, D. S. (2014). Can phenotypic plasticity in Rubisco performance contribute to photosynthetic acclimation? *Photosynthesis Research*, *119*(1-2), 203–214.

Chakrabarti, S., & Panchenko, A. R. (2010). Structural and functional roles of coevolved sites in proteins. *PloS One*. *5*(1), e8591.

Cheeseman, I. H., Miller, B., Tan, J. C., Tan, A., Nair, S., Nkhoma, S. C., De Donato, M., Rodulfo, H., Dondorp, A., Branch, O. H., Mesia, L. R., Newton, P., Mayxay, M., Amambua-Ngwa, A., Conway, D. J., Nosten, F., Ferdig, M. T., Anderson, T. J. C. (2016). Population structure shapes copy number variation in malaria parasites. *Molecular Biology and Evolution*. *33*(3), 603-620.

Cheng, S. H., Moore, B., & Seemann, J. R. (1998). Effects of short- and long-term elevated CO2 on the expression of ribulose-1,5-bisphosphate carboxylase/oxygenase genes and carbohydrate accumulation in leaves of *Arabidopsis thaliana* (L.) Heynh. *Plant Physiology*, *116*(2), 715–723.

Christin, P.-A., Besnard, G., Samaritani, E., Duvall, M. R., Hodkinson, T. R., Savolainen, V., & Salamin, N. (2008b). Oligocene $CO_2$ decline promoted C4 photosynthesis in grasses. *Current Biology: CB*, *18*(1), 37–43.

Christin, P.-A., Boxall, S. F., Gregory, R., Edwards, E. J., Hartwell, J., & Osborne, C. P. (2013). Parallel recruitment of multiple genes into c4 photosynthesis. *Genome Biology and Evolution*, *5*(11), 2174–2187.

Christin, P.-A., Sage, T. L., Edwards, E. J., Ogburn, R. M., Khoshravesh, R., & Sage, R. F. (2011). Complex evolutionary transitions and the significance of c(3)-c(4) intermediate forms of photosynthesis in Molluginaceae. *Evolution; International Journal of Organic Evolution*, *65*(3), 643–660.

Christin, P.-A., Salamin, N., Kellogg, E. a., Vicentini, A., & Besnard, G. (2009). Integrating phylogeny into studies of C4 variation in the grasses. *Plant Physiology*, *149*(1), 82–87.

Christin, P.-A., Salamin, N., Muasya, a. M., Roalson, E. H., Russier, F., & Besnard, G. (2008a). Evolutionary switch and genetic convergence on *rbcL* following the evolution of C4 photosynthesis. *Molecular Biology and Evolution*, *25*(11), 2361–2368.

Christin, P.-A., Salamin, N., Savolainen, V., Duvall, M. R., & Besnard, G. (2007). C4 Photosynthesis evolved in grasses via parallel adaptive genetic changes. *Current Biology: CB*, *17*(14), 1241–1247.

Clark, G. B., Sessions, A., Eastburn, D. J., & Roux, S. J. (2001). Differential expression of members of the annexin multigene family in *Arabidopsis*. *Plant Physiology*, *126*(July), 1072–1084.

Darriba, D., Taboada, G. L., Doallo, R., & Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods*, *9*(8), 772.

da Silva, J. M., & Arrabaça, M. C. (1995). Effects of Water Stress on Rubisco Activity of Setaria sphacelata (C4). In *Photosynthesis: from Light to Biosphere*, 3509–3512.

Davydov, I. I., Robinson-Rechavi, M., & Salamin, N. (2017). State aggregation for fast likelihood computations in molecular evolution. *Bioinformatics* , *33*(3), 354–362.

Dean, C., Pichersky, E., & Dunsmuir, P. (1989). STRUCTURE, EVOLUTION, AND REGULATION OF *RbcS* GENES IN HIGHER PLANTS SUMMARY OF GENE STRUCTURE AND EVOLUTION. *Annu. Rev. Plant Physiol. Plant Mol. Bioi*, *40*, 415–439.

Delwiche, C. F., & Palmer, J. D. (1996). Rampant horizontal transfer and duplication of rubisco genes in eubacteria and plastids. *Molecular Biology and Evolution*, *13*(6), 873–882.

Dib, L., Meyer, X., Artimo, P., Ioannidis, V., Stockinger, H., & Salamin, N. (2015). Coev-web: a web platform designed to simulate and evaluate coevolving positions along a phylogenetic tree. *BMC Bioinformatics*, *16*, 394.

Dib, L., Silvestro, D., & Salamin, N. (2014). Evolutionary footprint of coevolving positions in genes. *Bioinformatics* , *30*(9), 1241–1249.

Dumont, B. L., & Eichler, E. E. (2013). Signals of historical interlocus gene conversion in human segmental duplications. *PloS One*, *8*(10), e75949.

Edwards, E. J. (2012). New grass phylogeny resolves deep evolutionary relationships and discovers C4 origins. *The New Phytologist*, *193*(2), 304–312.

Edwards, E. J., Osborne, C. P., Strömberg, C. A. E., Smith, S. A., & Consortium, C. G. (2010). REVIEW The Origins of C 4 Grasslands: Integrating Evolutionary and Ecosystem Science, (April), 587–592.

Edwards, E. J., Still, C. J., & Donoghue, M. J. (2007). The relevance of phylogeny to studies of global change. *Trends in Ecology & Evolution*. *22*(5), 243-249.

