Faculty of Biological and Environmental Sciences
University of Helsinki
Finland

EVOLUTION OF THE DUF26-CONTAINING PROTEINS IN PLANTS

Aleksia Vaattovaara

ACADEMIC DISSERTATION

To be presented, with the permission of the Faculty of
Biological and Environmental Sciences of
the University of Helsinki, for public examination in auditorium 2041,
Biocenter 2 (Viikinkaari 5), on 28th June 2019, at 12 o'clock noon.

Helsinki 2019

Supervisors	Dr. Michael Wrzaczek University of Helsinki, Finland Asst. Prof. Jarkko Salojärvi Nanyang Technological University, Singapore University of Helsinki, Finland
Thesis committee	Dr. Saijaliisa Kangasjärvi University of Turku, Finland Prof. Jaakko Hyvönen (from September 2015) University of Helsinki, Finland Dr. Sarah Coleman (until September 2015) University of Helsinki, Finland
Reviewers	Prof. Hely Häggman University of Oulu, Finland Dr. Tanja Pyhäjärvi University of Oulu, Finland
Opponent	Prof. Shin-Han Shiu Michigan State University, USA
Custos	Prof. Jaakko Hyvönen University of Helsinki, Finland

ISBN 978-951-51-5302-9 (pbk.)
ISBN 978-951-51-5303-6 (PDF)

Unigrafia Oy
Helsinki 2019

ABSTRACT

For plants as sessile organisms effective signaling mechanisms are essential. Plants utilize signaling networks to receive cues from the environment and signal between cells. Various proteins and protein families are involved in the signaling networks in plants including receptor-like kinases (RLKs) and their related receptor-like proteins (RLPs). RLKs are typically located in the plasma membrane and transfer signals from the apoplastic space to the interior of the cell. The domain of unknown function 26 (DUF26) is a cysteine-rich protein domain involved in signaling. DUF26-containing proteins are a plant-specific protein family containing both RLKs and RLPs, including cysteine-rich receptor-like kinases (CRKs), plasmodesmata-localized proteins (PDLPs) and cysteine-rich receptor-like secreted proteins (CRRSPs).

To facilitate investigation of the functions of DUF26 proteins, comprehensive phylogenetic and evolutionary analyses were combined with broad phenotypic analyses of *crk* mutants and structural investigation of two PDLPs from the model species *Arabidopsis thaliana*. These analyses revealed that DUF26-containing genes have a complex evolutionary history, including several steps of domain rearrangements and differential expansion and contraction patterns in different groups of plants and between different groups of CRKs, PDLPs and CRRSPs. CRKs were found to be involved in stress responses and development based on their loss-of-function phenotypes. The crystal structure of the AtPDLPs revealed a close structural homology between the DUF26 domain and fungal lectins, suggesting that DUF26 could be a carbohydrate-binding unit in plants.

Annotation quality is crucial for virtually any type of sequence-based analysis, including phylogenetic estimation of relationships between genes, proteins and species. For this reason, the annotations of DUF26-containing genes were carefully curated in such a way as to facilitate the subsequent evolutionary analyses. Since most functional data is obtained from model species, only through thorough estimation of the relationships between proteins from different species we can reliably transfer information among species. In the future, as more functional information becomes available, the knowledge gained from this study will be applied in translational research between model species and crop species.

ACKNOWLEDGEMENTS

First of all, I want to thank my supervisors Michael Wrzaczek and Jarkko Salojärvi. You have been an amazing support during these years. Specifically, I wish to thank Michi for giving me the opportunity to study this interesting topic. I also appreciate enormously your efforts and enthusiasm as a supervisor, you always had time for discussions and encouragement. I am also thankful to Jarkko for always finding time to guide and help me during these years.

I wish to thank my follow-up group members Saijaliisa Kangasjärvi, Jaakko Hyvönen and Sarah Coleman. It was always a great pleasure to meet with you. I also thank Jaakko Hyvönen for agreeing to act as a custos in my defense. For very efficient pre-examination and helpful comments, I want to thank Hely Häggman and Tanja Pyhäjärvi. For funding, I want to thank DPPS and Finnish Cultural Foundation. I also express my gratitude to Karen Sims-Huopaniemi, for all help and advices during these years.

I am grateful to all collaborators and people that co-authored publications with me. It has been great to work with you. Especially thanks to Ari Löytynoja for the substantial help with phylogenetic analyses and all advices. Also thanks to Johanna Leppälä for all our conversations and support. I want to thank all current and former members in Rec-Sig group, BMBG group, Ox-stress me(t/g)a group led by Jaakko Kangasjärvi, CoE and ViPS. It has been wonderful to be part of these skillful communities and I have learned a lot from you. Especially I thank Kirk and Mikael for all discussions over lunch. I thank Matt for proof-reading my thesis. I want to thank people that shared office with me (Julia, Sitaram, Twinkle, David, Ali, Jarkko and Sachie), you have been great friends during these years. I also thank other PhD students belonging to DPPS for company and support.

I want to thank Outi Savolainen for time in her group and inspiration to continue to PhD studies. I thank Esa Aalto for guiding me to bioinformatics and Esa Borén and Esa Kekki for getting me so interested in biology in the first place.

Nästyspoppoo, my dear friends, I am very grateful of you all. I want to thank all my friends from stables, especially Titina, you and our common hobby are my delight and important balance for work.

To my family, there is no words to describe how grateful I am of everything. Kiitos. Jaakko, my dearest, I thank you from the bottom of my heart. You have been the greatest support all these years and this would not have been possible without you.

CONTENTS

LIST OF ORIGINAL PUBLICATIONS

This thesis is based on the following publications:

I **Vaattovaara A**, Brandt B, Rajaraman S, Safronov O, Veidenberg A, Luklová M, Kangasjärvi J, Löytynoja A, Hothorn M, Salojärvi J & Wrzaczek M, Mechanistic insights into the evolution of DUF26-containing proteins in land plants, 2019, Communications Biology, 2:56

II Bourdais G, Burdiak P, Gauthier A, Nitsch L, Salojärvi J, Rayapuram C, Idänheimo N, Hunter K, Kimura S, Merilo E, **Vaattovaara A**, Oracz K, Kaufholdt D, Pallon A, Anggoro D, Glow D, Lowe J, Zhou J, Mohammadi O, Puukko T, Albert A, Lang H, Ernst D, Kollist H, Brosche M, Durner J, Borst J W, Collinge D, Karpinski S, Lyngkjaer M, Robatzek S, Wrzaczek M & Kangasjärvi J, Large-scale phenomics identifies primary and fine-tuning roles for CRKs in responses related to oxidative stress, 2015, PLoS Genetics, 11(7): e1005373

III **Vaattovaara A**, Salojärvi J & Wrzaczek M, Extraction and curation of gene models for plant receptor kinases for phylogenetic analysis, 2017, Plant Receptor Kinases: Methods and Protocols, Aalen, R (ed.). New York: Springer (Methods in Molecular Biology; no. 1621), Pages 79-91

IV **Vaattovaara A**, Leppälä J, Salojärvi J & Wrzaczek M, High-throughput sequencing data and the impact of plant genome annotation quality, 2019, Journal of Experimental Botany, 70(4): 1069–1076

The publications are referred to in the text by their roman numerals.

ABBREVIATIONS

ATP	Adenosine triphosphate
CDS	Coding sequence
CRCK	Cysteine-rich receptor-like cytoplasmic kinase
CRK	Cysteine-rich receptor-like protein kinase
CRRSP	Cysteine-rich receptor-like secreted protein
DAMP	Damage-associated molecular pattern
DNA	Deoxyribonucleic acid
DUF	Domain of unknown function
EST	Expressed sequence tag
LRR	Leucine-rich repeat
MAMP	Microbe-associated molecular pattern
PRR	Pattern recognition receptor
PDLP	Plasmodesmata-localized protein
RLCK	Receptor-like cytoplasmic kinase
RLK	Receptor-like protein kinase
RLP	Receptor-like protein
RNA	Ribonucleic acid
SP	Signal peptide
SNP	Single nucleotide polymorphism
SSD	Small-scale duplication
TMR	Transmembrane region
UTR	Untranslated region
WGD	Whole genome duplication
WGM	Whole genome multiplication
At	*Arabidopsis thaliana*
bCRK	basal group CRK
dd	double domain
Hv	*Hordeum vulgare*
Os	*Oryza sativa*
sd	single domain
vCRK	variable group CRK
e.g.	exempli gratia
i.e.	id est

1 INTRODUCTION

The genetic information of organisms is encoded in their nuclear and organellar genomes. Genomes consist of non-coding and coding regions. Information in genomic DNA encodes different types of RNA molecules, like non-coding RNA (including transfer RNA and ribosomal RNA) and genes, which store protein-coding information. Eukaryotic genes are comprised of exons and introns which are flanked by regulatory sequences in 5' and 3' untranslated regions (UTRs). Exons contain the coding information for the amino acid sequence of the protein. Introns are spliced out during the process by which a gene is transcribed into messenger RNA (mRNA) and further translated into protein (Figure 1). Alternative splicing of exons and introns can produce different variants of proteins and provides flexibility that increases the adaptation potential in organisms (Lewin, 2008).

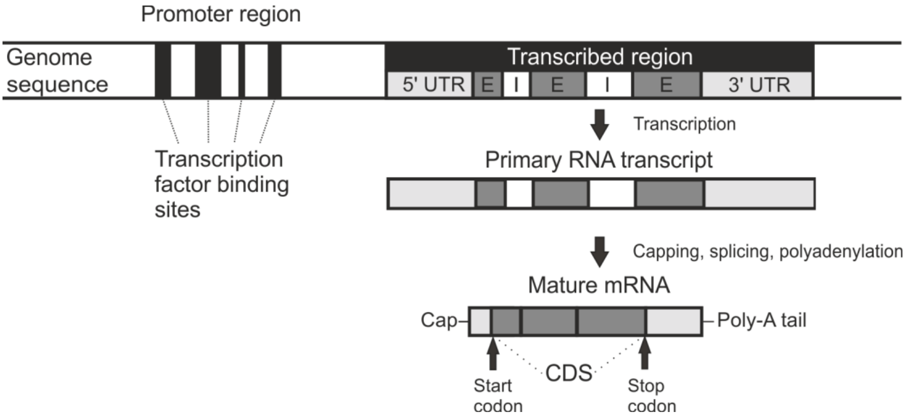

Figure 1 Eukaryotic gene and mRNA structure. The promoter region is located in 5' direction from the gene. The transcribed region contains 5' UTR, exons (E), introns (I) and 3'UTR. After transcription the primary mRNA is capped at the 5' end and a poly-A tail is added to the 3' end. Then the introns are spliced out. Start and stop codons define the coding region (CDS) which contains information on the amino acid sequence of the coded protein from exons.

1.1 SPECIFIC FEATURES OF PLANT GENOMES

Plants are sessile organisms and are therefore exposed continuously to external challenges to a different extent compared to mobile organisms, for example animals. This means that plants have to be able to adapt to the changes that occur at the location where they grow. Many features of plant genomes, like their large number of duplicated genes, repeats and transposable elements, and the modularity of protein domains, have been hypothesized to be linked to the need for rapid adaptation (Kersting *et al.*, 2012; Niu *et al.*, 2019). Polyploidy is also relatively common among plants. Approximately 31% of speciation events occured in ferns and 15% in angiosperms are estimated to be accompanied by an increase of the ploidy level (Wood *et al.*, 2009). When compared to the frequency of polyploidy today, ancient whole genome duplication (WGD) events are rare (Van de Peer *et al.*, 2009; Van de Peer *et al.*, 2017). In plants, two main paleopolyploidy events have taken place, one shared across angiosperms approximately 192 million years ago, and an older one which is common to all seed plants occuring about 319 million years ago (Jiao *et al.*, 2011). Over the course of evolution following the polyploidy event, polyploids evolve into diploids by a process called diploidization (Van de Peer *et al.*, 2009). Notably, plants can be autopolyploids, if polyploidy arises within a single species, or allopolyploids, if polypoidy results from hybridization between species. It is interesting and important to note that an individual plant species can also include both diploid and polyploid populations, for example *Arabidopsis arenosa* comprises populations of diploids and autopolyploids (Arnold *et al.*, 2015). Mixed-ploidy species provide an oppotunity to study the effects of polyploidization towards fitness and survival (Kolar *et al.*, 2017).

1.2 GENE FAMILIES AND THEIR EVOLUTION

Genes, and their products, proteins, can be categorized on the basis of various physiological and molecular functions, cellular localization, domains and expression patterns. Different systems have been created for the purpose of categorization like Gene Ontology (GO) for gene functions and subcellular locatization (Ashburner *et al.*, 2000; The Gene Ontology Consortium, 2019), as well as PFAM for defining protein domains (El-Gebali *et al.*, 2019; Sonnhammer *et al.*, 1997).

One method that has traditionally been used to group genes, is based on their ancestry and evolutionary relationships, which can be estimated according to their sequence similarity. Genes which share similar sequences and ancestry are often grouped into so-called gene families. In plants, the overwhelming majority of genes are members of gene families (Guo, 2013). Gene families consist of two or more genes which share the same ancestry. Gene families can be shared between species but they can also be specific to certain species or groups of species (Martinez, 2011). The members of multigene families are recognized based on the similarities in their sequences and consequently, they often have similar domain compositions. Gene families can frequently be clustered into the superfamilies that share one or more domains of single ancestry. The receptor-like kinases (RLKs) in plants (Shiu and Bleecker, 2001b; Shiu *et al.*, 2004) and the olfactory receptors (ORs) in vertebrates (Gaillard *et al.*, 2004; Olender *et al.*, 2008) are prominent examples of such superfamilies. However, it is important to note that the definition of a gene family is strongly context dependent.

1.2.1 BIRTH AND DEATH MECHANISMS

Over evolutionary timescales, gene families grow and shrink in size. These evolutionary events can take place in several different ways among different genomes. Small-scale duplications (SSDs) affect only small areas of a genome and emerge through various mechanisms (Panchy *et al.*, 2016). One or few genes can be locally duplicated by an unequal crossing-over event on a chromosome resulting in a tandem duplication. Duplicated gene pairs, that are not located in tandem repeats, can originate from transposon-mediated duplication or retroduplication where mRNA is reverse-transcribed into DNA and inserted into the genome. In addition, small parts of a genome can be copied into other parts of the same or a different chromosome. This kind of duplication is referred as segmental duplication. The precise mechanisms behind segmental duplications are not well understood in plants. Small-scale duplications can produce duplicates of full-length genes, but alternatively partial duplications can occur resulting in truncated genes. Such partial duplicates can produce non-functional proteins but occasionally they lead to the evolution of new genes with novel functions (Katju and Lynch, 2006). Partial duplicates can also recruit sequences from other genes or surrounding genomic sequences which can increase the potential for new functions (Katju and Lynch, 2006; Zhou *et al.*, 2008). Genes resulting from tandem duplications are more likely to diverge fast and be retained in the genome than the genes originating from multiplication of the whole genome (Qiao *et al.*, 2019).

Multiplications of whole genomes are far less common than SSDs; that take place frequently in the genome. The most common type of whole genome multiplication (WGM) is WGD which doubles the chromosomes and subsequently the gene number. WGM are often followed by rapid gene loss (Inoue *et al.*, 2015). Gene loss following WGM is called fractionation (Sankoff *et al.*, 2015). Genome rearrangement events may also increase after WGM, but rearrangement rates vary among species (Hufton and Panopoulou, 2009; Semon and Wolfe, 2007).

Closely related genes that are present in multiple species, but only with a single copy in each species, are called single-copy genes. Single-copy genes are under strong selection against copy number variation in the genome and usually duplicates are removed after WGM (De Smet *et al.*, 2013). This is likely related to their essential functions, e.g. maintaining genome stability and proper functions of organelles (Li *et al.*, 2016a). Proteins with conserved functions can also belong to protein interaction networks where changes in the copy-number of interacting proteins can cause stoichiometric imbalances of the protein complexes leading to deregulation of signaling mechanisms and thus be selected against. This phenomena is called dosage balance (Veitia, 2005). The outcome of the dosage balance depends on the duplication mode, SSDs usually duplicate only part of interaction genes and thus, the duplicates are removed. After WGD, all the members are duplicated in then the loss of one member is unadvantageous and the duplicates are retained in the genome for longer evolutionary times than other genes (Birchler and Veitia, 2012).