Ellis, R. J. (1979). The most abundant protein in the world. *Trends in Biochemical Sciences*, *4*(11), 241–244.

Epskamp, S., Cramer, A., Waldorp, L., Schmittmann, V., & Borsboom, D. (2012). qgraph: Network Visualizations of Relationships in Psychometric Data. *Journal of Statistical Software*, *48*(1), 1–18.

Erb, T., Zarzycki, J. (2017), A short history of RubisCO: the rise and fall (?) of Nature's predominant $CO_2$ fixing enzyme. *Current Opinion in Biotechnology*, *49*, 100-107.

Esquível, M. G., Ferreira, R. B., & Teixeira, A. R. (2000). Protein degradation in C3 and C4 plants subjected to nutrient starvation. Particular reference to ribulose bisphosphate carboxylase/oxygenase and glycolate oxidase. *Plant Science: An International Journal of Experimental Plant Biology*, *153*(1), 15–23.

Eswar, N., Eramian, D., Webb, B., Shen, M.-Y., & Sali, A. (2008). Protein structure modeling with MODELLER. *Methods in Molecular Biology* , *426*, 145–159.

Eyun, S.-I. (2013). The origin and molecular evolution of two multigene families: G-protein coupled receptors and glycoside hydrolase families (Doctoral dissertation) Retrieved from DigitalCommons@University of Nebraska - Lincoln.

Flagel, L. E., & Wendel, J. F. (2009). Gene duplication and evolutionary novelty in plants. *New Phytologist*. *183*(3), 557-564.

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L., & Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, *151*(4), 1531–1545.

Furbank, R. T., Hatch, M. D., & Jenkins, C. L. D. (2000). C4 Photosynthesis: Mechanism and Regulation. *Photosynthesis*, 435–457.

Galmés, J., Hermida-Carrera, C., Laanisto, L., & Niinemets, Ü. (2016). A compendium of temperature responses of Rubisco kinetic traits: variability among and within photosynthetic groups and impacts on photosynthesis modeling. *Journal of Experimental Botany*, *67*(17), 5067–5091.

Genkov, T., Meyer, M., Griffiths, H., & Spreitzer, R. J. (2010). Functional hybrid rubisco enzymes with plant small subunits and algal large subunits: Engineered rbcS cDNA for expression in chlamydomonas. *The Journal of Biological Chemistry*. *285*(26), 19833-19841.

Genkov, T., & Spreitzer, R. J. (2009). Highly conserved small subunit residues influence Rubisco large subunit catalysis. *The Journal of Biological Chemistry*. *284*(44), 30105-30112.

Giussani, L. M., Cota-Sánchez, J. H., Zuloaga, F. O., & Kellogg, E. A. (2001). A molecular phylogeny of the grass subfamily Panicoideae (Poaceae) shows multiple origins of C4 photosynthesis. *American Journal of Botany*, *88*(11), 1993–2012.

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N, Rokhsar, D. S. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research*, *40*(Database issue), D1178–D1186.

Grula, J. W., & Hudspeth, R. L. (1987). The phosphoenolpyruvate carboxylase gene family of maize. *Plant Gene Systems and Their Biology*. *99*(1), 87-94.

Guindon, S., & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, *52*(5), 696–704.

Hakes, L., Lovell, S. C., Oliver, S. G., & Robertson, D. L. (2007). Specificity in protein interactions and its relationship with sequence diversity and coevolution. *Proceedings of the National Academy of Sciences* , *104*(19) 7999-8004.

Hatch, M. D., & Slack, C. R. (1968). A new enzyme for the interconversion of pyruvate and phosphopyruvate and its role in the C4 dicarboxylic acid pathway of photosynthesis. *Biochemical Journal*, *106*(1), 141–146.

Horn, J. W., Xi, Z., Riina, R., Peirson, J. a., Yang, Y., Dorsey, B. L., Berry, P. E., Davis, C. C., Wurdack, K. J. (2014). Evolutionary bursts in Euphorbia (Euphorbiaceae) are linked with photosynthetic pathway. *Evolution; International Journal of Organic Evolution*, *68*(12), 3485–3504.

Hudsona, G. S., Dengler, R. E., Hattersleya, P. W., & Dengler, N. G. (1992). Cell-specific Expression of Rubisco Small Subunit and Rubisco Activase Genes in C3 and C4 Species of Atriplex. *Australian Journal of Plant Physiology*, *19*, 89–96.

Hudson, G. S., Evans, J. R., von Caemmerer, S., Arvidsson, Y. B., & Andrews, T. J. (1992). Reduction of ribulose-1,5-bisphosphate carboxylase/oxygenase content by antisense RNA reduces photosynthesis in transgenic tobacco plants. *Plant Physiology*, *98*(1), 294–302.

Hughes, A. L. (1994). The evolution of functionally novel proteins after gene duplication. *Proceedings. Biological Sciences / The Royal Society*, *256*(1346), 119–124.

Innan, H., & Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews. Genetics*, *11*. 97-108.

Isono, Y., Fukushima, K., Kawakatsu, T., & Nakajima, M. (1997). Integration of charged membrane into perstraction system for separation of amino acid derivatives. *Biotechnology and Bioengineering*, *56*(2), 162–167.

Jordan, D. B., & Ogren, W. L. (1983). Species variation in kinetic properties of ribulose 1,5-bisphosphate carboxylase/oxygenase. *Archives of Biochemistry and Biophysics*, *227*(2), 425–433.