Duplicates from SSDs and WGMs that are kept in the genome can retain part of their ancestral function (subfunctionalization) or acquire novel functions (neofunctionalization). For example in fish, neofunctionalization has been detected in a receptor protein (Braasch *et al.*, 2006) and subfunctionalization in the transcription factor PAX6 (Kleinjan *et al.*, 2008), both duplicated in WGD. Sub- and neofunctionalization can both take place in the same duplicated gene, and subfunctionalization has been hypothesized to serve also as transition state to neofunctionalization (Rastogi and Liberles, 2005). Regulatory changes in gene expression can precede or indicate neo- and subfunctionalization. Gene expression data from *Arabidopsis thaliana* indicates that 30% of duplicates originating from WGD and 38% of tandemly duplicated genes are neo- or subfunctionalized based on divergence on their gene expression (Liu *et al.*, 2011). Duplicates can also turn into nonfunctional pseudogenes or can be lost with a fragment deleted from the genome.

Several different models of gene family evolution have been suggested. These models include divergent evolution, concerted evolution and the birth-death model (Eirin-Lopez *et al.*, 2012; Nei and Rooney, 2005). Divergent evolution is oldest model of the gene family evolution and was based on the evolutionary patterns observed from hemoglobin α, β, γ, and δ chains and myoglobin (Ingram, 1961). It assumes that genes diverge gradually after duplication and gain new functions. When the gene family of ribosomal RNA was found to contain tandemly duplicated genes that were more similar within species than between species the model of concerted evolution was proposed (Arnheim *et al.*, 1980; Coen *et al.*, 1982). In concerted evolution, duplicated genes are kept similar by gene conversion which involves non-reciprocal recombination of a segment of DNA from the donor gene to the recipient gene (Nei and Rooney, 2005). Neither convergent nor concerted evolution models were able to explain all the features of gene family evolution and they were followed by the birth-death model (Ota and Nei, 1994). The birth-death model represents the gene gain and loss events in the gene family and the hallmarking features include the presence of pseudogenes and interspecific gene clusters in phylogenetic tree (Eirin-Lopez *et al.*, 2012).

1.2.2 DOMAIN COMPOSITION CHANGES

Proteins can be considered to be operational units in a cell. However, many proteins can be further subdivided into protein domains. A protein domain is a part of the protein sequence that can be defined based on a conserved amino acid composition or structural region (Kelley and Sternberg, 2015). Proteins can consist of several domains, but there are also some proteins for which no domains have been defined to date. The domain composition of members of a single gene family is frequently similar, but domain swaps, gains and losses can occur, leading to the emergence of proteins with novel domain compositions (Moore *et al.*, 2008). These new domain compositions may produce proteins with novel physiological or biochemical functions. Partial duplications can result in loss of the domains. Partial duplications and genome rearrangements can also lead to exchanges of domains (domain swaps) and the acquisition of novel domains (domain gains). Domains within genes can also multiply by internal duplication, resulting in several copies of a domain within a single protein (Nacher *et al.*, 2010). The exact mechanism producing internal duplications is not known, but the exon shuffling by nonhomologous recombination is one possible cause (Björklund *et al.*, 2006).

Novel domains may appear as the consequence of accumulated mutations over evolutionary time. As mutations accumulate, the similarity between domains decreases and new domains may be defined. However, novel domains can also arise *de novo* from previously non-coding genome sequences (Klasberg *et al.*, 2018). This way of domain appearance is similar to the *de novo* appearance of novel genes. The exact mechanism behind the *de novo* origin of genes is not well understood, but it has been suggested that they evolve from non-coding sequences through a transitory protogene period (Carvunis *et al.*, 2012). DNA sequences coding novel protein domain usually integrate at the terminal ends of genes, and can subsequently migrate to different positions resulting in different domain compositions according to a study analyzing insect genomes (Klasberg *et al.*, 2018). In plants, the largest emergence of novel protein domains has taken place in the lineage leading to Embryophytes (Kersting *et al.*, 2012). This could correlate with anatomical and physiological adaptations which enabled plants to conquer land and may have allowed ancestral plants to conquer new habitats and adopt new survival strategies.

1.2.3 SELECTION AND GENETIC DRIFT

Selection affects gene families at the gene but also at the genome level. According to the neutral theory (Kimura, 1983), most of the genetic changes contributing to sequence divergence between and within species are neutral or slightly deleterious. The theory assumes that most of these neutral or nearly neutral mutations may be fixed or lost via random genetic drift. The deleterious mutations happen, but they are removed by natural selection. Thus, in genes coding proteins with conserved protein structure and function, a major selective mechanism is purifying selection which acts against changes that would negatively affect the structure or the function (Cooper and Brown, 2008). In contrast to this, positive directional selection is a rare but necessary step towards neo- or subfunctionalization and is critical for evolutionary change on the genetic, biochemical and physiological level. In particular, genes in tandem repeats are known to typically evolve first under positive selection before their fixation and conservation (Persi *et al.*, 2016). Based on theoretical model, the recently duplicated gene, excluding genes under dosage balance, produces a weak selective advantage and can be positively selected and thus retained in the genome (Rodrigo and Fares, 2018). The copy number variation of the gene family, via both gain and loss of the genes, can increase the adaptive potential of the species and be therefore positively selected (Katju and Bergthorsson, 2013). Similar to the point mutations with neutral effect, also copy number variation of the gene family can have so mild effect that the fixation of copies in the genome is dependent on the random genetic drift (Moore and Purugganan, 2003).

1.2.4 PHYLOGENETIC RELATIONSHIPS

The evolutionary relationships between different members of gene families or genes sharing common domains can be studied using phylogenetic approaches. Phylogenetic estimation is based on the similarities and differences between the subjects compared. In modern molecular biology the relationships between genes, individuals or species is frequently defined based on sequences or variable sequence based genetic markers like single nucleotide polymorphisms (SNPs) or restriction sites. In the case of genes this estimation can be done using either DNA or amino acid sequences.

Phylogenetic trees illustrate evolutionary distances between species, which can be estimated in different ways. For genes and proteins the distances are estimated from the sequence alignments. The different methods for constructing phylogenetic trees utilize distance matrices, parsimony, maximum-likelihood and Bayesian approaches (Yang and Rannala, 2012). Phylogeny construction for large datasets is a so-called NP-hard problem; meaning that the computation of the correct tree is close to impossible with the current means. Thus, a phylogenetic tree representing the evolutionary relationships across a large dataset is an estimate. The reliability of this estimate can be evaluated using different methods, including testing the reproducibility of the tree by bootstrapping (Felsenstein, 1985) or using a different algorithm to estimate the tree. Estimating trees using different subsets of the data can also shed light on the reliability of a particular phylogenetic tree.

Once a phylogenetic tree for a gene family is constructed, the relationships between gene family members can be interpreted. Comparison of a phylogeny of the species, which were included in the phylogenetic gene family tree, and the phylogeny of gene family members, will help to infer the origin of the genes. Between species, gene family members can be defined as orthologs or paralogs based on their origin (Koonin, 2005). Orthologous genes derive from the same ancestral gene and have been separated to different species due to speciation. Orthologs often retain similar functions in different species (Jensen *et al.*, 2003). Recognizing orthologs is essential for functional studies where gene functions in one species are inferred and extrapolated based on findings in another species. The molecular functions of genes are mainly known in model species like *Arabidopsis thaliana* or rice (*Oryza sativa*) and the available data can be transferred to other species when common orthologs are known. Paralogs are genes originating from duplication events within species. They may also have similar functions, depending on the definition of function as physiological or biochemical: however, because the duplicates can sub- or neofunctionalize, the more time that has passed since a duplication event the more likely it is that the function of paralog has changed.

1.3 RECEPTOR-LIKE KINASES AND RELATED PROTEINS

Receptor kinases are a large gene family in plants. The genome of the model plant species *Arabidopsis thaliana* contains almost 950 kinases, close to 600 membrane-bound kinases (RLKs) and almost 400 soluble kinases (Zulawski *et al.*, 2014), and monocots also contain similar numbers (Dardick *et al.*, 2007). The existence of so many plant RLKs is possibly related to the sessile nature of plants. Protein kinases are involved in the very precise transduction of a plethora of signals through protein phosphorylation in plant cells but also most other eukaryotes (Cock *et al.*, 2002; Stone and Walker, 1995). Plant RLKs are related to animal kinases, producing a superfamily containing the serine-threonine-tyrosine kinases (Shiu and Bleecker, 2001b).

Plant RLKs usually contain a signal peptide (SP), an ectodomain, a transmembrane region (TMR) and an intracellular region. Genes encoding receptor kinases usually contain a signal peptide in the beginning of the coding region. This part is cleaved off during synthesis and is not present in the mature protein. Most RLKs are localized to the plasma membrane and the ligand-binding ectodomain resides outside of the cell membrane in the extracellular space of the plant cell, commonly referred to as the apoplast. The TMR of the protein is the region which resides within the cell membrane. The intracellular region containing the kinase domain is located in the intracellular space of the plant cell.

All RLKs have an intracellular protein kinase domain (Hanks *et al.*, 1988). The main function of any protein kinase is phosphorylation of a substrate protein. Phosphorylation adds a phosphoryl group, which carries negative charge, to target protein (Stone and Walker, 1995). In plants, the amino acids phosphorylated by protein kinases, including RLKs, are mostly serines and threonines, but also tyrosines (Klaus-Heisen *et al.*, 2011; Oh *et al.*, 2009). Phosphorylation of the target protein can lead to structural rearrangements, changes in enzymatic activity, subcellular relocalization, protein stability or control of protein-protein interactions. Phosphorylation is a cost-effective way to initiate signalling as one ATP makes one signal. A kinase is commonly classified as active, if it contains all catalytic sites needed for phosphorylation, or inactive, if it has lost one or more of its residues required for activity in its catalytic sites due mutations (Kornev *et al.*, 2006). Notably, protein phosphorylation is one of the most researched post-translational modifications and has been studied for over a century (Pawson and Scott, 2005). Yet, despite all the research effort the complexity of signal transduction by protein phosphorylation still provides a plethora of questions for researchers.

The different types of RLKs are related by their kinase domains. The kinases are connected to various different ectodomain types containing different domains which are typically used to group RLKs into smaller subfamilies (Shiu and Bleecker, 2003). Usually RLK subfamilies are named according to their ectodomain region. The most abundant RLK families in the plant genomes are leucine-rich receptor-like protein kinases (LRR-RLKs) (Liu *et al.*, 2017; Torii, 2004). Kinases have fused to different ectodomains at different evolutionary different times, for example some LRR-RLKs were already present in algae (Liu *et al.*, 2017), but CRKs are only present in vascular plants (I).

There are proteins in plant genomes that resemble the ectodomains of several of the RLK subfamilies. These proteins are referred as receptor-like proteins (RLPs) (Shiu and Bleecker, 2003). RLPs are related to the RLK group and may either share their ancestry with a protein that originally fused with the kinase domain, or result from the loss of the kinase domain. In addition to RLPs there are receptor-like cytoplasmic kinases (RLCKs), which lack the ectodomain, but can function as co-receptors for RLKs (Liang and Zhou, 2018).

1.3.1 FUNCTIONS OF RLKS

Plants use receptors to sense extracellular signals. In RLKs the ectodomain perceives signal in the apoplast leading to RLK activation and subsequent signal transduction events in the cytosol. RLKs have roles in many central processes including plant development, stress responses and hormone perception (De Smet *et al.*, 2009; Osakabe *et al.*, 2013; Shiu and Bleecker, 2001a; Tör *et al.*, 2009). Some of the best studied functions of RLKs in stress responses are related to the innate immunity of plants (Greeff *et al.*, 2012; Wu and Zhou, 2013). The RLKs and RLPs can function as pattern recognition receptors (PRRs) that can recognize pathogen-originated or pathogen induced molecules, so-called microbe-associated molecular patterns (MAMPs) or damage-associated molecular patterns (DAMPs) (Boutrot and Zipfel, 2017; Couto and Zipfel, 2016; Ranf, 2017). One of the best known examples of such PRR is FLS2 that perceives flagellin (Gómez-Gómez and Boller, 2000). Often the receptor does not function alone. FLS2 also requires the co-receptor BAK1 for ligand-induced activation (Chinchilla *et al.*, 2007; Sun *et al.*, 2013). Actually, various ectodomain-mediated interaction networks of different LRR-RLKs have evolved for specific responses to different extracellular cues (Smakowska-Luzan *et al.*, 2018).

1.4 DUF26-CONTAINING PROTEINS

The domain of unknown function 26 (DUF26) is defined by a conserved cysteine motif (C-8X-C-2X-C) in its core (Chen, 2001). The domain is also referred to as stress-antifung domain, Gnk2-domain and is identified by the PFAM code PF01657. Genes containing this domain can be categorized into three main groups: cysteine-rich receptor-like secreted proteins (CRRSPs), plasmodesmata-localized proteins (PDLPs) and cysteine-rich receptor-like protein kinases (CRKs) (Figure 2).

CRRSP contain a signal peptide (SP) followed by one or two DUF26 domains. CRRSPs containing a single DUF26 are referred as sdCRRSPs. The best know example of an sdCRRSP is the GNK2 protein from the gymnosperm tree *Ginkgo biloba*. It has antifungal properties *in vitro* and the crystal structure of the protein has been resolved (Miyakawa *et al.*, 2009). GNK2 functions as a mannose-binding lectin (Miyakawa *et al.*, 2014). CRRSPs from maize with a configuration of two DUF26 domains (ddCRRSPs) have recently been found to exhibit similar mannose-binding properties to GNK2 (Ma *et al.*, 2018) and may be involved in its response to pathogen infection. In addition, a rice CRRSP has been associated with the response to salt stress (Zhang *et al.*, 2009).

PDLPs contain a single transmembrane region (TMR) in addition to SP and two DUF26 domains. They are found to localize to the plasmodesmata (Lee *et al.*, 2011; Thomas *et al.*, 2008), which are pore structures connecting the cytoplasm of plant cells through cell walls. The functions of PDLPs are likely to be related to the regulation of plasmodesmata and plasmodesmal permeability (Brunkard and Zambryski, 2017; Cui and Lee, 2016; De Storme and Geelen, 2014; Lee *et al.*, 2011; Lim *et al.*, 2016; Thomas *et al.*, 2008).

CRKs contain a SP, in most cases two DUF26 domains, TMR and a cytosolic protein kinase domain. A group of CRKs, which is unique to *Selaginella moellendorffii*, contains only a single DUF26 domain. Those *Selaginella* CRKs are referred to as sdCRKs (I). In addition, there are a few examples of CRKs with three or four DUF26 domains in the extracellular domain. The functions of CRKs are related to plant development and stress responses in *Arabidopsis thaliana* (Acharya *et al.*, 2007; Burdiak *et al.*, 2015; Chen *et al.*, 2003; Chen *et al.*, 2004; Hunter *et al.*, 2019; Idänheimo *et al.*, 2014; Lee *et al.*, 2017; Tanaka *et al.*, 2012; Wrzaczek *et al.*, 2010; Xu *et al.*, 2019; Yadeta *et al.*, 2017; Yeh *et al.*, 2015), *Oryza sativa* (Chern *et al.*, 2016) and *Hordeum vulgare* (Rayapuram *et al.*, 2012). CRKs encoding only the intracellular part of CRK (i.e. mainly kinase domain) are called cysteine-rich receptor-like cytoplasmic kinases (CRCKs) (Figure 2).

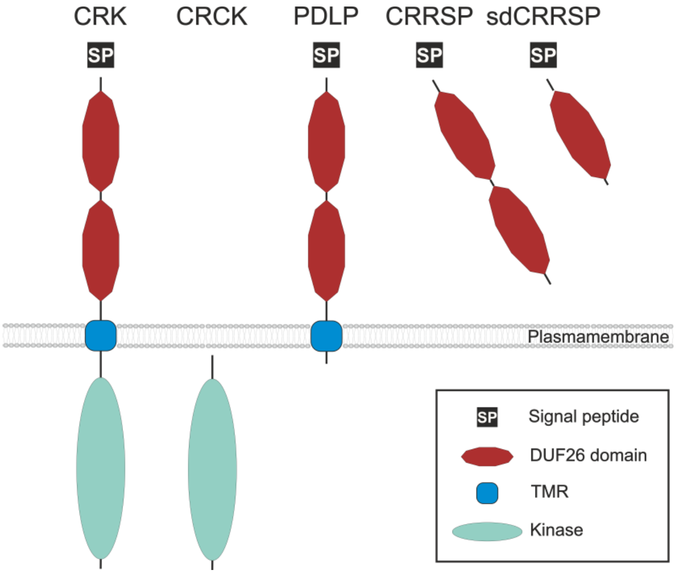

Figure 2 Main domain compositions of the DUF26 proteins. The CRKs have a typical RLK domain composition including the ectodomain (in case of CRKs it contains DUF26 domains), transmembrane region (TMR) and kinase domain. CRCKs have only the intracellular kinase domain. PDLPs contain the ectodomain and TMR and CRRSPs contain only one or two DUF26 domains. Signal peptides are present in gene sequences but cleaved of from the mature proteins.

1.5 GENE ANNOTATION

1.5.1 GENE MODELS

When a genome is sequenced, the sequence information, the reads, are assembled into contigs and pseudochromosomes and subsequently the coding portion of the genome is identified and annotated (Dominguez Del Angel *et al.*, 2018). Protein coding genes are annotated as gene models. Based on the genome assembly, *ab initio* annotation programs (e.g. Augustus (Stanke *et al.*, 2008), GeneMark (Ter-Hovhannisyan *et al.*, 2008), Glimmer (Delcher *et al.*, 1999)) can predict the most likely exon-intron structure and regulatory regions of a gene. The *ab initio* models and evidence from expressed genes, i.e transcripts and known genes from other species are combined using combiner programs, e.g. MAKER (Cantarel *et al.*, 2008), thereby generating a predicted gene model. The gene model is an estimate of the likely gene structure and may not always be a correct representation of the real gene. The exon-intron structure can be verified by sequencing the mRNA product of the gene.

One gene can be represented by more than one gene model as many genes have splice variants. This means that when mRNA is produced the exon-intron structure can be utilized in more than one way by alternatively splicing together different exons. This offers more variability to protein products from one gene and may allow faster adaptation to different conditions. Thus, databases can contain more than one splice variant for one gene. These models are usually ranked according to evidence from the transcriptome so that the most prominent splice variant is called the primary transcript. Sometimes the longest transcript is named as primary, if no information is available about the presence of different transcripts is not available.

1.5.2 AUTOMATED AND MANUAL ANNOTATION

Gene models can be annotated either manually or produced by automation. Manual annotation involves a person searching for potential coding regions and annotating genes based on the available evidence manually themselves or by using a software program to help define exon-intron borders. Automated annotation in principle works the same way using a pipeline of programs and evidence defined by the user. Nowadays, genomes are annotated automatically as manual annotation would be far too time consuming. Even though the annotation programs are constantly improving, incorrectly annotated gene models are still present in databases. It is good practice to verify annotation quality by comparison with well-known sets of conserved genes or by manually verifying a subset of automatically annotated genes.

Several types of evidence are used to support the process of automated annotation. The evidence used may be data from RNA sequencing, expressed sequence tags (ESTs) or existing gene models from related species. The quality of automated gene annotationof the genomecan be checked using a few well known gene sets of conserved single-copy genes: 'Benchmarking Universal Single-Copy Orthologs' (BUSCO) (Simao *et al.*, 2015) and 'Core Eukaryotic Genes Mapping Approach' (CEGMA) (Parra *et al.*, 2007). This approach reflects the success of annotation of the orthologs of well-known single-copy genes and the quality of those annotations in terms of full-length versus partial annotation. This approach does not tell us about the quality and annotation coverage of the gene families. In plant genomes, the presence of a gene family can be compared to that of a set of known core gene families, to get an estimate of the annotation quality of the gene families (Li *et al.*, 2016a; Veeckman *et al.*, 2016).

1.5.3 ANNOTATION ISSUES AND MANUAL CURATION

Gene models can contain several types of annotation defects. Most common types of annotation errors in DUF26 proteins were partial gene models or gene models with additional sequences (I and III). Partial gene models lack full or partial exons. The most common reasons for such annotation errors are gaps or assembly problems in the genome sequence. Sometimes a sequencing error in a single base can introduce a premature stop codon in the genome sequence causing problems for annotation algorithms. In addition, the evidence used in annotation can be partial and thus resulting in a partial gene model. An additional sequence, which does not really belong to the gene, can be introduced to a gene model over short distances to nearby genes or open reading frames. In particular, tandemly repeated genes which are close to each other can cause challenges for annotation programs; as shown in the case of AtCRK16 and AtCRK17 where in previous annotations in TAIR10 the ectodomain of AtCRK16 was annotated as part of the AtCRK17 (Figure 3A).

In rare cases different combinations of exons from tandemly repeated genes can also lead to a gene model that looks like the correct gene, but is actually a fused gene model built from exons of two or more separate genes (Figure 3B). Thus, examination of the intron lengths and genomic region of the tandem genes might be necessary to find all gene models.

Genes can also escape annotation entirely: there are examples of re-annotations of gene families where a large number of previously unannotated genes have been found. For example, in the re-annotation of the NB-LRR genes from potato (*Solanum tuberosum*), 317 new gene models were identified (Jupe *et al.*, 2013) and in genome-scale re-sequencing, and similarly annotation of the wild strawberry (*Fragaria vesca*) lead to identification of 631 new gene models and corrections to 39.3% of the existing gene models (Li *et al.*, 2017).

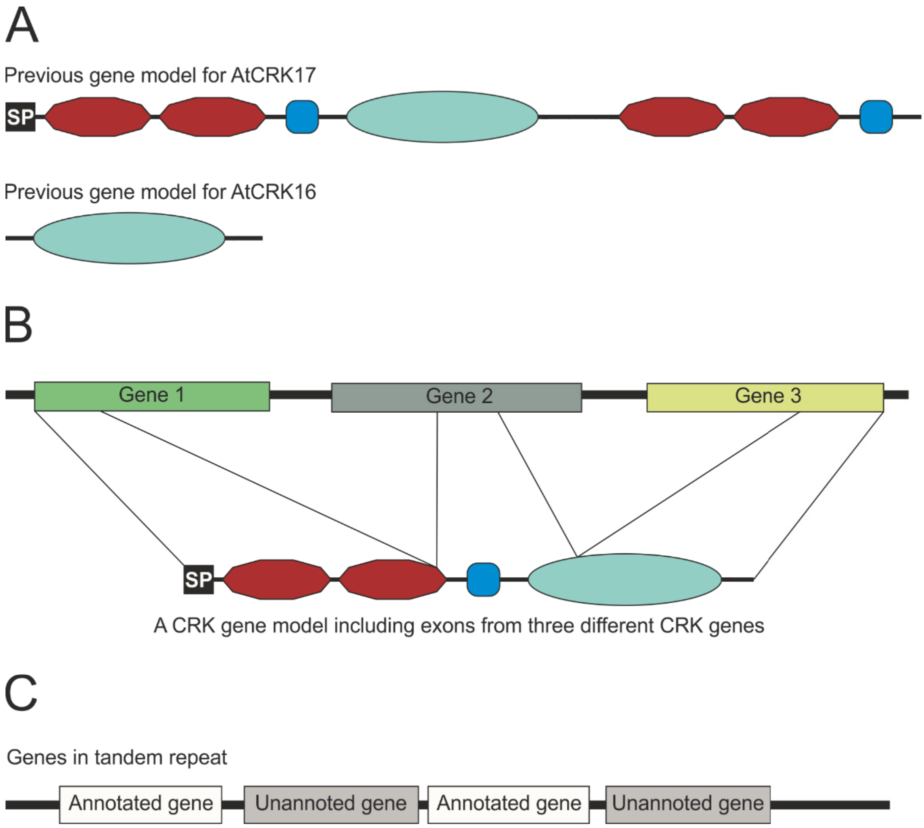
A
Previous gene model for AtCRK17
SP
Previous gene model for AtCRK16
B
Gene 1
Gene 2
Gene 3
SP
A CRK gene model including exons from three different CRK genes
C
Genes in tandem repeat
Annotated gene
Unannoted gene
Annotated gene
Unannoted gene

Figure 3 Specific issues in gene annotation on tandemly duplicated genes located close to each other. A) The beginning of CRK16 was annotated as part of CRK17 resulting a truncated gene model for AtCRK16 and falsely identifying additional domains in the gene model for AtCRK17. B) The gene model was composed of the exons from 3 different genes. C) Only some of the genes located in the tandem repeat were annotated.

2 AIMS OF THE STUDY

The main aim of this study was to resolve the evolutionary history of DUF26-containing genes, to be able to give functional analyses an evolutionary perspective, and provide future opportunities to transfer functional information from model species to other species. The evolution of DUF26 genes in plant kingdom was analysed in I and connected to the functions of CRKs in *Arabidopsis thaliana* described in II. To achieve these goals, it was essential to use data of high quality. Thus, the annotation and curation of the gene models for evolutionary analyses was carried out carefully and this process is described in III, with annotation quality issues highlighted in IV.

3 MATERIAL AND METHODS

3.1 PLANT GENOMES

Study species were selected to cover as broad a diversity of the plant kingdom as possible from available sequenced plant genomes. Phylogenetic relationships among the species are presented in the Figure 4. A few species were excluded from the further analyses based on poor annotation quality (Table 1). To specifically check, whether CRK2 orthologs are also found to be tandemly duplicated in the wild relatives of cultivated tomato (*Solanum lycopersicum*), the genomes of *Solanum lycopersicoides* (The Solanum lycopersicoides Genome Consortium) and *Solanum penellii* (Bolger *et al.*, 2014) were searched using BLAST and by verifying the found gene models.

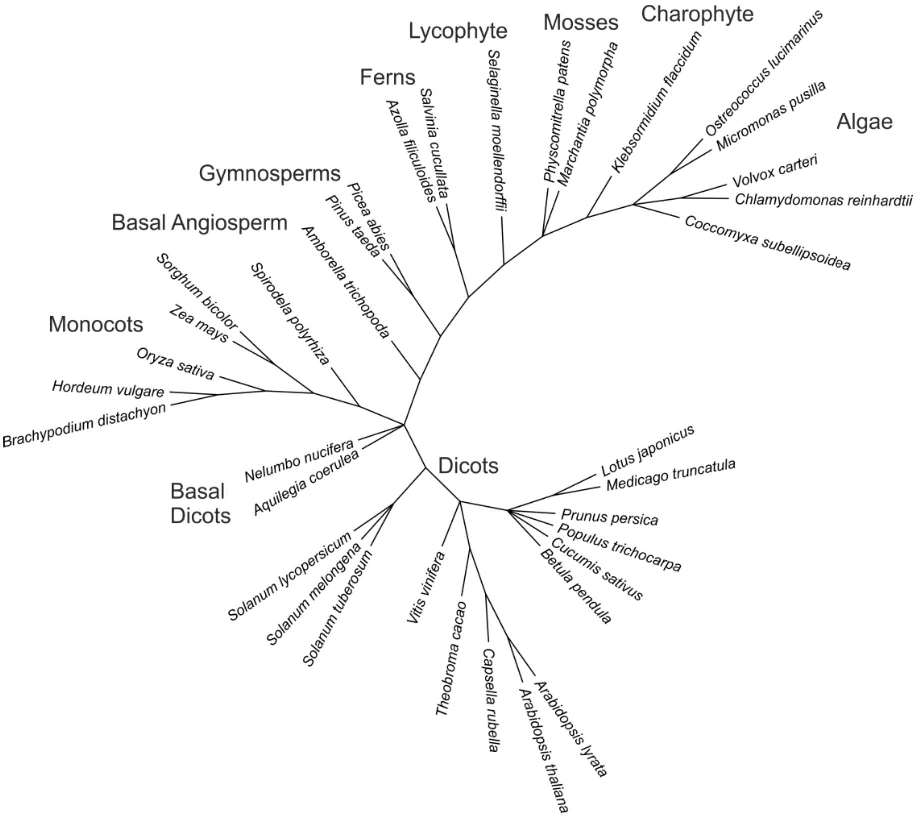

Figure 4 Species tree presenting plant and algal genomes analysed for DUF26 genes. Phylogenetic relationships are based on the definitions from NCBI.

Table 1. Plant and algae genomes analysed for DUF26 genes. DUF26 genes were not identified from the algal and charophyte species. DUF26 genes from the genomes of *Betula pendula, Hordeum vulgare* and *Nelumbo nucifera* were manually annotated as prior gene models did not exist at the time of the annotation. The annotation quality was defined as good, if less than 20% of the final dataset of annotated DUF26 genes needed curation or manual annotation (i.e. were erroneously annotated or missing from the genome annotation version mentioned in the table), modarate if less than 60% and poor if more than 60%. The species marked with gray were not included in the evolutionary analyses due to quality problems.

Species name	Genome annotation version	Annotation quality of DUF26 genes	Reference
Amborella trichopoda	v1.0	Moderate	(Amborella Genome Project, 2013)
Aquilegia coerulea	v1.1	Moderate	(Filiault *et al.*, 2018)
Arabidopsis lyrata	v1.0	Moderate	(Hu *et al.*, 2011)
Arabidopsis thaliana	TAIR10	Good	(The Arabidopsis Genome Initiative, 2000)
Betula pendula	v1.0	Manually annotated	(Salojärvi *et al.*, 2017)
Brachypodium distachyon	v3.0	Good	(Int Brachypodium Initiative, 2010)
Capsella rubella	v1.0	Good	(Slotte *et al.*, 2013)
Chlamydomonas reinhardtii	v5.5	No DUF26 genes	(Merchant *et al.*, 2007)
Coccomyxa subellipsoidea	v2.0	No DUF26 genes	(Blanc *et al.*, 2012)
Cucumis sativus	v1.0	Moderate	Unpublished
Hordeum vulgare	v2.2	Manually annotated	(Mascher *et al.*, 2017)
Klebsormidium flaccidum	v1	No DUF26 genes	(Hori *et al.*, 2014)
Marchantia polymorpha	v3.1	Good	(Bowman *et al.*, 2017)
Medicago truncatula	Mt4.0v1	Moderate	(Tang *et al.*, 2014; Young *et al.*, 2011)
Micromonas pusilla	v3.0	No DUF26 genes	(Worden *et al.*, 2009)
Nelumbo nucifera	v1	Manually annotated	(Ming *et al.*, 2013)

Oryza sativa	v7	Good	(Goff *et al.*, 2002)
Ostreococcus lucimarinus	v2.0	No DUF26 genes	(Palenik *et al.*, 2007)
Physcomitrella patens	v3.0	Good	(Lang *et al.*, 2018)
Picea abies	v1.0	Moderate	(Nystedt *et al.*, 2013)
Populus trichocarpa	v3.0	Moderate	(Tuskan *et al.*, 2006)
Prunus persica	v1.0	Moderate	(International Peach Genome Initiative, 2013)
Selaginella moellendorffii	v1.0	Poor	(Banks *et al.*, 2011)
Solanum lycopersicum	iTAG2.4	Moderate	(Tomato Genome Consortium, 2012)
Solanum melongena	r2.5.1	Moderate	(Hirakawa *et al.*, 2014)
Solanum tuberosum	v3.4	Moderate	(Potato Genome Sequencing Consortium, 2011)
Sorghum bicolor	v2.1	Moderate	(Paterson *et al.*, 2009)
Spirodela polyrhiza	v1	Moderate	(Wang *et al.*, 2014)
Theobroma cacao	v1.1	Moderate	(Motamayor *et al.*, 2013)
Vitis vinifera	Genoscope.12X	Poor	(Jaillon *et al.*, 2007)
Volvox carteri	v2.0	No DUF26 genes	(Prochnik *et al.*, 2010)
Zea mays	AGPv3	Moderate	(Schnable *et al.*, 2009)
Azolla filiculoides	v1.1	Poor	(Li *et al.*, 2018)
Salvinia cucullata	v1.2	Poor	(Li *et al.*, 2018)
Lotus japonicus	v3.0	Poor	(Sato *et al.*, 2008)
Pinus taeda	v1.01	Poor	(Zimin *et al.*, 2014)

3.2 GENE ANNOTATION

Gene annotation was carried out as described in I and III. The importance of the annotation quality for analyses utilizing gene model information is discussed in IV.