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. *The Journal of Biological Chemistry*, *290*(2), 1066–1074.

Kafri, R., Dahan, O., Levy, J., & Pilpel, Y. (2008). Preferential protection of protein interaction network hubs in yeast: evolved functionality of genetic redundancy. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(4), 1243–1248.

Kanai, R., & Edwards, G. E. (1999). The biochemistry of C4 photosynthesis. *C4 Plant Biology*. 49-87.

Kapralov, M. V., & Filatov, D. a. (2007). Widespread positive selection in the photosynthetic Rubisco enzyme. *BMC Evolutionary Biology*, *7*, 73.

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*(4), 1491-1503.

Kapralov, M. V., Smith, J. A. C., & Filatov, D. a. (2012). Rubisco evolution in c(4) eudicots: an analysis of amaranthaceae sensu lato. *PloS One*, *7*(12), e52974.

Karkehabadi, S., Peddi, S. R., & Anwaruzzaman, M. (2005). Chimeric small subunits influence catalysis without causing global conformational changes in the crystal structure of ribulose-1, 5-bisphosphate carboxylase/oxygenase. *Biochemistry*. *44*(29), 9851-9861.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., Drummond, A. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* , *28*(12), 1647–1649.

Knight, S., Andersson, I., & Brändén, C. I. (1990). Crystallographic analysis of ribulose 1,5-bisphosphate carboxylase from spinach at 2.4 A resolution. Subunit interactions and active site. *Journal of Molecular Biology*. *215*(1), 113-160.

Kokubun, N., Ishida, H., Makino, A., & Mae, T. (2002). The Degradation of the Large Subunit of Ribulose-1 , 5-bisphosphate Carboxylase / oxygenase into the 44-kDa Fragment in the Lysates of Chloroplasts Incubated in Darkness, *43*(11), 1390–1395.

Kubien, D. S., Whitney, S. M., Moore, P. V., & Jesson, L. K. (2008). The biochemistry of Rubisco in Flaveria. In *Journal of Experimental Botany*. *59*(7), 1767-1777.

Külahoglu, C., Denton, A. K., Sommer, M., Maß, J., Schliesky, S., Wrobel, T. J., Berckmans, B., Gongora-Castillo, E., Buell, C. R., Simon, R., De Veylder, L., Bräutigam, A., Weber, A. P. M (2014). Comparative transcriptome atlases reveal altered gene expression modules between two Cleomaceae C3 and C4 plant species. *The Plant Cell*, *26*(8), 3243–3260.

Ku, M. S., Kano-Murakami, Y., & Matsuoka, M. (1996). Evolution and expression of C4 photosynthesis genes. *Plant Physiology*, *111*(4), 949–957.

Lal, A., Ku, M. S., & Edwards, G. E. (1996). Analysis of inhibition of photosynthesis due to water stress in the C3 species Hordeum vulgare and Vicia faba: Electron transport, $CO_2$ fixation and carboxylation capacity. *Photosynthesis Research*, *49*(1), 57–69.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359.

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, *15*(2), R29.

Leegood, R. C. (2002). C4 photosynthesis: principles of CO2 concentration and prospects for its introduction into C3 plants. *Journal of Experimental Botany*, *53*(369), 581–590.

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., & Storey, J. D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* , *28*(6), 882–883.

Leinonen, R., Sugawara, H., Shumway, M., & International Nucleotide Sequence Database Collaboration. (2011). The sequence read archive. *Nucleic Acids Research*, *39*(Database issue), D19–D21.

Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics,* 12:323

Lu, A., Guindon, S., Performance of standard and stochastic branch-site models for detecting positive selection among coding sequences. *Molecular Biology and Evolution*. *31*(2), 484-495.

Ludwig, M. (2012). Carbonic anhydrase and the molecular evolution of C4 photosynthesis. *Plant, Cell & Environment*, *35*(1), 22–37.

Ludwig, M. (2016). The Roles of Organic Acids in C4 Photosynthesis. *Frontiers in Plant Science*, *7*:647.

Luo, C., Tsementzi, D., Kyrpides, N., Read, T., & Konstantinidis, K. T. (2012). Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. *PloS One*, *7*(2), e30087.

Lynch, M., & Force, A. (2000). The probability of duplicate gene preservation by subfunctionalization. *Genetics*, *154*(1), 459–473.

Magallón, S., Gómez-Acevedo, S., Sánchez-Reyes, L. L., & Hernández-Hernández, T. (2015). A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. *The New Phytologist*, *207*(2), 437–453.

Makino, A., Mae, T., & Ohira, K. (1985). Photosynthesis and ribulose-1,5-bisphosphate carboxylase/oxygenase in rice leaves from emergence through senescence. Quantitative analysis by carboxylation/oxygenation and regeneration of ribulose 1,5-bisphosphate. *Planta*, *166*(3), 414–420.

Mano, S., & Innan, H. (2008). The evolutionary rate of duplicated genes under concerted evolution. *Genetics*. *180*(1), 493-505.

Mansai, S. P., & Innan, H. (2010). The power of the methods for detecting interlocus gene conversion. *Genetics*. *184*(2), 517-527.

Manzara, T., Carrasco, P., & Gruissem, W. (1991). Developmental and organ-specific changes in promoter DNA-protein interactions in the tomato rbcS gene family. *The Plant Cell*, *3*(12), 1305–1316.