3.3 EVOLUTIONARY ANALYSES OF DUF26 PROTEINS

Sequence-based analyses of protein family evolution and adaptation processes are described in I.

3.4 PHENOTYPING OF CRKS

The phenotypes of *Arabidopsis thaliana crk* loss-of-function mutants were characterized as described in II.

4 RESULTS AND DISCUSSION

4.1 ANNOTATION QUALITY

Annotation quality is an important factor in all downstream analyses utilizing the annotated gene models (Stein, 2001). The annotation quality of analysed plant genomes varied considerably. Genomes with several annotation versions like *Arabidopsis thaliana* and *Oryza sativa* had overall high annotation quality but still few annotation errors were nevertheless present. At the other end of the annotation quality spectrum were genomes including *Lotus japonicus* and the ferns *Azolla filiculoides* and *Salvinia cucullata*: the quality of their genome assembly and annotation was insufficient for gene family analyses, therefore they were excluded from the further analyses (Table 1). *Pinus taeda* was excluded due to the unreliable high number of gene models for DUF26 genes (364) compared to the other conifer species *Picea abies*.

The annotation quality issues, such as partial and missing annotations, of DUF26-containing genes were overcome by manual curation and re-annotation of these genes in the analysed plant genomes (the annotation process described in detail in III). The effects of the annotation errors have been described in detail in IV. In the phylogenetic analyses, the erroneous or missing gene models can lead to false interpretation of the relationships between the analysed genes.

In the analysis of the DUF26-containing genes, a main challenge of the annotation was to separate different domain compositions as gene model looking like CRRSP could erroneously be annotated PDLP or CRK. To overcome this problem species-specific trees were produced to spot the CRRSPs located within CRK or PDLP clades and these gene models were still separately verified to be CRRSPs. Part of CRKs and CRRSPs are clustered in tandem repeats in the chromosomes and this easily leads to problems in the gene annotation.

4.2 EVOLUTIONARY HISTORY OF DUF26 GENES

4.2.1 DUF26 IS SPECIFIC TO LAND PLANTS ON THE SEQUENCE LEVEL

The 36 plant and algal genomes were analyzed for DUF26 domains. Based on the amino acid sequence, the DUF26 domain is only present in land plants and not found from algae, charophytes, fungi or animals (I). This specificity to plants at the sequence level was verified by also querying fungal and animal genomes. The non-coding sequence was also analysed, but nothing that could be ancestral to DUF26 was identified from algae or charophyta species. The most ancestral form of DUF26 genes are sdCRRSPs. They might be *de novo* genes and their short length supports this hypothesis. The main features of *de novo* genes are their short length, low expression and high diversification compared to other genes (Li *et al.*, 2016b). Alternatively, they could have evolved from other genes or domains, but that would have involved a drastic change at the sequence level. One more option is horizontal gene transfer from other species (Soucy *et al.*, 2015). sdCRRSPs from the moss *Physcomitrella patens* and liverwort *Marchantia polymorha* are different from each other and from other plants indicating high diversification in early forms of the DUF26 domain between species. This is in line with the hypothesis that young domains evolve in less constrained manner than evolutionary older domains (Toll-Riera and Alba, 2013). In lycophytes and the lineages leading to ferns, gymnosperms and angiosperms the main features of the DUF26 domain are clearly conserved at the sequence level, this could indicate that they have conserved structure and function. After the duplication of the DUF26 domain, the DUF26-A and DUF26-B forms can be separated from each others (Figure 2d in I) even though they both are still recognized as the same PFAM domain. The divergence of the two DUF26 domains may be the result of their evolution towards different functions.

4.2.2 ORIGIN OF DIFFERENT DOMAIN COMPOSITIONS

In the global phylogenetic tree of the DUF26 genes, two main groups, named as α and β, were found. The more basal genes grouped to α group and recently expanded genes produced β-group that branches out from the α-group. The phylogenetic tree was rooted to the sdCRRSPs from *Selaginella moellendorffii* as the various domain compositions of DUF26-containing proteins (ddCRRSPs, CRKs, PDLPs) have arisen from these relatively simple sdCRRSPs. In addition to the sdCRRSPs, the α-group included basal type of CRKs (bCRKs) and PDLPs. Gymnosperm-specific CRKs also belonged to this group. They were named as variable CRKs (vCRKs) to separate them from the basal clade of the CRKs, which includes genes from all species that have CRKs. The β-group included vCRKs from monocots and dicots and ddCRRSPs.

CRKs are already present in lycophytes, with one DUF26 domain, indicating the fusion of sdCRRSP with the TMR and kinase domain. Lycophyte genomes also encode the typical CRK configuration with two DUF26 domains in the extracellular region. It is likely that TMR and kinase originate from the same source, for example another type of RLK since intermediate domain architectures are absent from the moss or lycophyte genomes. In the case of CRKs with the double-DUF26-domain configuration, it remains unclear whether the duplication of the DUF26 domain occurred in CRRSPs prior to fusion with the TMR and kinase domain or whether this duplication took place after the appearance of sdCRKs. However, the genome of *Selaginella moellendorffii* does not contain CRRSPs with double DUF26 domains, suggesting that this duplication is likely to have taken place in CRKs.

PDLPs are present in the genomes of gymnosperms and angiosperms. They are likely to have evolved from ddCRKs by loss of the protein kinase domain. They group close to gymnosperm-specific vCRKs in phylogenetic trees which could indicate that they have emerged from those CRKs before the split to gymnosperms and angiosperms. They could already be present in ferns, but there is not enough evidence to definitively draw this conclusion. The only supporting evidence is the partial gene model from fern *Marsilea quandrifolia* and it unfortunately lacks the TMR region. The latest fern genomes do not shed any light on this issue as the quality of their DUF26 annotation is very low.

The variable group of DUF26 genes contains CRRSPs and few sdCRRSPs which have emerged following loss of the TMR and kinase domains. In the case of secondary sdCRRSP, one DUF26 is also lost. These losses may have happened once or in several steps: for example the kinase domain could have been lost prior the loss of TMR. However, we did not find clear intermediate forms, which would suggest that losses mainly happened in one step.

Domain loss is more likely for domains located at the end of the gene (Weiner *et al.*, 2006), as the C-terminal loss does not affect transcriptional regulation and the transcription starting site. Still, the CRCKs have emerged at least three times by losing the ectodomain region and TMR of CRKs. Most common of these CRCKs is the CRCK-I group shared across the angiosperms. They seem to be very strictly selected against duplications as they are present in angiosperms as a single copy gene.

Fusion of CRRSPs with the TMR and kinase domain has taken place at least once. Based on their sequence similarity, the main kinase domain of CRKs shares the same ancestry as S-locus lectin and Leucine-rich repeat receptor-like kinases from LRR clade 3 as defined in (Zulawski *et al.*, 2014). However, grasses contain few CRKs that have a different protein kinase domain related to the kinase domain of Concanavalin-A RLKs and a different exon-intron structure compared to the majority of the CRKs (Figure 6b in I). This suggests that the kinase domain in this subgroup of CRKs has been exchanged with another kinase. Alternatively, the typical CRK kinase domain has been lost, followed by the subsequent fusion of ectodomain with a different kinase domain.

Overall, the evolution of the domain compositions in DUF26-containing genes has been complex. It has included several domain fusions, losses, duplications and at least one domain swap (Figure 5). Domain rearrangements seem to be a continuous feature of this gene family, whereas the losses and internal domain duplications take place in more recently evolved genes. These modifications are likely to have provided material for adaptive evolution.

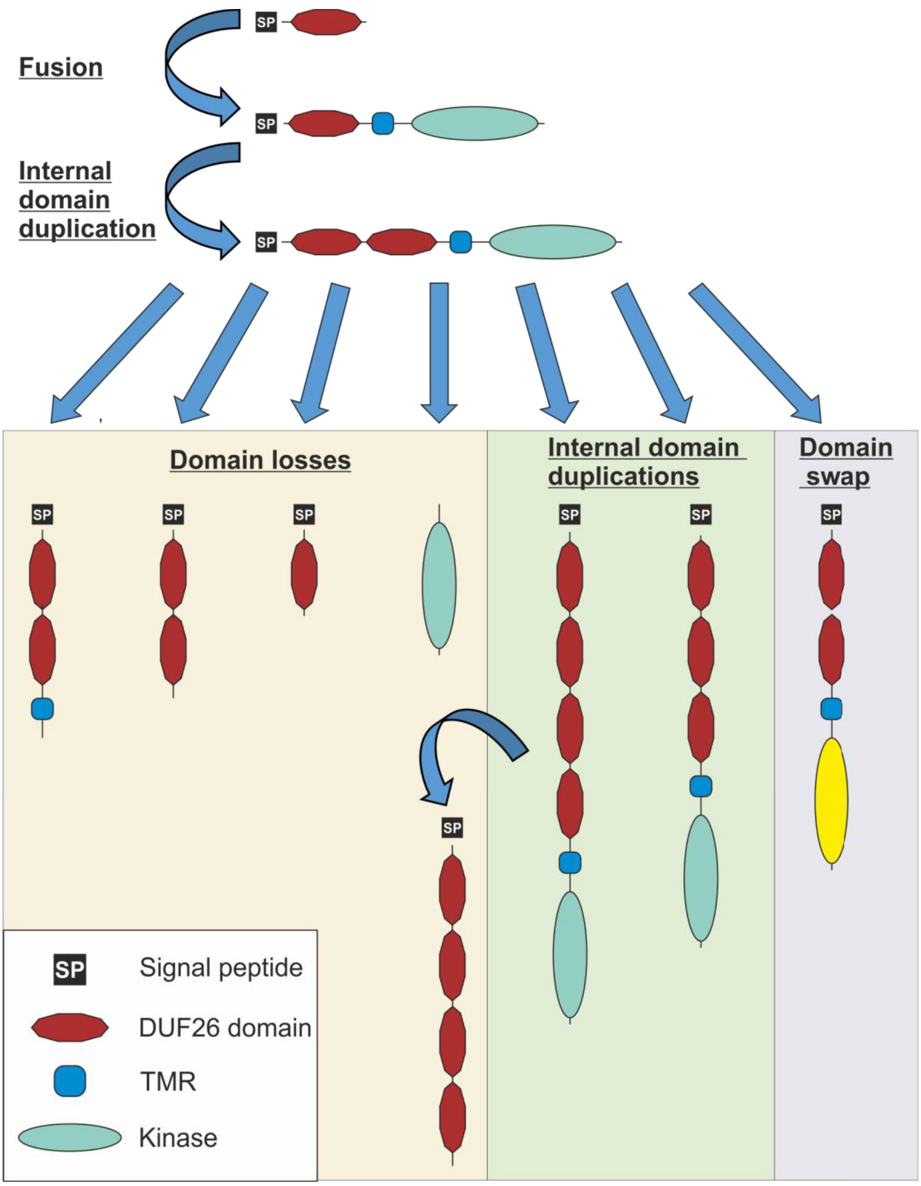

Figure 5 Domain composition rearrangements in DUF26-containing genes. DUF26 genes have undergone several domain fusion, loss, duplication and exchange events during their evolution.

4.3 FUNCTIONS OF THE DUF26 DOMAIN AND THE DUF26-CONTAINING PROTEINS

The DUF26-containing proteins are involved in the regulation of plant development and stress responses. Least is known about the CRRSPs, but the few studies available associate them with defense against fungal pathogens (Ma *et al.*, 2018; Miyakawa *et al.*, 2014) and response to salt stress (Zhang *et al.*, 2009). PDLPs have been described as localizing to the plasmodesmata, which is also the origin of their name. They are involved in the regulation of plasmodesmatal function for example via the control of callose deposition which controls plasmodesmal permeability (Cui and Lee, 2016). Thereby PDLPs participate in pathogen response (Caillaud *et al.*, 2014) and symplastic intercellular signaling (Brunkard and Zambryski, 2017; Lim *et al.*, 2016). They have also been identified as targets for viral movement proteins (Amari *et al.*, 2010). CRKs are perhaps the best characterized subgroup of DUF26-containing proteins. Based on gene expression analysis, *CRKs* have been linked with stress responses and reactive oxygen species (ROS) signaling (Wrzaczek *et al.*, 2010). The phenotypes of *crk* mutants support their involvement in stress signaling but also in the regulation of plant development (II). In particular, several CRKs have been linked to immunity. Those CRKs include AtCRK13 (Acharya *et al.*, 2007), AtCRK5 (Chen *et al.*, 2003), AtCRK4, AtCRK19 and AtCRK20 (Chen *et al.*, 2004), AtCRK4, AtCRK6 and AtCRK36 (Yeh *et al.*, 2015), AtCRK28 and AtCRK29 (Yadeta *et al.*, 2017), AtCRK36 (Lee *et al.*, 2017), HvCRK1 (Rayapuram *et al.*, 2012) and OsCRK6 and OsCRK10 (Chern *et al.*, 2016). Other CRKs have been connected to ROS signaling, like AtCRK6 and AtCRK7 (Idänheimo *et al.*, 2014) and AtCRK5 (Burdiak *et al.*, 2015). They are also involved in the responses to abiotic stress, for example AtCRK36 (Tanaka *et al.*, 2012) and AtCRK2 (Hunter *et al.*, 2019), while AtCRK16 and AtCRK36 may be associated with adaptation to climate conditions (Xu *et al.*, 2019).

A comprehensive analysis of the functions of the *Arabidopsis thaliana* CRKs was carried out by large-scale phenotyping of a loss-of-function mutant collection (II). This study included most AtCRKs; with the exceptions of only the putative pseudogene AtCRK35, the truncated AtCRK9 and three other CRKs, AtCRK27, AtCRK34 and AtCRK44, for which no homozygous T-DNA insertion lines were identified. Different biotic and abiotic stress treatments were used to investigate the role of CRKs. Furthermore, parameters assessing stomatal regulation, plant development and also photosynthesis were collected and analyzed. Overall, *crk* mutants displayed phenotypes in the response to various treatments compared to wild type plants (Figure 2 in II). Out of all the CRKs, bCRK *crk2* showed the most striking phenotypes in comparison to wildtype plants. Even under control conditions, without any treatments, *crk2* exhibited a dwarf-sized phenotype. It was also the only *crk* mutant to flower later than wild type plants. CRK2 has also been shown to play a role in seed germination (Bassel *et al.*, 2011) and salt stress tolerance (Hunter *et al.*, 2019). Based on the analyses in I, bCRKs evolved according to the dosage balance hypothesis and notably CRK2 is present as a single or double copy in most plant species with exception of species from the family Solanaceae; tomato (*Solanum lycopersicum*) and potato (*Solanum tuberosum*), which contain five and nine copies of CRK2, respectively. The reason for its expansion in these species in unknown, but could be linked to adaptation or random duplication rather than domestication, as the expansion was not found from the eggplant (*Solanum melongena*) genome but is present in the wild relatives of cultivated tomato, *Solanum penellii* and *Solanum lycopersicoides*. It is hypothesized, that the proteins coded by genes evolving under dosage balance have essential functions and a high number of protein-protein interactions. Thus, the loss-of-function mutants for these genes are predicted to show stronger phenotypes compared to mutants in other genes (Veitia, 2005). This was also the case for *bcrk* mutants when phenotypic data from the *crk* mutant collection (II) was re-analyzed based on the phylogenetic groupings in I (Figure 7e in I). Accordingly, the phenotypes of *crk2* as bCRK support this hypothesis.