Marin-Navarro, J., & Moreno, J. (2006). Cysteines 449 and 459 modulate the reduction-oxidation conformational changes of ribulose 1.5-bisphosphate carboxylase/oxygenase and the translocation of the enzyme to membranes during stress. *Plant, Cell & Environment*, *29*(5), 898–908.

Martin, D. P., Murrell, B., Golden, M., Khoosal, A., & Muhire, B. (2015). RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evolution*, *1*(1), 1.

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*(1), 75–86.

Matsuoka, M. (1995). The gene for pyruvate, orthophosphate dikinase in C4 plants: structure, regulation and evolution. *Plant & Cell Physiology*, *36*(6), 937–943.

Mcglothlin, J. W., Kobiela, M. E., Feldman, C. R., Castoe, T. A., Geffeney, S. L., Hanifin, C. T., Toledo, G., Vonk, F. J., Richardson, M. K., Brodie Jr., E. D., Pfrender, M. E., Brodie III, E. D.

(2016). Historical Contingency in a Multigene Family Facilitates Adaptive Evolution of Toxin Resistance Report. *Current Biology: 26*(12), 1616-1621.

Mckown, A. D., & Dengler, N. G. (2007). Key innovations in the evolution of Kranz anatomy and C4 vein pattern in Flaveria (Asteraceae). *American Journal of Botany*. *94*(3), 382-399.

Miller, D. D., Ota, Y., Sumaila, U. R., Cisneros-Montemayor, A. M., & Cheung, W. W. L. (2018). Adaptation strategies to climate change in marine systems. *Global Change Biology*, *24*(1), e1–e14.

Moore, R. C., & Purugganan, M. D. (2005). The evolutionary dynamics of plant duplicate genes. *Current Opinion in Plant Biology*. *8*(2), 122-128.

Morita, K., Hatanaka, T., Misoo, S., & Fukayama, H. (2014). Unusual small subunit that is not expressed in photosynthetic cells alters the catalytic properties of rubisco in rice. *Plant Physiology*, *164*(1), 69–79.

Morita, K., Hatanaka, T., Misoo, S., & Fukayama, H. (2016). Identification and expression analysis of non-photosynthetic Rubisco small subunit, *OsRbcS1*-like genes in plants. *Plant Gene*, *8*, 26–31.

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*(7), e1002764.

Muse, S. V., Gaut, B. S., & Carolina, N. (1994). A Likelihood Approach for Comparing Synonymous and Nonsynonymous Nucleotide Substitution Rates with Application to the Chloroplast Genome, *11*(5), 715-724.

Nei, M., Gu, X., & Sitnikova, T. (1997). Evolution by the birth-and-death process in multigene families of the vertebrate immune system, *94*, 7799–7806.

Nei, M., & Rooney, A. P. (2005). Concerted and Birth-and-Death Evolution of Multigene Families. *Annual Review of Genetics. 39*, 121-152.

Niimura, Y. (2009). Evolutionary dynamics of olfactory receptor genes in chordates: interaction between environments and genomic contents. *Human Genomics*, *4*(2), 107–118.

Miller. R. (2014). Evolution of the *rbcS* gene family in Solanaceae: concerted evolution and gain and loss of introns, with a description of new statistical guidelines for determining the

number of unique gene copies. (Doctoral dissertation) Retrieved from University of Washington.

Ohta, T. (1977). Genetic variation in multigene families. *Nature*, *267*(5611), 515–517.

Ohta, T. (1979). An Extension of a Model for the Evolution of Multigene Families by Unequal Crossing over. *Genetics*, *91*(3), 591–607.

Ohta, T. (1980). Evolution and variation of multigene families. Lecture notes in biomathematics. Vol. 37. Springer-Verlag.

Ohta, T. (1983). On the evolution of multigene families. *Theoretical Population Biology*, *23*(2), 216–240.

Ohta, T. (1988). Time for acquiring a new gene by duplication. *Proceedings of the National Academy of Sciences of the United States of America*, *85*(10), 3509–3512.

Ohta, T. (1991). Multigene families and the evolution of complexity. *Journal of Molecular Evolution*, *33*(1), 34–41.

O'Neal, J. K., Pokalsky, A. R., Kiehne, K. L., & Shewmaker, C. K. (1987). Isolation of tobacco SSU genes: characterization of a transcriptionally active pseudogene. *Nucleic Acids Research*, *15*(21), 8661–8677.

Panchy, N., Lehti-Shiu, M. D., & Shiu, S.-H. (2016). Evolution of gene duplication in plants. *Plant Physiology*, *171*(4), 2294-2316.

Papp, B., Pál, C., & Hurst, L. D. (2003). Dosage sensitivity and the evolution of gene families in yeast. *Nature*, *424*(6945), 194–197.

Parry, M. A. J. (2002). Rubisco Activity: Effects of Drought Stress. *Annals of Botany*, *89*(7), 833–839.

Parry, M. A. J., Andralojc, P. J., Mitchell, R. A. C., Madgwick, P. J., & Keys, A. J. (2003). Manipulation of Rubisco: the amount, activity, function and regulation. *Journal of Experimental Botany*, *54*(386), 1321–1333.

Paterson, A. H., Bowers, J. E., & Chapman, B. A. (2004). Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proceedings of the National Academy of Sciences*, *101*(26), 9903–9908.