Even though the DUF26 genes have been linked with different biological responses and processes including stress response and development, their biochemical and mechanistic functions are poorly understood. The CRKs possess an intracellular protein kinase domain which is predicted to phosphorylate a set of substrate proteins. However, the substrates for most CRKs have not been identified and the large number of CRKs and their high sequence similarity suggests that they may have strongly overlapping sets of substrates, which may further complicate substrate identification. Unlike the protein kinase domain, very little is known about the DUF26 domain. Previously, structural data has only been available for a DUF26 protein with a single DUF26 domain, the sdCRRSP GNK2 (Miyakawa *et al.*, 2009). Unlike GNK2, most DUF26 proteins contain DUF26 domains in a tandem arrangement. The crystal structures of two PDLPs from *Arabidopsis thaliana* were resolved, facilitating functional analysis of the PDLP and CRK ectodomain (I). These two PDLPs belong to two different subclades of the PDLP-II main clade in higher plants and their structures have very high similarity. The individual DUF26 domains of both PDLPs also share strong structural similarity with GNK2. Unexpectedly, they also share strong structural similarity with two fungal lectins (i.e. carbohydrate-binding proteins) which on the sequence level were neither similar to each other nor with DUF26 proteins. This outcome can be result of convergent evolution producing similar structures and functions, or alternatively because they share a very ancient common origin but have since undergone a lot of changes in sequence level.

GNK2 as well as two sdCRRSPs from maize have been described to bind mannose (Ma *et al.*, 2018; Miyakawa *et al.*, 2014). However, the tandem DUF26 arrangement found in AtPDLP5 and AtPDLP8 does not bind mannose *in vitro* (Supplementary Figure 13a in I). Even though the structures of

AtPDLP5 and AtPDLP8 are highly similar, the surface charges are very different (Figure 5b in I). This indicates that they could potentially bind different ligands. Many different carbohydrates or glycopeptides are present in the apoplast and PDLPs and accordingly DUF26-containing proteins, including CRKs, might bind a very diverse set of ligands. The specificity of a lectin is usually high (Iskratsch *et al.*, 2009) and therefore identifying different ligands for all DUF26 proteins will be challenging. The situation is further complicated by recent reports that the RLK FERONIA, which contains malectin domains in its ectodomain, can bind peptide ligands (Haruta *et al.*, 2014; Stegmann *et al.*, 2017) but also polygalacturonic acid, a carbohydrate and a key component of pectin (Feng *et al.*, 2018). Taken together the evidence from structural analysis and known carbohydrate-binding of sdCRRSPs may suggest that the DUF26 domain functions as a lectin. However, this deduction will have to be verified in the future by detailed interaction studies.

4.4 THE MODE OF EVOLUTION WITHIN DUF26 GENES

In *Arabidopsis thaliana*, the CRKs were reported already by Shiu and Bleecker (2001b) as an extreme case of RLKs located in tandem repeats. These tandemly duplicated CRKs belong to vCRKs (I). Similarly, the AtCRRSPs are mostly located in tandem arrays. These tandem duplications are also lineage-specific (Figure 3e in I). This indicates that more recently evolved genes in the β-group expand through lineage-specific tandem duplications. For example, *Arabidopsis thaliana* contains such an extremely recent CRRSP tandem repeat, that the genes within this region still have identical amino acid sequences.

Genes in the α-group, sdCRRSPs, bCRKs, PDLPs and CRCK-Is, are of more ancient origin and typically distributed throughout the genomes. For the α-group, WGDs are likely to be the main source of duplications with a few exceptions (for example CRK2 homologs in tomato and potato). They likely have more conserved functions and interaction partners and therefore evolve according to the dosage balance model. The duplicability of the genes belonging to gene family is also known to indicate their essentiality. The size-conserved families, like the subgroups in the α-group, have lower evolutionary rates, a higher proportion of essential genes, higher expression levels and a higher proportion of broadly expressed genes, when compared to the members of families fluctuating in size (Chen *et al.*, 2010).

The different evolutionary modes of the α- and β-group of DUF26-containing genes suggest that subfamilies within a single gene family can evolve in drastically different ways. This might be associated with the age and functional specialization of the genes involved. Older members of the gene family may have more specific functions and interaction partners, including ligands, substrates or co-receptors. This causes selective pressure to keep these proteins more similar and thus they evolve according to the dosage balance model. More recently evolved genes are expanding through tandem duplications in a lineage-specific manner. Those genes have not yet diverged and likely overlap in their functions. During evolutionary time, these genes may acquire more specific functions which will subsequently drive them towards a dosage-balance mode of evolution.

5 CONCLUSIONS

Adaptation and precise responses to environmental changes are absolutely necessary for plants as sessile organisms and therefore, specialized yet dynamic signaling networks are essential. As CRKs, PDLPs and CRRSPs all have important biological functions in plant stress signaling and development, their evolution reflects the need to adapt and further optimize signal transduction. The evolutionary patterns observed for the DUF26 genes were far more complex than anticipated. This study demonstrated that recurrent domain rearrangements within the DUF26 family provide ample material for adaptation. This study also revealed that different subgroups of the DUF26 genes evolve differently depending on their age and consequently also their functional conservation. The more ancient genes with more conserved functions and likely conserved interactions with other proteins are typically not retained if they are tandemly duplicated. In contrast to this, the DUF26 genes of more recent evolutionary origin are likely to possess partially redundant functions and tandem duplications in these genes might give origin to variation that enables neofunctionalization or subfunctionalization and facilitates adaptation of the plant.

Despite the important biological functions of DUF26-containing proteins, their biochemical interactions with ligands, substrates or other proteins are poorly understood. The crystal structures of the ectodomains of AtPDLP5 and AtPDLP8 and their high structural similarity to fungal lectins, together with existing evidence, suggest that the DUF26 could function as a lectin domain. Their structural similarity with fungal lectins, without any sequence level similarity, is intriguing and allows several scenarios for the origin of the DUF26 domain. It could have emerged from non-coding DNA *de novo* and due to convergent evolution evolved to resemble fungal lectins. Alternatively, it could be the result of horizontal gene transfer between plant and fungal genomes. A third option is the common origin in the ancestor of plants and fungi that has diversified to a level, where sequence-level similarity is no longer high enough to identify homologs. Currently, there is insufficient data to determine the origin of the DUF26 domain, but in the future, as more high quality plant, algal and fungal genomes as well as more structural information about DUF26 and other lectin domains becomes available, it will likely be possible to identify the evolutionary origin of the DUF26 domain.

The majority of information on the biochemical and physiological roles of signaling proteins has been obtained experimentally using model plant species. Therefore, it is central to provide a firm phylogenetic basis for the relationships in gene families, in order to identify orthologs to transfer this functional information from model species to crops. Thus, this field of research can be utilized in the crop breeding for productivity or to enhance stress tolerance, which is centrally important to human survival in a time of anthropogenic climate change. To this end, high-quality gene family annotations allow identification of orthologs with high confidence. It is not usually possible to identify all orthologs in a protein family between distant plant species. In the case of the DUF26 proteins, it is possible to identify orthologs in the conserved and ancient α-group where tandem duplicates are rare. In the β-group, with its lineage-specific tandem duplications, ortholog recognition, especially between distant species, is frequently impossible in the absence of physiological and biochemical data. Thus, it is imperative to highlight, that genes with the highest sequence similarity are not necessarily orthologs.

Finally, gene annotation quality is an important but underappreciated aspect which is critical for any sequence-based analyses involving gene models - especially now, as genome assemblies and gene annotations are still being continually improved. Annotation programs and sequencing techniques for genomes and transcriptomes are constantly improving. Better sequencing techniques will provide longer high-quality reads. This will lead to better assemblies and better transcriptome data and will eventually lead also to drastically improved gene annotations, which will facilitate even more detailed evolutionary and functional analyses. These analyses will subsequently provide more information for applied research on improving crop plants based on research in model species.

REFERENCES

Acharya BR, Raina S, Maqbool SB, Jagadeeswaran G, Mosher SL, Appel HM, Schultz JC, Klessig DF, Raina R. 2007. Overexpression of CRK13, an Arabidopsis cysteine-rich receptor-like kinase, results in enhanced resistance to *Pseudomonas syringae*. *Plant Journal* **50**, 488-499.

Amari K, Boutant E, Hofmann C, Schmitt-Keichinger C, Fernandez-Calvino L, Didier P, Lerich A, Mutterer J, Thomas CL, Heinlein M, Mély Y, Maule AJ, Ritzenthaler C. 2010. A family of plasmodesmal proteins with receptor-like properties for plant viral movement proteins. *PLoS Pathogens* **6**, e1001119.

Amborella Genome Project. 2013. The *Amborella* genome and the evolution of flowering plants. *Science* **342**, 1241089.

Arnheim N, Krystal M, Schmickel R, Wilson G, Ryder O, Zimmer E. 1980. Molecular evidence for genetic exchanges among ribosomal genes on nonhomologous chromosomes in man and apes. *Proceedings of the National Academy of Sciences of the United States of America* **77**, 7323-7327.

Arnold B, Kim ST, Bomblies K. 2015. Single geographic origin of a widespread autotetraploid *Arabidopsis arenosa* lineage followed by interploidy admixture. *Molecular Biology and Evolution* **32**, 1382-1395.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. 2000. Gene ontology: tool for the unification of biology. *Nature Genetics* **25**, 25-29.

Banks JA, Nishiyama T, Hasebe M, Bowman JL, Gribskov M, dePamphilis C, Albert VA, Aono N, Aoyama T, Ambrose BA, Ashton NW, Axtell MJ, Barker E, Barker MS, Bennetzen JL, Bonawitz ND, Chapple C, Cheng C, Correa LG, Dacre M, DeBarry J, Dreyer I, Elias M, Engstrom EM, Estelle M, Feng L, Finet C, Floyd SK, Frommer WB, Fujita T, Gramzow L, Gutensohn M, Harholt J, Hattori M, Heyl A, Hirai T, Hiwatashi Y, Ishikawa M, Iwata M, Karol KG, Koehler B, Kolukisaoglu U, Kubo M, Kurata T, Lalonde S, Li K, Li Y, Litt A, Lyons E, Manning G, Maruyama T, Michael TP, Mikami K, Miyazaki S, Morinaga S, Murata T, Mueller-Roeber B, Nelson DR, Obara M, Oguri Y, Olmstead RG, Onodera N, Petersen BL, Pils B, Prigge M, Rensing SA, Riano-Pachon DM, Roberts AW, Sato Y, Scheller HV, Schulz B, Schulz C, Shakirov EV, Shibagaki N, Shinohara N, Shippen DE, Sorensen I, Sotooka R, Sugimoto N, Sugita M, Sumikawa N, Tanurdzic M, Theissen G, Ulvskov P, Wakazuki S, Weng JK, Willats WW, Wipf D, Wolf PG, Yang L, Zimmer AD, Zhu Q, Mitros T, Hellsten U, Loque D, Otillar R, Salamov A, Schmutz J, Shapiro H, Lindquist E, Lucas S, Rokhsar D, Grigoriev IV. 2011. The Selaginella genome identifies genetic changes associated with the evolution of vascular plants. *Science* **332**, 960-963.

Bassel GW, Glaab E, Marquez J, Holdsworth MJ, Bacardit J. 2011. Functional network construction in *Arabidopsis* using rule-based machine learning on large-scale data sets. *Plant Cell* **23**, 3101-3116.
Birchler JA, Veitia RA. 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**, 14746-14753.
Björklund AK, Ekman D, Elofsson A. 2006. Expansion of protein domain repeats. *Plos Computational Biology* **2**, 959-970.
Blanc G, Agarkova I, Grimwood J, Kuo A, Brueggeman A, Dunigan DD, Gurnon J, Ladunga I, Lindquist E, Lucas S, Pangilinan J, Proschold T, Salamov A, Schmutz J, Weeks D, Yamada T, Lomsadze A, Borodovsky M, Claverie JM, Grigoriev IV, Van Etten JL. 2012. The genome of the polar eukaryotic microalga *Coccomyxa subellipsoidea* reveals traits of cold adaptation. *Genome Biology* **13,** R93.
Bolger A, Scossa F, Bolger ME, Lanz C, Maumus F, Tohge T, Quesneville H, Alseekh S, Sorensen I, Lichtenstein G, Fich EA, Conte M, Keller H, Schneeberger K, Schwacke R, Ofner I, Vrebalov J, Xu YM, Osorio S, Aflitos SA, Schijlen E, Jimenez-Gomez JM, Ryngajllo M, Kimura S, Kumar R, Koenig D, Headland LR, Maloof JN, Sinha N, van Ham RCHJ, Lankhorst RK, Mao LY, Vogel A, Arsova B, Panstruga R, Fei ZJ, Rose JKC, Zamir D, Carrari F, Giovannoni JJ, Weigel D, Usadel B, Fernie AR. 2014. The genome of the stress-tolerant wild tomato species *Solanum pennellii*. *Nature Genetics* **46**, 1034-1038.
Boutrot F, Zipfel C. 2017. Function, discovery, and exploitation of plant pattern recognition receptors for broad-spectrum disease resistance. *Annual Review of Phytopatholy* **55**, 257-286.
Bowman JL, Kohchi T, Yamato KT, Jenkins J, Shu SQ, Ishizaki K, Yamaoka S, Nishihama R, Nakamura Y, Berger F, Adam C, Aki SS, Althoff F, Araki T, Arteaga-Vazquez MA, Balasubrmanian S, Barry K, Bauer D, Boehm CR, Briginshaw L, Caballero-Perez J, Catarino B, Chen F, Chiyoda S, Chovatia M, Davies KM, Delmans M, Demura T, Dierschke T, Dolan L, Dorantes-Acosta AE, Eklund DM, Florent SN, Flores-Sandoval E, Fujiyama A, Fukuzawa H, Galik B, Grimanelli D, Grimwood J, Grossniklaus U, Hamada T, Haseloff J, Hetherington AJ, Higo A, Hirakawa Y, Hundley HN, Ikeda Y, Inoue K, Inoue SI, Ishida S, Jia QD, Kakita M, Kanazawa T, Kawai Y, Kawashima T, Kennedy M, Kinose K, Kinoshita T, Kohara Y, Koide E, Komatsu K, Kopischke S, Kubo M, Kyozuka J, Lagercrantz U, Lin SS, Lindquist E, Lipzen AM, Lu CW, De Luna E, Martienssen RA, Minamino N, Mizutani M, Mizutani M, Mochizuki N, Monte I, Mosher R, Nagasaki H, Nakagami H, Naramoto S, Nishitani K, Ohtani M, Okamoto T, Okumura M, Phillips J, Pollak B, Reinders A, Rovekamp M, Sano R, Sawa S, Schmid MW, Shirakawa M, Solano R, Spunde A, Suetsugu N, Sugano S, Sugiyama A, Sun R, Suzuki Y, Takenaka M, Takezawa D, Tomogane H, Tsuzuki M, Ueda T, Umeda M, Ward JM, Watanabe Y, Yazaki K, Yokoyama R, Yoshitake Y, Yotsui I, Zachgo S, Schmutz J. 2017. Insights into land plant evolution garnered from the *Marchantia polymorpha* genome. *Cell* **171**, 287-304.

Braasch I, Salzburger W, Meyer A. 2006. Asymmetric evolution in two fish-specifically duplicated receptor tyrosine kinase paralogons involved in teleost coloration. *Molecular Biology and Evolution* **23**, 1192-1202.

Brunkard JO, Zambryski PC. 2017. Plasmodesmata enable multicellularity: new insights into their evolution, biogenesis, and functions in development and immunity. *Current Opinion in Plant Biology* **35**, 76-83.

Burdiak P, Rusaczonek A, Witoń D, Głów D, Karpiński S. 2015. Cysteine-rich receptor-like kinase CRK5 as a regulator of growth, development, and ultraviolet radiation responses in *Arabidopsis thaliana*. *Journal of Experimental Botany* **66**, 3325-3337.

Caillaud MC, Wirthmueller L, Sklenar J, Findlay K, Piquerez SJM, Jones AME, Robatzek S, Jones JDG, Faulkner C. 2014. The plasmodesmal protein PDLP1 localises to haustoria-associated membranes during downy mildew infection and regulates callose deposition. *Plos Pathogens* **10,** e1004496.

Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Sánchez Alvarado A, Yandell M. 2008. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research* **18**, 188-196.

Carvunis AR, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, Charloteaux B, Hidalgo CA, Barbette J, Santhanam B, Brar GA, Weissman JS, Regev A, Thierry-Mieg N, Cusick ME, Vidal M. 2012. Proto-genes and *de novo* gene birth. *Nature* **487**, 370-374.

Chen FC, Chen CJ, Li WH, Chuang TJ. 2010. Gene family size conservation is a good indicator of evolutionary rates. *Molecular Biology and Evolution* **27**, 1750-1758.

Chen K, Du L, Chen Z. 2003. Sensitization of defense responses and activation of programmed cell death by a pathogen-induced receptor-like protein kinase in *Arabidopsis*. *Plant Molecular Biology* **53**, 61-74.

Chen K, Fan B, Du L, Chen Z. 2004. Activation of hypersensitive cell death by pathogen-induced receptor-like protein kinases from *Arabidopsis*. *Plant Molecular Biology* **56**, 271-283.

Chen Z. 2001. A superfamily of proteins with novel cysteine-rich repeats. *Plant Physiology* **126**, 473-476.

Chern M, Xu QF, Bart RS, Bai W, Ruan DL, Sze-To WH, Canlas PE, Jain R, Chen XW, Ronald PC. 2016. A genetic screen identifies a requirement for cysteine-rich-receptor-like kinases in rice NH1 (OsNPR1)-mediated immunity. *Plos Genetics* **12**, e1006182.

Chinchilla D, Zipfel C, Robatzek S, Kemmerling B, Nürnberger T, Jones JD, Felix G, Boller T. 2007. A flagellin-induced complex of the receptor FLS2 and BAK1 initiates plant defence. *Nature* **448**, 497-500.

Cock JM, Vanoosthuyse V, Gaude T. 2002. Receptor kinase signalling in plants and animals: distinct molecular systems with mechanistic similarities. *Current Opinion in Cell Biology* **14**, 230-236.

Coen E, Strachan T, Dover G. 1982. Dynamics of concerted evolution of ribosomal DNA and histone gene families in the *melanogaster* species subgroup of *Drosophila*. *Journal of Molecular Biology* **158**, 17-35.

Cooper GM, Brown CD. 2008. Qualifying the relationship between sequence conservation and molecular function. *Genome Research* **18**, 201-205.

Couto D, Zipfel C. 2016. Regulation of pattern recognition receptor signalling in plants. *Nature Reviews Immunology* **16**, 537-552.

Cui W, Lee JY. 2016. *Arabidopsis* callose synthases CalS1/8 regulate plasmodesmal permeability during stress. *Nature Plants* **2**, 16034.

Dardick C, Chen J, Richter T, Ouyang S, Ronald P. 2007. The rice kinase database. A phylogenomic database for the rice kinome. *Plant Physiology* **143**, 579-586.

De Smet I, Voss U, Jürgens G, Beeckman T. 2009. Receptor-like kinases shape the plant. *Nature Cell Biology* **11**, 1166-1173.

De Smet R, Adams KL, Vandepoele K, van Montagu MCE, Maere S, Van de Peer Y. 2013. Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 2898-2903.

De Storme N, Geelen D. 2014. Callose homeostasis at plasmodesmata: molecular regulators and developmental relevance. *Frontiers in Plant Science* **5**, 138.

Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. 1999. Improved microbial gene identification with GLIMMER. *Nucleic Acids Research* **27**, 4636-4641.

Dominguez Del Angel V, Hjerde E, Sterck L, Capella-Gutierrez S, Notredame C, Vinnere Pettersson O, Amselem J, Bouri L, Bocs S, Klopp C, Gibrat JF, Vlasova A, Leskosek BL, Soler L, Binzer-Panchal M, Lantz H. 2018. Ten steps to get started in genome sssembly and annotation. *F1000Research* 7, ELIXIR-148.

Eirin-Lopez JM, Rebordinos L, Rooney AP, Rozas J. 2012. The birth-and-death evolution of multigene families revisited. *Repetitive DNA* 7, 170-196.

El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer ELL, Hirsh L, Paladin L, Piovesan D, Tosatto SCE, Finn RD. 2019. The Pfam protein families database in 2019. *Nucleic Acids Research* **47**, D427-D432.

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**, 783-791.

Feng W, Kita D, Peaucelle A, Cartwright HN, Doan V, Duan QH, Liu MC, Maman J, Steinhorst L, Schmitz-Thom I, Yvon R, Kudla J, Wu HM, Cheung AY, Dinneny JR. 2018. The FERONIA receptor kinase maintains cell-wall integrity during salt stress through Ca^{2+} signaling. *Current Biology* **28**, 666-675.

Filiault DL, Ballerini ES, Mandakova T, Akoz G, Derieg NJ, Schmutz J, Jenkins J, Grimwood J, Shu SQ, Hayes RD, Hellsten U, Barry K, Yan JY, Mihaltcheva S, Karafiatova M, Nizhynska V, Kramer EM, Lysak MA, Hodges SA, Nordborg M. 2018. The *Aquilegia* genome provides insight into adaptive radiation and reveals an extraordinarily polymorphic chromosome with a unique history. *Elife* 7, e36426.

Gaillard I, Rouquier S, Giorgi D. 2004. Olfactory receptors. *Cellular and Molecular Life Scieces* **61**, 456-469.

Goff SA, Ricke D, Lan TH, Presting G, Wang RL, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, Hadley D, Hutchinson

D, Martin C, Katagiri F, Lange BM, Moughamer T, Xia Y, Budworth P, Zhong JP, Miguel T, Paszkowski U, Zhang SP, Colbert M, Sun WL, Chen LL, Cooper B, Park S, Wood TC, Mao L, Quail P, Wing R, Dean R, Yu YS, Zharkikh A, Shen R, Sahasrabudhe S, Thomas A, Cannings R, Gutin A, Pruss D, Reid J, Tavtigian S, Mitchell J, Eldredge G, Scholl T, Miller RM, Bhatnagar S, Adey N, Rubano T, Tusneem N, Robinson R, Feldhaus J, Macalma T, Oliphant A, Briggs S. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp *japonica*). *Science* **296**, 92-100.

Gómez-Gómez L, Boller T. 2000. FLS2: an LRR receptor-like kinase involved in the perception of the bacterial elicitor flagellin in *Arabidopsis*. *Molecular Cell* **5**, 1003-1011.

Greeff C, Roux M, Mundy J, Petersen M. 2012. Receptor-like kinase complexes in plant innate immunity. *Frontiers in Plant Sciece* **3**, 209.

Guo YL. 2013. Gene family evolution in green plants with emphasis on the origination and evolution of *Arabidopsis thaliana* genes. *Plant Journal* **73**, 941-951.

Hanks SK, Quinn AM, Hunter T. 1988. The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. *Science* **241**, 42-52.

Haruta M, Sabat G, Stecker K, Minkoff BB, Sussman MR. 2014. A peptide hormone and its receptor protein kinase regulate plant cell expansion. *Science* **343**, 408-411.

Hirakawa H, Shirasawa K, Miyatake K, Nunome T, Negoro S, Ohyama A, Yamaguchi H, Sato S, Isobe S, Tabata S, Fukuoka H. 2014. Draft genome sequence of eggplant (*Solanum melongena* L.): the representative *Solanum* species indigenous to the old world. *DNA Research* **21**, 649-660.

Hori K, Maruyama F, Fujisawa T, Togashi T, Yamamoto N, Seo M, Sato S, Yamada T, Mori H, Tajima N, Moriyama T, Ikeuchi M, Watanabe M, Wada H, Kobayashi K, Saito M, Masuda T, Sasaki-Sekimoto Y, Mashiguchi K, Awai K, Shimojima M, Masuda S, Iwai M, Nobusawa T, Narise T, Kondo S, Saito H, Sato R, Murakawa M, Ihara Y, Oshima-Yamada Y, Ohtaka K, Satoh M, Sonobe K, Ishii M, Ohtani R, Kanamori-Sato M, Honoki R, Miyazaki D, Mochizuki H, Umetsu J, Higashi K, Shibata D, Kamiya Y, Sato N, Nakamura Y, Tabata S, Ida S, Kurokawa K, Ohta H. 2014. *Klebsormidium flaccidum* genome reveals primary factors for plant terrestrial adaptation. *Nature Communications* **5**, 3978.

Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H, Haberer G, Hollister JD, Ossowski S, Ottilar RP, Salamov AA, Schneeberger K, Spannagl M, Wang X, Yang L, Nasrallah ME, Bergelson J, Carrington JC, Gaut BS, Schmutz J, Mayer KF, Van de Peer Y, Grigoriev IV, Nordborg M, Weigel D, Guo YL. 2011. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nature Genetics* **43**, 476-481.

Hufton AL, Panopoulou G. 2009. Polyploidy and genome restructuring: a variety of outcomes. *Current Opinion in Genetics & Development* **19**, 600-606.

Hunter K, Kimura S, Rokka A, Tran C, Toyota M, Kukkonen JP, Wrzaczek M. 2019. CRK2 enhances salt tolerance by regulating callose deposition in connection with PLDα1. *Plant Physiology*, DOI: 10.1104/pp.19.00560.

Idänheimo N, Gauthier A, Salojärvi J, Siligato R, Brosche M, Kollist H, Mähönen AP, Kangasjärvi J, Wrzaczek M. 2014. The *Arabidopsis thaliana* cysteine-rich receptor-like kinases CRK6 and CRK7 protect against apoplastic oxidative stress. *Biochemical and Biophysical Research Communications* **445**, 457-462.

Ingram VM. 1961. Gene evolution and the haemoglobins. *Nature* **189**, 704-708.

Inoue J, Sato Y, Sinclair R, Tsukamoto K, Nishida M. 2015. Rapid genome reshaping by multiple-gene loss after whole-genome duplication in teleost fish suggested by mathematical modeling. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 14918-14923.

International Brachypodium Initiative. 2010. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **463**, 763-768.

International Peach Genome Initiative. 2013. The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nature Genetics* **45**, 487-494.

Iskratsch T, Braun A, Paschinger K, Wilson IBH. 2009. Specificity analysis of lectins and antibodies using remodeled glycoproteins. *Analytical Biochemistry* **386**, 133-146.

Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, Vezzi A, Legeai F, Hugueney P, Dasilva C, Horner D, Mica E, Jublot D, Poulain J, Bruyere C, Billault A, Segurens B, Gouyvenoux M, Ugarte E, Cattonaro F, Anthouard V, Vico V, Del Fabbro C, Alaux M, Di Gaspero G, Dumas V, Felice N, Paillard S, Juman I, Moroldo M, Scalabrin S, Canaguier A, Le Clainche I, Malacrida G, Durand E, Pesole G, Laucou V, Chatelet P, Merdinoglu D, Delledonne M, Pezzotti M, Lecharny A, Scarpelli C, Artiguenave F, Pe ME, Valle G, Morgante M, Caboche M, Adam-Blondon AF, Weissenbach J, Quetier F, Wincker P, Public F-I. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463-465.

Jensen LJ, Ussery DW, Brunak S. 2003. Functionality of system components: Conservation of protein function in protein feature space. *Genome Research* **13**, 2444-2449.

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS, Soltis DE, Clifton SW, Schlarbaum SE, Schuster SC, Ma H, Leebens-Mack J, dePamphilis CW. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**, 97-100.

Jupe F, Witek K, Verweij W, Sliwka J, Pritchard L, Etherington GJ, Maclean D, Cock PJ, Leggett RM, Bryan GJ, Cardle L, Hein I, Jones JD. 2013. Resistance gene enrichment sequencing (RenSeq) enables reannotation of the NB-LRR gene family from sequenced plant genomes and rapid mapping of resistance loci in segregating populations. *Plant Journal* **76**, 530-544.

Katju V, Bergthorsson U. 2013. Copy-number changes in evolution: rates, fitness effects and adaptive significance. *Frontiers in Genetics* **4**, 273.

Katju V, Lynch M. 2006. On the formation of novel genes by duplication in the *Caenorhabditis elegans* genome. *Molecular Biology and Evolution* **23**, 1056-1067.

Kelley LA, Sternberg MJ. 2015. Partial protein domains: evolutionary insights and bioinformatics challenges. *Genome Biology* **16**, 100.

Kersting AR, Bornberg-Bauer E, Moore AD, Grath S. 2012. Dynamics and adaptive benefits of protein domain emergence and arrangements during plant genome evolution. *Genome Biology and Evolution* **4**, 316-329.

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

Klasberg S, Bitard-Feildel T, Callebaut I, Bornberg-Bauer E. 2018. Origins and structural properties of novel and *de novo* protein domains during insect evolution. *The FEBS Journal* **285**, 2605-2625.

Klaus-Heisen D, Nurisso A, Pietraszewska-Bogiel A, Mbengue M, Camut S, Timmers T, Pichereaux C, Rossignol M, Gadella TWJ, Imberty A, Lefebvre B, Cullimore JV. 2011. Structure-function similarities between a plant receptor-like kinase and the human interleukin-1 receptor-associated kinase-4. *Journal of Biological Chemistry* **286**, 11202-11210.

Kleinjan DA, Bancewicz RM, Gautier P, Dahm R, Schonthaler HB, Damante G, Seawright A, Hever AM, Yeyati PL, van Heyningen V, Coutinho P. 2008. Subfunctionalization of duplicated zebrafish *pax6* genes by *cis*-regulatory divergence. *PLoS Genetics* **4**, e29.

Kolar F, Certner M, Suda J, Schonswetter P, Husband BC. 2017. Mixed-ploidy species: progress and opportunities in polyploid research. *Trends in Plant Sciece* **22**, 1041-1055.

Koonin EV. 2005. Orthologs, paralogs, and evolutionary genomics. *Annual Review of Genetics* **39**, 309-338.

Kornev AP, Haste NM, Taylor SS, Eyck LF. 2006. Surface comparison of active and inactive protein kinases identifies a conserved activation mechanism. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 17783-17788.

Lang D, Ullrich KK, Murat F, Fuchs J, Jenkins J, Haas FB, Piednoel M, Gundlach H, Van Bel M, Meyberg R, Vives C, Morata J, Symeonidi A, Hiss M, Muchero W, Kamisugi Y, Saleh O, Blanc G, Decker EL, van Gessel N, Grimwood J, Hayes RD, Graham SW, Gunter LE, McDaniel SF, Hoernstein SNW, Larsson A, Li FW, Perroud PF, Phillips J, Ranjan P, Rokshar DS, Rothfels CJ, Schneider L, Shu S, Stevenson DW, Thummler F, Tillich M, Villarreal Aguilar JC, Widiez T, Wong GK, Wymore A, Zhang Y, Zimmer AD, Quatrano RS, Mayer KFX, Goodstein D, Casacuberta JM, Vandepoele K, Reski R, Cuming AC, Tuskan GA, Maumus F, Salse J, Schmutz J, Rensing SA. 2018. The *Physcomitrella patens* chromosome-scale assembly reveals moss genome structure and evolution. *Plant Journal* **93**, 515-533.

Lee DS, Kim YC, Kwon SJ, Ryu CM, Park OK. 2017. The Arabidopsis cysteine-rich receptor-like kinase CRK36 regulates immunity through interaction with the cytoplasmic kinase BIK1. *Frontiers in Plant Science* **8**, 1856.

Lee JY, Wang X, Cui W, Sager R, Modla S, Czymmek K, Zybaliov B, van Wijk K, Zhang C, Lu H, Lakshmanan V. 2011. A plasmodesmata-localized protein mediates crosstalk between cell-to-cell communication and innate immunity in *Arabidopsis*. *Plant Cell* **23**, 3353-3373.

Lewin B. 2008. *Genes IX*. Sundbury, Mass.: Jones and Bartlett Publishers.