Paterson, A. H., Bowers, J. E., Van de Peer, Y., & Vandepoele, K. (2005). Ancient duplication of cereal genomes. *The New Phytologist*, *165*(3), 658–661.

Paterson, A. H., Chapman, B. A., Kissinger, J. C., Bowers, J. E., Feltus, F. A., & Estill, J. C. (2006). Many gene and domain families have convergent fates following independent whole-genome duplication events in Arabidopsis, Oryza, Saccharomyces and Tetraodon. *Trends in Genetics: TIG*, *22*(11), 597–602.

Pearson, P. N., Foster, G. L., & Wade, B. S. (2009). Atmospheric carbon dioxide through the Eocene–Oligocene climate transition. *Nature*, *461*, 1110-1113.

Peterhansel, C., Horst, I., Niessen, M., Blume, C., Kebeish, R., Kürkcüoglu, S., & Kreuzaler, F. (2010). Photorespiration. *The Arabidopsis Book / American Society of Plant Biologists*, *8*, e0130.

Petter, M., Bonow, I., & Klinkert, M. Q. (2008). Diverse expression patterns of subgroups of the rif multigene family during Plasmodium falciparum gametocytogenesis. *PloS One*, *3*(11), e3779.

Pichersky, E., & Cashmore, A. R. (1986). Evidence for selection as a mechanism in the concerted evolution of Lycopersicon esculentum ( tomato ) genes encoding the small subunit of ribulose-1 , 5-bisphosphate carboxylase / oxygenase, *83*(June), 3880–3884.

Plata, G., & Vitkup, D. (2014). Genetic robustness and functional evolution of gene duplicates. *Nucleic Acids Research*, *42*(4), 2405–2414.

Pond, S. L. K., Frost, S. D. W., & Muse, S. V. (2005). HyPhy: hypothesis testing using phylogenies. *Bioinformatics* , *21*(5), 676–679.

Quick, W. P., Schurr, U., Scheibe, R., Schulze, E. D., Rodermel, S. R., Bogorad, L., & Stitt, M. (1991). Decreased ribulose-1,5-bisphosphate carboxylase-oxygenase in transgenic tobacco transformed with “antisense” *rbcS* : I. Impact on photosynthesis in ambient growth conditions. *Planta*, *183*(4), 542–554.

Rawsthorne, S. (1992). C3–C4 intermediate photosynthesis: linking physiology to gene expression. *The Plant Journal*. https://doi.org/10.1111/j.1365-313X.1992.00267.x

Rensing, S. a. (2014). Gene duplication as a driver of plant morphogenetic evolution. *Current Opinion in Plant Biology*, *17*, 43–48.

Revell, L. J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution / British Ecological Society*, *3*(2), 217–223.

Rice, A. M., & McLysaght, A. (2017). Dosage-sensitive genes in evolution and disease. *BMC Biology*, *15*(1), 78.

Roberts, A., Pimentel, H., Trapnell, C., & Pachter, L. (2011). Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* , *27*(17), 2325–2329.

Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L., & Pachter, L. (2011). Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biology*, *12*(3), R22.

Robert Tabita, F. (1999). Microbial ribulose 1,5-bisphosphate carboxylase/oxygenase: A different perspective. *Photosynthesis Research*, *60*(1), 1–28.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* , *26*(1), 139–140.

Roulin, A., Auer, P. L., Libault, M., Schlueter, J., Farmer, A., May, G., Stacey, G., Doerge, R.W., Jackson, S. A. (2012). The fate of duplicated genes in a polyploid plant genome. *The Plant Journal: For Cell and Molecular Biology*, 143–153.

Sage, R. F. (2002). Variation in the k(cat) of Rubisco in C(3) and C(4) plants and some implications for photosynthetic performance at high and low temperature. *Journal of Experimental Botany*, *53*(369), 609–620.

Sage, R. F. (2004). The evolution of C4 photosynthesis. *The New Phytologist*. 161(2), 341-371.

Sage, R. F., Christin, P. A., & Edwards, E. J. (2011). The C 4 plant lineages of planet Earth. *Journal of Experimental Botany*. *62*(9), 3155-3169.

Sage, R. F., & Coleman, J. R. (2001). Effects of low atmospheric CO(2) on plants: more than a thing of the past. *Trends in Plant Science*, *6*(1), 18–24.

Sánchen-Ken, J. G., Gabriel Sánchen-Ken, J., Clark, L. G., Kellogg, E. A., & Kay, E. E. (2007). Reinstatement and Emendation of Subfamily Micrairoideae (Poaceae). *Systematic Botany*, *32*(1), 71–80.

Sasanuma, T. (2001). Characterization of the rbcS multigene family in wheat: subfamily classification, determination of chromosomal location and evolutionary analysis. *Molecular Genetics and Genomics: MGG*, *265*(1), 161–171.

Schaffner, A. R., & Sheen, J. (1992). Maize C4 photosynthesis involves differential regulation of phosphoenolpyruvate carboxylase genes. *The Plant Journal: For Cell and Molecular Biology*, *2*(2), 221–232.

Schrödinger, LLC. (2015, November). The PyMOL Molecular Graphics System, Version 1.8.

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., & Serrano, L. (2005). The FoldX web server: an online force field. *Nucleic Acids Research*, *33*(Web Server issue), W382–W388.

Seemann, J. R., Badger, M. R., & Berry, J. A. (1984). Variations in the Specific Activity of Ribulose-1,5-bisphosphate Carboxylase between Species Utilizing Differing Photosynthetic Pathways. *Plant Physiology*, *74*(4), 791–794.

Sela, I., Ashkenazy, H., Katoh, K., & Pupko, T. (2015). GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic Acids Research*, *43*(W1), W7–W14.

Sen, L., Fares, M. a., Liang, B., Gao, L., Wang, B., Wang, T., & Su, Y.-J. (2011). Molecular evolution of rbcL in three gymnosperm families: identifying adaptive and coevolutionary patterns. *Biology Direct*, *6*(1), 29.

Simpson, G. G. (1953). *The Major Features of Evolution*.

Slater, G. S. C., & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, *6*, 31.

Song, G., Riemer, C., Dickins, B., Kim, H. L., Zhang, L., Zhang, Y., Hsu, C.-H., Hardison, R. C. Nisc Comparative Sequencing Program, Green, E. D. Miller, W. (2012). Revealing mammalian evolutionary relationships by comparative analysis of gene clusters. *Genome Biology and Evolution*, *4*(4), 586–601.

Spreitzer, R. J. (2003). Role of the small subunit in ribulose-1, 5-bisphosphate carboxylase/oxygenase. *Archives of Biochemistry and Biophysics. 414*(2), 141-149.

Spreitzer, R. J., & Salvucci, M. E. (2002). RUBISCO: Structure, Regulatory Interactions, and Possibilities for a Better Enzyme. *Annual Review of Plant Biology*, *53*, 449–475.

Stitt, M., Quick, W. P., Schurr, U., Schulze, E. D., Rodermel, S. R., & Bogorad, L. (1991). Decreased ribulose-1,5-bisphosphate carboxylase-oxygenase in transgenic tobacco transformed with “antisense” *rbcS* : II. Flux-control coefficients for photosynthesis in varying light, CO2, and air humidity. *Planta*, *183*(4), 555–566.

Studer, R. a., Christin, P.-A., Williams, M. a., & Orengo, C. a. (2014). Stability-activity tradeoffs constrain the adaptive evolution of RubisCO. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(6), 2223–2228.

Studer, R. a., Penel, S., Duret, L., & Robinson-Rechavi, M. (2008). Pervasive positive selection on duplicated and nonduplicated vertebrate protein coding genes. *Genome Research*, *18*(9), 1393–1402.

Sugita, M., Manzara, T., Pichersky, E., Cashmore, A., & Gruissem, W. (1987). Genomic organization, sequence analysis and expression of all five genes encoding the small subunit of ribulose-l,5-bisphosphate carboxylase/oxygenase from tomato. *Molecular & General Genetics: MGG*, *209*, 247–256.

Suyama, M., Torrents, D., & Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, *34*(Web Server issue), W609–W612.

Suzuki, Y., Kihara-Doi, T., Kawazu, T., Miyake, C., & Makino, A. (2010). Differences in Rubisco content and its synthesis in leaves at different positions in Eucalyptus globulus seedlings. *Plant, Cell & Environment*, *33*(8), 1314–1323.

Suzuki, Y., & Makino, A. (2012). Availability of Rubisco small subunit up-regulates the transcript levels of large subunit for stoichiometric assembly of its holoenzyme in rice. *Plant Physiology*, *160*(1), 533–540.

Szöllősi, G. J., Davín, A. A., Tannier, E., Daubin, V., & Boussau, B. (2015). Genome-scale phylogenetic analysis finds extensive gene transfer among fungi. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *370*(1678), 20140335.

Tanz, S. K., Tetu, S. G., Vella, N. G. F., & Ludwig, M. (2009). Loss of the transit peptide and an increase in gene expression of an ancestral chloroplastic carbonic anhydrase were instrumental in the evolution of the cytosolic C4 carbonic anhydrase in *Flaveria*. *Plant Physiology*, *150*(3), 1515–1529.

Tcherkez, G. G. B., Farquhar, G. D., Andrews, T. J., & Lorimer, G. H. (2006). Despite slow catalysis and confused substrate specificity, all ribulose bisphosphate carboxylases may be nearly perfectly optimized. *103*(19), 7246-7251.

Tezara, W., Mitchell, V., Driscoll, S. P., & Lawlor, D. W. (2002). Effects of water deficit and its interaction with CO(2) supply on the biochemistry and physiology of photosynthesis in sunflower. *Journal of Experimental Botany*, *53*(375), 1781–1791.

Thomas-Hall, S., Campbell, P. R., Carlens, K., Kawanishi, E., Swennen, R., Sági, L., & Schenk, P. M. (2007). Phylogenetic and molecular analysis of the ribulose-1,5-bisphosphate carboxylase small subunit gene family in banana. *Journal of Experimental Botany*, *58*(10), 2685–2697.

Tierney, L. (2012). The R Statistical Computing Environment. *Lecture Notes in Statistics*. 435–447.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, *7*(3), 562-578.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, *28*(5), 511–515.

Uecker, H., & Hermisson, J. (2011). On the fixation process of a beneficial mutation in a variable environment. *Genetics*, *188*(4), 915–930.