Li FW, Brouwer P, Carretero-Paulet L, Cheng SF, de Vries J, Delaux PM, Eily A, Koppers N, Kuo LY, Li Z, Simenc M, Small I, Wafula E, Angarita S, Barker MS, Brautigam A, dePamphilis C, Gould S, Hosmani PS, Huang YM, Huettel B, Kato Y, Liu X, Maere S, McDowell R, Mueller LA, Nierop KGJ, Rensing SA, Robison T, Rothfels CJ, Sigel EM, Song Y, Timilsena PR, Van de Peer Y, Wang HL, Wilhelmsson PKI, Wolf PG, Xu X, Der JP, Schluepmann H, Wong GKS, Pryer KM. 2018. Fern genomes elucidate land plant evolution and cyanobacterial symbioses. *Nature Plants* **4**, 460-472.

Li Y, Wei W, Feng J, Luo H, Pi M, Liu Z, Kang C. 2017. Genome re-annotation of the wild strawberry *Fragaria vesca* using extensive Illumina- and SMRT-based RNA-seq datasets. *DNA Research* **25**, 61–70.

Li Z, Defoort J, Tasdighian S, Maere S, Van de Peer Y, De Smet R. 2016a. Gene duplicability of core genes is highly consistent across all angiosperms. *Plant Cell* **28**, 326-344.

Li ZW, Chen X, Wu Q, Hagmann J, Han TS, Zou YP, Ge S, Guo YL. 2016b. On the origin of de novo genes in *Arabidopsis thaliana* populations. *Genome Biology and Evolution* **8**, 2190-2202.

Liang X, Zhou JM. 2018. Receptor-like cytoplasmic kinases: central players in plant receptor kinase-mediated signaling. *Annual Review of Plant Biology* **69**, 267-299.

Lim GH, Shine MB, de Lorenzo L, Yu KS, Cui WE, Navarre D, Hunt AG, Lee JY, Kachroo A, Kachroo P. 2016. Plasmodesmata localizing proteins regulate transport and signaling during systemic acquired immunity in plants. *Cell Host & Microbe* **19**, 541-549.

Liu PL, Du L, Huang Y, Gao SM, Yu M. 2017. Origin and diversification of leucine-rich repeat receptor-like protein kinase (LRR-RLK) genes in plants. *BMC Evolutionary Biology* **17**, 47.

Liu SL, Baute GJ, Adams KL. 2011. Organ and cell type-specific complementary expression patterns and regulatory neofunctionalization between duplicated genes in *Arabidopsis thaliana*. *Genome Biology and Evolution* **3**, 1419-1436.

Ma LS, Wang L, Trippel C, Mendoza-Mendoza A, Ullmann S, Moretti M, Carsten A, Kahnt J, Reissmann S, Zechmann B, Bange G, Kahmann R. 2018. The *Ustilago maydis* repetitive effector Rsp3 blocks the antifungal activity of mannose-binding maize proteins. *Nature Communicatios* **9**, 1711.

Martinez M. 2011. Plant protein-coding gene families: emerging bioinformatics approaches. *Trends in Plant Sciece* **16**, 558-567.

**Mascher M, Gundlach H, Himmelbach A, Beier S, Twardziok SO, Wicker T, Radchuk V, Dockter C, Hedley PE, Russell J, Bayer M, Ramsay L, Liu H, Haberer G, Zhang XQ, Zhang QS, Barrero RA, Li L, Taudien S, Groth M, Felder M, Hastie A, Simkova H, Stankova H,

Vrana J, Chan S, Munoz-Amatrian M, Ounit R, Wanamaker S, Bolser D, Colmsee C, Schmutzer T, Aliyeva-Schnorr L, Grasso S, Tanskanen J, Chailyan A, Sampath D, Heavens D, Clissold L, Cao SJ, Chapman B, Dai F, Han Y, Li H, Li X, Lin CY, McCooke JK, Tan C, Wang PH, Wang SB, Yin SY, Zhou GF, Poland JA, Bellgard MI, Borisjuk L, Houben A, Dolezel J, Ayling S, Lonardi S, Kersey P, Lagridge P, Muehlbauer GJ, Clark MD, Caccamo M, Schulman AH, Mayer KFX, Platzer M, Close TJ, Scholz U, Hansson M, Zhang GP, Braumann I, Spannagl M, Li CD, Waugh R, Stein N. 2017. A chromosome conformation capture ordered sequence of the barley genome. *Nature* **544**, 426-433.

Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, Witman GB, Terry A, Salamov A, Fritz-Laylin LK, Marechal-Drouard L, Marshall WF, Qu LH, Nelson DR, Sanderfoot AA, Spalding MH, Kapitonov VV, Ren QH, Ferris P, Lindquist E, Shapiro H, Lucas SM, Grimwood J, Schmutz J, Cardol P, Cerutti H, Chanfreau G, Chen CL, Cognat V, Croft MT, Dent R, Dutcher S, Fernandez E, Fukuzawa H, Gonzalez-Ballester D, Gonzalez-Halphen D, Hallmann A, Hanikenne M, Hippler M, Inwood W, Jabbari K, Kalanon M, Kuras R, Lefebvre PA, Lemaire SD, Lobanov AV, Lohr M, Manuell A, Meir I, Mets L, Mittag M, Mittelmeier T, Moroney JV, Moseley J, Napoli C, Nedelcu AM, Niyogi K, Novoselov SV, Paulsen IT, Pazour G, Purton S, Ral JP, Riano-Pachon DM, Riekhof W, Rymarquis L, Schroda M, Stern D, Umen J, Willows R, Wilson N, Zimmer SL, Allmer J, Balk J, Bisova K, Chen CJ, Elias M, Gendler K, Hauser C, Lamb MR, Ledford H, Long JC, Minagawa J, Page MD, Pan JM, Pootakham W, Roje S, Rose A, Stahlberg E, Terauchi AM, Yang PF, Ball S, Bowler C, Dieckmann CL, Gladyshev VN, Green P, Jorgensen R, Mayfield S, Mueller-Roeber B, Rajamani S, Sayre RT, Brokstein P, Dubchak I, Goodstein D, Hornick L, Huang YW, Jhaveri J, Luo YG, Martinez D, Ngau WCA, Otillar B, Poliakov A, Porter A, Szajkowski L, Werner G, Zhou KM, Grigoriev IV, Rokhsar DS, Grossman AR, Annotation C, Team JA. 2007. The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science* **318**, 245-251.

Ming R, VanBuren R, Liu YL, Yang M, Han YP, Li LT, Zhang Q, Kim MJ, Schatz MC, Campbell M, Li JP, Bowers JE, Tang HB, Lyons E, Ferguson AA, Narzisi G, Nelson DR, Blaby-Haas CE, Gschwend AR, Jiao YN, Der JP, Zeng FC, Han J, Min XJ, Hudson KA, Singh R, Grennan AK, Karpowicz SJ, Watling JR, Ito K, Robinson SA, Hudson ME, Yu QY, Mockler TC, Carroll A, Zheng Y, Sunkar R, Jia RZ, Chen N, Arro J, Wai CM, Wafula E, Spence A, Han YN, Xu LM, Zhang JS, Peery R, Haus MJ, Xiong WW, Walsh JA, Wu J, Wang ML, Zhu YJ, Paull RE, Britt AB, Du CG, Downie SR, Schuler MA, Michael TP, Long SP, Ort DR, Schopf JW, Gang DR, Jiang N, Yandell M, dePamphilis CW, Merchant SS, Paterson AH, Buchanan BB, Li SH, Shen-Miller J. 2013. Genome of the long-living sacred lotus (*Nelumbo nucifera* Gaertn.). *Genome Biology* **14**.

Miyakawa T, Hatano K, Miyauchi Y, Suwa Y, Sawano Y, Tanokura M. 2014. A secreted protein with plant-specific cysteine-rich motif functions as a mannose-binding lectin that exhibits antifungal activity. *Plant Physiology* **166**, 766-778.

Miyakawa T, Miyazono K, Sawano Y, Hatano K, Tanokura M. 2009. Crystal structure of ginkbilobin-2 with homology to the extracellular domain of plant cysteine-rich receptor-like kinases. *Proteins* **77**, 247-251.

Moore AD, Björklund AK, Ekrnan D, Bornberg-Bauer E, Elofsson A. 2008. Arrangements in the modular evolution of proteins. *Trends in Biochemical Sciences* **33**, 444-451.

Moore RC, Purugganan MD. 2003. The early stages of duplicate gene evolution. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 15682-15687.

Motamayor JC, Mockaitis K, Schmutz J, Haiminen N, Livingstone D, Cornejo O, Findley SD, Zheng P, Utro F, Royaert S, Saski C, Jenkins J, Podicheti R, Zhao MX, Scheffler BE, Stack JC, Feltus FA, Mustiga GM, Amores F, Phillips W, Marelli JP, May GD, Shapiro H, Ma JX, Bustamante CD, Schnell RJ, Main D, Gilbert D, Parida L, Kuhn DN. 2013. The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. *Genome Biology* **14**.

Nacher JC, Hayashida M, Akutsu T. 2010. The role of internal duplication in the evolution of multi-domain proteins. *Biosystems* **101**, 127-135.

Nei M, Rooney AP. 2005. Concerted and birth-and-death evolution of multigene families. *Annu Review in Genetics* **39**, 121-152.

Niu XM, Xu YC, Li ZW, Bian YT, Hou XH, Chen JF, Zou YP, Jiang J, Wu Q, Ge S, Balasubramanian S, Guo YL. 2019. Transposable elements drive rapid phenotypic variation in *Capsella rubella*. *Proceedings of the National Academy of Sciences of the United States of America* **116**, 6908-6913.

Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin YC, Scofield DG, Vezzi F, Delhomme N, Giacomello S, Alexeyenko A, Vicedomini R, Sahlin K, Sherwood E, Elfstrand M, Gramzow L, Holmberg K, Hallman J, Keech O, Klasson L, Koriabine M, Kucukoglu M, Kaller M, Luthman J, Lysholm F, Niittyla T, Olson A, Rilakovic N, Ritland C, Rossello JA, Sena J, Svensson T, Talavera-Lopez C, Theissen G, Tuominen H, Vanneste K, Wu ZQ, Zhang B, Zerbe P, Arvestad L, Bhalerao R, Bohlmann J, Bousquet J, Garcia Gil R, Hvidsten TR, de Jong P, MacKay J, Morgante M, Ritland K, Sundberg B, Thompson SL, Van de Peer Y, Andersson B, Nilsson O, Ingvarsson PK, Lundeberg J, Jansson S. 2013. The Norway spruce genome sequence and conifer genome evolution. *Nature* **497**, 579-584.

Oh MH, Wang XF, Kota U, Goshe MB, Clouse SD, Huber SC. 2009. Tyrosine phosphorylation of the BRI1 receptor kinase emerges as a component of brassinosteroid signaling in *Arabidopsis*. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 658-663.

Olender T, Lancet D, Nebert DW. 2008. Update on the olfactory receptor (OR) gene superfamily. *Human Genomics* **3**, 87-97.

Osakabe Y, Yamaguchi-Shinozaki K, Shinozaki K, Tran LS. 2013. Sensing the environment: key roles of membrane-localized kinases in plant perception and response to abiotic stress. *Journal of Experimental Botany* **64**, 445-458.

Ota T, Nei M. 1994. Divergent evolution and evolution by the birth-and-death process in the immunoglobulin V-H gene family. *Molecular Biology and Evolution* **11**, 469-482.

Palenik B, Grimwood J, Aerts A, Rouze P, Salamov A, Putnam N, Dupont C, Jorgensen R, Derelle E, Rombauts S, Zhou K, Otillar R, Merchant SS, Podell S, Gaasterland T, Napoli C, Gendler K, Manuell A, Tai V, Vallon O, Piganeau G, Jancek S, Heijde M, Jabbari K, Bowler C, Lohr M, Robbens S, Werner G, Dubchak I, Pazour GJ, Ren Q, Paulsen I, Delwiche C, Schmutz J, Rokhsar D, Van de Peer Y, Moreau H, Grigoriev IV. 2007. The tiny eukaryote *Ostreococcus* provides genomic insights into the paradox of plankton speciation. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 7705-7710.

Panchy N, Lehti-Shiu M, Shiu SH. 2016. Evolution of gene duplication in plants. *Plant Physiology* **171**, 2294-2316.

Parra G, Bradnam K, Korf I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061-1067.

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang HB, Wang XY, Wicker T, Bharti AK, Chapman J, Feltus FA, Gowik U, Grigoriev IV, Lyons E, Maher CA, Martis M, Narechania A, Otillar RP, Penning BW, Salamov AA, Wang Y, Zhang LF, Carpita NC, Freeling M, Gingle AR, Hash CT, Keller B, Klein P, Kresovich S, McCann MC, Ming R, Peterson DG, Mehboob-ur-Rahman, Ware D, Westhoff P, Mayer KFX, Messing J, Rokhsar DS. 2009. The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**, 551-556.

Pawson T, Scott JD. 2005. Protein phosphorylation in signaling- 50 years and counting. *Trends in Biochemical Sciences* **30**, 286-290.

Persi E, Wolf YI, Koonin EV. 2016. Positive and strongly relaxed purifying selection drive the evolution of repeats in proteins. *Nature Communications* **7**, 13570.

Potato Genome Sequencing Consortium. 2011. Genome sequence and analysis of the tuber crop potato. *Nature* **475**, 189-U194.

Prochnik SE, Umen J, Nedelcu AM, Hallmann A, Miller SM, Nishii I, Ferris P, Kuo A, Mitros T, Fritz-Laylin LK, Hellsten U, Chapman J, Simakov O, Rensing SA, Terry A, Pangilinan J, Kapitonov V, Jurka J, Salamov A, Shapiro H, Schmutz J, Grimwood J, Lindquist E, Lucas S, Grigoriev IV, Schmitt R, Kirk D, Rokhsar DS. 2010. Genomic analysis of organismal complexity in the multicellular green alga *Volvox carteri*. *Science* **329**, 223-226.

Qiao X, Li Q, Yin H, Qi K, Li L, Wang R, Zhang S, Paterson AH. 2019. Gene duplication and evolution in recurring polyploidization-diploidization cycles in plants. *Genome Biology* **20**, 38.

Ranf S. 2017. Sensing of molecular patterns through cell surface immune receptors. *Current Opinion in Plant Biology* **38**, 68-77.

Rastogi S, Liberles DA. 2005. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evolutionary Biology* **5**.

Rayapuram C, Jensen MK, Maiser F, Shanir JV, Hornshoj H, Rung JH, Gregersen PL, Schweizer P, Collinge DB, Lyngkjaer MF. 2012.