UniProt Consortium. (2015). UniProt: a hub for protein information. *Nucleic Acids Research*, *43*(Database issue), D204–D212.

van Lun, M., Hub, J. S., van der Spoel, D., & Andersson, I. (2014). $CO_2$ and $O_2$ distribution in Rubisco suggests the small subunit functions as a $CO_2$ reservoir. *Journal of the American Chemical Society*, *136*(8), 3165–3171.

Vicentini, A., Barber, J. C., & Aliscioni, S. S. (2008). The age of the grasses and clusters of origins of C4 photosynthesis. *Global Change Biology*. *14,* 2963-2977.

Von Caemmerer, S., Millgate, A., Farquhar, G. D., & Furbank, R. T. (1997). Reduction of Ribulose-1,5-Bisphosphate Carboxylase/Oxygenase by Antisense RNA in the C4 Plant Flaveria bidentis Leads to Reduced Assimilation Rates and Increased Carbon Isotope Discrimination. *Plant Physiology*, *113*(2), 469–477.

von Caemmerer, S., & Paul Quick, W. (2000). Rubisco: Physiology in Vivo. *Advances in Photosynthesis and Respiration*. 85–113.

Vu, J. C. V., Gesch, R. W., Hartwell Allen, L., Boote, K. J., & Bowes, G. (1999). C02 Enrichment Delays a Rapid, Drought-Induced Decrease in Rubisco Small Subunit Transcript Abundance. *Journal of Plant Physiology*, *155*(1), 139–142.

Wagner, A. (1998). The fate of duplicated genes: loss or new function? *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, *20*(10), 785–788.

Wagner, A. (2005). Robustness, evolvability, and neutrality. *FEBS Letters*, *579*(8), 1772–1778.

Wang, X., Shi, X., Hao, B., Ge, S., & Luo, J. (2005). Duplication and DNA segmental loss in the rice genome: implications for diploidization. *The New Phytologist*, *165*(3), 937–946.

Wanner ', L. A., & Gruissem2, W. (1991). Expression Dynamics of the Tomato rbcS Gene Family during Development. *The Plant Cell*, *3*, 1289–1303.

Washburn, J. D., Schnable, J. C., Conant, G. C., Brutnell, T. P., Shao, Y., Zhang, Y., Ludwig, M., Davidse, G., Pires J. C. (2017). Genome-Guided Phylo-Transcriptomic Methods and the Nuclear Phylogentic Tree of the Paniceae Grasses. *Scientific Reports*, *7*(1), 13528.

Watson, G. M. F., & Tabita, F. R. (2006). Microbial ribulose 1,5-bisphosphate carboxylase/oxygenase: a molecule for phylogenetic and enzymological investigation. *FEMS Microbiology Letters*, *146*(1), 13–22.

Wessinger, M. E., Edwards, G. E., & Ku, M. S. B. (1989). Quantity and Kinetic Properties of Ribulose 1,5-Bisphosphate Carboxylase in C3, C4, and C3-C4 Intermediate Species of Flaveria (Asteraceae). *Plant & Cell Physiology*, *30*(5), 665–671.

Wickham, H. (2009). ggplot2: elegant graphics for data analysis. *Springer New York*, *1*(2), 3.

Wiens, J. J., & Morrill, M. C. (2011). Missing data in phylogenetic analysis: reconciling results from simulations and empirical data. *Systematic Biology*, *60*(5), 719–731.

Wolter, F. P., Fritz, C. C., Willmitzer, L., Schell, J., & Schreier, P. H. (1988). *rbcS* genes in Solanum tuberosum: Conservation of transit peptide and exon shuffling during evolution (plant gene family/ribulose bisphosphate carboxylase/potato/intron). *Evolution; International Journal of Organic Evolution*, *85*, 846–850.

Yamori, W., & Von Caemmerer, S. (2009). Effect of Rubisco Activase Deficiency on the Temperature Response of $CO_2$ Assimilation Rate and Rubisco Activation State: Insights from Transgenic Tobacco with Reduced Amounts of Rubisco Activase. *151*(4), 2073-2082.

Yang, X., Coulombe-Huntington, J., Kang, S., Sheynkman, G. M., Hao, T., Richardson, A., Sun, S., Yang, F., Shen, Y. A., Murray, R. R., Spirohn, K., Begg, B. E., Duran-Frigola, M., MacWilliams, A., Pevzner, S. J., Zhong, Q., Trigg, S. A., Tam, S., Ghamsari, L., Sahni, N., Yi, S., Rodriguez, M. D., Balcha, D., Tan, G., Costanzo, M., Andrews, B., Boone, C., Zhou, X. J., Salehi-Ashtiani, K., Charloteaux, B., Chen, A. A., Calderwood, M. A., Aloy, P., Roth, F. P., Hill, D. E., Iakoucheva, L. M., Xia, Y., Vidal, M. (2016). Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. *Cell*, *164*(4), 805–817.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, *24*(8), 1586–1591.

Yeoh, H.-H., Badger, M. R., & Watson, L. (1980). Variations in Km($CO_2$) of Ribulose-1,5-bisphosphate Carboxylase among Grasses. *Plant Physiology*, *66*(6), 1110–1112.

Yeoh, H. H., Badger, M. R., & Watson, L. (1981). Variations in Kinetic Properties of Ribulose-1,5-bisphosphate Carboxylases among Plants. *Plant Physiology*, *67*(6), 1151–1155.

Yoon, Y., Lee, Y., Kim, T., Ahn, J. S., Jung, Y., Kim, B., & Lee, S. (2001). High resolution resonance enhanced two photon ionization spectroscopy of *RbCs* in a cold molecular beam. *The Journal of Chemical Physics*, *114*(20), 8926–8931.

Yu, J., Wang, J., Lin, W., Li, S., Li, H., Zhou, J., Dong, W., Hu, S., Zeng, C., Zhang, J., Zhang, Y., Li, R., Xu, Z., Li, S., Li, X., Zheng, H., Cong, L., Lin, L., Yin, J., Geng, J., Li, G., Shi, J., Liu, J., Lv, H., Li, J., Wang, J., Deng, Y., Ran, L., Shi, X., Wang, X., Wu, Q., Li, C., Ren, X., Wang, J., Wang, X., Li, D., Liu, D., Zhang, X., Ji, Z., Zhao, W., Sun, Y., Zhang, Z., Bao, J., Han, Y., Dong, L., Ji, J., Chen, P., Wu, S., Liu, J., Xiao, Y., Bu, D., Tan, J., Yang, L., Ye, C., Zhang, J., Xu, J., Zhou, Y., Yu, Y., Zhang, B., Zhuang, S., Wei, H., Liu, B., Lei, M., Yu, H., Li, Y., Xu, H., Wei, S., He, X., Fang, L., Zhang, Z., Zhang, Y., Huang, X., Su, Z., Tong, W., Li, J., Tong, Z., Li, S., Ye, J., Wang, L., Fang, L., Chen, C., Chen, H., Xu, Z., Li, H., Huang, H., Zhang, F., Xu, H., Li, N., Zhao, C., Li, S., Dong, L., Huang, Y., Li, L., Xi, Y., Qi, Q., Li, W., Zhang, B., Hu, W., Zhang, Y., Tian, X., Jiao, Y., Liang, X., Jin, J., Gao, L., Zheng, W., Hao, B., Liu, S., Wang, W., Yuan, L., Cao, M., McDermott, J., Samudrala, R., Wang, J,. Wong, G. K., Yang, H. (2005). The Genomes of *Oryza sativa*: A History of Duplications. *PLoS Biology*, *3*(2), e38.

Zhang, J. (2003). Evolution by gene duplication: an update. *Trends in Ecology & Evolution*, *18*(6), 292–298.

Zhang, J., Zhang, Y. P., & Rosenberg, H. F. (2002). Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. *Nature Genetics*, *30*(4), 411–415.

Zhang, L., Zhang, L., Sun, J., Zhang, Z., Ren, H., & Sui, X. (2013). Rubisco gene expression and photosynthetic characteristics of cucumber seedlings in response to water deficit. *Scientia Horticulturae*. *161*, 81-87.

Zhong, Y., Jia, Y., Gao, Y., Tian, D., Yang, S., & Zhang, X. (2013). Functional requirements driving the gene duplication in 12 *Drosophila* species. *BMC Genomics*, *14*(1), 555.

Zmieńko, A., Samelak, A., Kozłowski, P., & Figlerowicz, M. (2014). Copy number polymorphism in plant genomes. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, *127*(1), 1–18.

Zuo, T., Zhang, J., Lithio, A., Dash, S., Weber, D. F., Wise, R., Nettleton, D., Peterson, T. (2016). Genes and Small RNA Transcripts Exhibit Dosage-Dependent Expression Pattern in Maize Copy-Number Alterations. *Genetics*, *203*(3), 1133–1147.

# Contributions

## Chapter 1

Initially, I ran all the analyses including the reconstruction of the phylogenetic trees, positive selection, coevolution and homology modelling by myself. One of the difficulties of this project was the number of the *rbcS* gene copies changes according to the updated genome assembly to the public database. Each time the assembly is updated, I should have run each analysis for multiple times using most recently updated public data. It was especially required for the publications.

Following collaborator helped me to keep my analyses to up to date for the publication. Victor Rossier helped to run Exonerate. Romain Studer taught me the method of homology modelling. He re-ran all the homology modelling analysis by FoldX 4.0 because the technical support of FoldX3.0 has finished when I wanted to include the newly annotated gene copies for the analysis. Iakov Davydov used Hyphy to re-run the positive selection analysis that had already run by myself using CodeML.

## Chapter 2

The extracted DNA of Poaceae was provided by Guillaume Besnard and Pascal-Antoine Christin. I calculated the estimated cost for 454 sequencing. I designed the primers of the *rbcS* referring to the advice of Guillaume and Pascal-Antoine. I performed the wet-lab experiments including PCR, purification, the preparation for the 454 sequencing. The 454 sequencing has run by Microsynth AG. I created the new pipeline for sorting, clustering and assembly of the reads. I did the alignment and reconstruction of the phylogenetic tree. Iakov Davydov is a developer of the software to test positive selection, called Godon. I followed his instruction to run positive selection analysis by Godon. Lab technicians, Dessislava Savova Bianchi, and Catherine Berney gave me some advice when I was working in wet-lab.

## Chapter 3

I discussed the design of transcriptome analyses with Andrea Komljenovic. I ran all the transcriptome analyses by myself.

## In the end

It was not a simple way to analyze the *rbcS* gene. From sequencing to clustering of the reads, there were always some problems to solve. I needed to have patience. The project took a longer time than I expected. However, it was a great pleasure to work on the evolution of the interesting *rbcS* gene encoding part of the most abundant protein on Earth.

Thank you *rbcS*. ☺