Regulation of basal resistance by a powdery mildew-induced cysteine-rich receptor-like protein kinase in barley. *Molecular Plant Pathology* **13**, 135-147.
Rodrigo G, Fares MA. 2018. Intrinsic adaptive value and early fate of gene duplication revealed by a bottom-up approach. *Elife* 7, e29739.
Salojärvi J, Smolander OP, Nieminen K, Rajaraman S, Safronov O, Safdari P, Lamminmäki A, Immanen J, Lan TY, Tanskanen J, Rastas P, Amiryousefi A, Jayaprakash B, Kammonen JI, Hagqvist R, Eswaran G, Ahonen VH, Serra JA, Asiegbu FO, Barajas-Lopez JD, Blande D, Blokhina O, Blomster T, Broholm S, Brosche M, Cui FQ, Dardick C, Ehonen SE, Elomaa P, Escamez S, Fagerstedt KV, Fujii H, Gauthier A, Gollan PJ, Halimaa P, Heino PI, Himanen K, Hollender C, Kangasjärvi S, Kauppinen L, Kelleher CT, Kontunen-Soppela S, Koskinen JP, Kovalchuk A, Karenlampi SO, Kärkönen AK, Lim KJ, Leppälä J, Macpherson L, Mikola J, Mouhu K, Mähönen AP, Niinemets U, Oksanen E, Overmyer K, Palva ET, Pazouki L, Pennanen V, Puhakainen T, Poczai P, Possen BJHM, Punkkinen M, Rahikainen MM, Rousi M, Ruonala R, van der Schoot C, Shapiguzov A, Sierla M, Sipilä TP, Sutela S, Teeri TH, Tervahauta AI, Vaattovaara A, Vahala J, Vetchinnikova L, Welling A, Wrzaczek M, Xu EJ, Paulin LG, Schulman AH, Lascoux M, Albert VA, Auvinen P, Helariutta Y, Kangasjärvi J. 2017. Genome sequencing and population genomic analyses provide insights into the adaptive landscape of silver birch. *Nature Genetics* **49**, 904-912.
Sankoff D, Zheng C, Wang B, Abad Najar C. 2015. Structural vs. functional mechanisms of duplicate gene loss following whole genome doubling. *BMC Bioinformatics* **16 Suppl 17**, S9.
Sato S, Nakamura Y, Kaneko T, Asamizu E, Kato T, Nakao M, Sasamoto S, Watanabe A, Ono A, Kawashima K, Fujishiro T, Katoh M, Kohara M, Kishida Y, Minami C, Nakayama S, Nakazaki N, Shimizu Y, Shinpo S, Takahashi C, Wada T, Yamada M, Ohmido N, Hayashi M, Fukui K, Baba T, Nakamichi T, Mori H, Tabata S. 2008. Genome Structure of the Legume, *Lotus japonicus*. *DNA Research* **15**, 227-239.
Tomato Genome Consortium. 2012. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**, 635-641.
Schnable PS, Ware D, Fulton RS, Stein JC, Wei FS, Pasternak S, Liang CZ, Zhang JW, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du FY, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, Chen WZ, Yan L, Higginbotham J, Cardenas M, Waligorski J, Applebaum E, Phelps L, Falcone J, Kanchi K, Thane T, Scimone A, Thane N, Henke J, Wang T, Ruppert J, Shah N, Rotter K, Hodges J, Ingenthron E, Cordes M, Kohlberg S, Sgro J, Delgado B, Mead K, Chinwalla A, Leonard S, Crouse K, Collura K, Kudrna D, Currie J, He RF, Angelova A, Rajasekar S, Mueller T, Lomeli R, Scara G, Ko A, Delaney K, Wissotski M, Lopez G, Campos D, Braidotti M, Ashley E, Golser W, Kim H, Lee S, Lin JK, Dujmic Z, Kim W, Talag J, Zuccolo A, Fan C, Sebastian A, Kramer M, Spiegel L, Nascimento L, Zutavern T, Miller B, Ambroise C, Muller S, Spooner W,

Narechania A, Ren LY, Wei S, Kumari S, Faga B, Levy MJ, McMahan L, Van Buren P, Vaughn MW, Ying K, Yeh CT, Emrich SJ, Jia Y, Kalyanaraman A, Hsia AP, Barbazuk WB, Baucom RS, Brutnell TP, Carpita NC, Chaparro C, Chia JM, Deragon JM, Estill JC, Fu Y, Jeddeloh JA, Han YJ, Lee H, Li PH, Lisch DR, Liu SZ, Liu ZJ, Nagel DH, McCann MC, SanMiguel P, Myers AM, Nettleton D, Nguyen J, Penning BW, Ponnala L, Schneider KL, Schwartz DC, Sharma A, Soderlund C, Springer NM, Sun Q, Wang H, Waterman M, Westerman R, Wolfgruber TK, Yang LX, Yu Y, Zhang LF, Zhou SG, Zhu Q, Bennetzen JL, Dawe RK, Jiang JM, Jiang N, Presting GG, Wessler SR, Aluru S, Martienssen RA, Clifton SW, McCombie WR, Wing RA, Wilson RK. 2009. The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**, 1112-1115.

Semon M, Wolfe KH. 2007. Rearrangement rate following the whole-genome duplication in teleosts. *Molecular Biology and Evolution* **24**, 860-867.

Shiu SH, Bleecker AB. 2001a. Plant receptor-like kinase gene family: diversity, function, and signaling. *Science's STKE* **2001**, re22.

Shiu SH, Bleecker AB. 2001b. Receptor-like kinases from *Arabidopsis* form a monophyletic gene family related to animal receptor kinases. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 10763-10768.

Shiu SH, Bleecker AB. 2003. Expansion of the receptor-like kinase/Pelle gene family and receptor-like proteins in Arabidopsis. *Plant Physiology* **132**, 530-543.

Shiu SH, Karlowski WM, Pan R, Tzeng YH, Mayer KF, Li WH. 2004. Comparative analysis of the receptor-like kinase family in Arabidopsis and rice. *The Plant Cell* **16**, 1220-1234.

Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210-3212.

Slotte T, Hazzouri KM, Agren JA, Koenig D, Maumus F, Guo YL, Steige K, Platts AE, Escobar JS, Newman LK, Wang W, Mandakova T, Vello E, Smith LM, Henz SR, Steffen J, Takuno S, Brandvain Y, Coop G, Andolfatto P, Hu TT, Blanchette M, Clark RM, Quesneville H, Nordborg M, Gaut BS, Lysak MA, Jenkins J, Grimwood J, Chapman J, Prochnik S, Shu SQ, Rokhsar D, Schmutz J, Weigel D, Wright SI. 2013. The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. *Nature Genetics* **45**, 831-835.

Smakowska-Luzan E, Mott GA, Parys K, Stegmann M, Howton TC, Layeghifard M, Neuhold J, Lehner A, Kong J, Grunwald K, Weinberger N, Satbhai SB, Mayer D, Busch W, Madalinski M, Stolt-Bergner P, Provart NJ, Mukhtar MS, Zipfel C, Desveaux D, Guttman DS, Belkhadir Y. 2018. An extracellular network of *Arabidopsis* leucine-rich repeat receptor kinases. *Nature* **553**, 342-346.

Sonnhammer EL, Eddy SR, Durbin R. 1997. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* **28**, 405-420.

Soucy SM, Huang JL, Gogarten JP. 2015. Horizontal gene transfer: building the web of life. *Nature Reviews Genetics* **16**, 472-482.

Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. *Bioinformatics* **24**, 637-644.

Stegmann M, Monaghan J, Smakowska-Luzan E, Rovenich H, Lehner A, Holton N, Belkhadir Y, Zipfel C. 2017. The receptor kinase FER is a RALF-regulated scaffold controlling plant immune signaling. *Science* **355**, 287-289.

Stein L. 2001. Genome annotation: From sequence to biology. *Nature Reviews Genetics* **2**, 493-503.

Stone JM, Walker JC. 1995. Plant protein kinase families and signal transduction. *Plant Physiology* **108**, 451-457.

Sun Y, Li L, Macho AP, Han Z, Hu Z, Zipfel C, Zhou JM, Chai J. 2013. Structural basis for flg22-induced activation of the *Arabidopsis* FLS2-BAK1 immune complex. *Science* **342**, 624-628.

Tanaka H, Osakabe Y, Katsura S, Mizuno S, Maruyama K, Kusakabe K, Mizoi J, Shinozaki K, Yamaguchi-Shinozaki K. 2012. Abiotic stress-inducible receptor-like kinases negatively control ABA signaling in Arabidopsis. *Plant Journal* **70**, 599-613.

Tang HB, Krishnakumar V, Bidwell S, Rosen B, Chan AN, Zhou SG, Gentzbittel L, Childs KL, Yandell M, Gundlach H, Mayer KFX, Schwartz DC, Town CD. 2014. An improved genome release (version Mt4.0) for the model legume *Medicago truncatula*. *BMC Genomics* **15**.

Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M. 2008. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Research* **18**, 1979-1990.

The Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796-815.

The Gene Ontology Consortium. 2019. The Gene Ontology resource: 20 years and still GOing strong. *Nucleic Acids Research* **47**, D330-D338.

The *Solanum lycopersicoides* Genome Consortium,
https://solgenomics.net/

Thomas CL, Bayer EM, Ritzenthaler C, Fernandez-Calvino L, Maule AJ. 2008. Specific targeting of a plasmodesmal protein affecting cell-to-cell communication. *Plos Biology* **6**, 180-190.

Toll-Riera M, Alba MM. 2013. Emergence of novel domains in proteins. *BMC Evolutionary Biology* **13**, 47.

Tör M, Lotze MT, Holton N. 2009. Receptor-mediated signalling in plants: molecular patterns and programmes. *Journal of Experimental Botany* **60**, 3645-3654.

Torii KU. 2004. Leucine-rich repeat receptor kinases in plants: structure, function, and signal transduction pathways. *International Review of Cytology* **234**, 1-46.

Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, Schein J, Sterck L, Aerts A, Bhalerao RR, Bhalerao RP, Blaudez D, Boerjan W, Brun A, Brunner A, Busov V, Campbell M, Carlson J, Chalot M, Chapman J, Chen GL, Cooper D, Coutinho PM, Couturier J, Covert S, Cronk Q, Cunningham R, Davis J, Degroeve S, Dejardin A, Depamphilis C, Detter J, Dirks B, Dubchak I, Duplessis S, Ehlting J, Ellis B, Gendler

K, Goodstein D, Gribskov M, Grimwood J, Groover A, Gunter L, Hamberger B, Heinze B, Helariutta Y, Henrissat B, Holligan D, Holt R, Huang W, Islam-Faridi N, Jones S, Jones-Rhoades M, Jorgensen R, Joshi C, Kangasjärvi J, Karlsson J, Kelleher C, Kirkpatrick R, Kirst M, Kohler A, Kalluri U, Larimer F, Leebens-Mack J, Leple JC, Locascio P, Lou Y, Lucas S, Martin F, Montanini B, Napoli C, Nelson DR, Nelson C, Nieminen K, Nilsson O, Pereda V, Peter G, Philippe R, Pilate G, Poliakov A, Razumovskaya J, Richardson P, Rinaldi C, Ritland K, Rouze P, Ryaboy D, Schmutz J, Schrader J, Segerman B, Shin H, Siddiqui A, Sterky F, Terry A, Tsai CJ, Uberbacher E, Unneberg P, Vahala J, Wall K, Wessler S, Yang G, Yin T, Douglas C, Marra M, Sandberg G, Van de Peer Y, Rokhsar D. 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**, 1596-1604.

Van de Peer Y, Maere S, Meyer A. 2009. The evolutionary significance of ancient genome duplications. *Nature Reviews Genetics* **10**, 725-732.

Van de Peer Y, Mizrachi E, Marchal K. 2017. The evolutionary significance of polyploidy. *Nature Reviews Genetics* **18**, 411-424.

Wang W, Haberer G, Gundlach H, Glässer C, Nussbaumer T, Luo MC, Lomsadze A, Borodovsky M, Kerstetter RA, Shanklin J, Byrant DW, Mockler TC, Appenroth KJ, Grimwood J, Jenkins J, Chow J, Choi C, Adam C, Cao XH, Fuchs J, Schubert I, Rokhsar D, Schmutz J, Michael TP, Mayer KFX, Messing J. 2014. The *Spirodela polyrhiza* genome reveals insights into its neotenous reduction fast growth and aquatic lifestyle. *Nature Communications* **5**, 3311 .

Veeckman E, Ruttink T, Vandepoele K. 2016. Are we there yet? Reliably estimating the completeness of plant genome sequences. *Plant Cell* **28**, 1759-1768.

Weiner J, Beaussart F, Bornberg-Bauer E. 2006. Domain deletions and substitutions in the modular protein evolution. *The FEBS Journal* **273**, 2037-2047.

Veitia RA. 2005. Gene dosage balance: deletions, duplications and dominance. *Trends in Genetics* **21**, 33-35.

Wood TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon PB, Rieseberg LH. 2009. The frequency of polyploid speciation in vascular plants. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 13875-13879.

Worden AZ, Lee JH, Mock T, Rouze P, Simmons MP, Aerts AL, Allen AE, Cuvelier ML, Derelle E, Everett MV, Foulon E, Grimwood J, Gundlach H, Henrissat B, Napoli C, McDonald SM, Parker MS, Rombauts S, Salamov A, Von Dassow P, Badger JH, Coutinho PM, Demir E, Dubchak I, Gentemann C, Eikrem W, Gready JE, John U, Lanier W, Lindquist EA, Lucas S, Mayer KFX, Moreau H, Not F, Otillar R, Panaud O, Pangilinan J, Paulsen I, Piegu B, Poliakov A, Robbens S, Schmutz J, Toulza E, Wyss T, Zelensky A, Zhou K, Armbrust EV, Bhattacharya D, Goodenough UW, Van de Peer Y, Grigoriev IV. 2009. Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes *Micromonas*. *Science* **324**, 268-272.

Wrzaczek M, Brosché M, Salojärvi J, Kangasjärvi S, Idänheimo N, Mersmann S, Robatzek S, Karpiński S, Karpińska B, Kangasjärvi J.

2010. Transcriptional regulation of the CRK/DUF26 group of Receptor-like protein kinases by ozone and plant hormones in Arabidopsis. *BMC Plant Biology* **10**, 95.

Wu Y, Zhou JM. 2013. Receptor-like kinases in plant innate immunity. *Journal of Integrated Plant Biology* **55**, 1271-1286.

Xu Y-C, Niu X-M, Li X-X, He W, Chen J-F, Zou Y-P, Wu Q, Zhang YE, Busch W, Guo Y-L. 2019. Adaptation and phenotypic diversification through loss-of-function mutations in Arabidopsis protein-coding genes. *The Plant Cell* tpc.00791.02018.

Yadeta KA, Elmore JM, Creer AY, Feng B, Franco JY, Rufian JS, He P, Phinney B, Coaker G. 2017. A cysteine-rich protein kinase associates with a membrane immune complex and the cysteine residues are required for cell death. *Plant Physiology* **173**, 771-787.

Yang Z, Rannala B. 2012. Molecular phylogenetics: principles and practice. *Nature Reviews Genetics* **13**, 303-314.

Yeh YH, Chang YH, Huang PY, Huang JB, Zimmerli L. 2015. Enhanced Arabidopsis pattern-triggered immunity by overexpression of cysteine-rich receptor-like kinases. *Frontiers in Plant Scieces* **6**, 322.

Young ND, Debelle F, Oldroyd GED, Geurts R, Cannon SB, Udvardi MK, Benedito VA, Mayer KFX, Gouzy J, Schoof H, Van de Peer Y, Proost S, Cook DR, Meyers BC, Spannagl M, Cheung F, De Mita S, Krishnakumar V, Gundlach H, Zhou SG, Mudge J, Bharti AK, Murray JD, Naoumkina MA, Rosen B, Silverstein KAT, Tang HB, Rombauts S, Zhao PX, Zhou P, Barbe V, Bardou P, Bechner M, Bellec A, Berger A, Berges H, Bidwell S, Bisseling T, Choisne N, Couloux A, Denny R, Deshpande S, Dai XB, Doyle JJ, Dudez AM, Farmer AD, Fouteau S, Franken C, Gibelin C, Gish J, Goldstein S, Gonzalez AJ, Green PJ, Hallab A, Hartog M, Hua A, Humphray SJ, Jeong DH, Jing Y, Jocker A, Kenton SM, Kim DJ, Klee K, Lai HS, Lang CT, Lin SP, Macmil SL, Magdelenat G, Matthews L, McCorrison J, Monaghan EL, Mun JH, Najar FZ, Nicholson C, Noirot C, O'Bleness M, Paule CR, Poulain J, Prion F, Qin BF, Qu CM, Retzel EF, Riddle C, Sallet E, Samain S, Samson N, Sanders I, Saurat O, Scarpelli C, Schiex T, Segurens B, Severin AJ, Sherrier DJ, Shi RH, Sims S, Singer SR, Sinharoy S, Sterck L, Viollet A, Wang BB, Wang KQ, Wang MY, Wang XH, Warfsmann J, Weissenbach J, White DD, White JD, Wiley GB, Wincker P, Xing YB, Yang LM, Yao ZY, Ying F, Zhai JX, Zhou LP, Zuber A, Denarie J, Dixon RA, May GD, Schwartz DC, Rogers J, Quetier F, Town CD, Roe BA. 2011. The *Medicago* genome provides insight into the evolution of rhizobial symbioses. *Nature* **480**, 520-524.

Zhang L, Tian LH, Zhao JF, Song Y, Zhang CJ, Guo Y. 2009. Identification of an apoplastic protein involved in the initial phase of salt stress response in rice root by two-dimensional electrophoresis. *Plant Physiology* **149**, 916-928.

Zhou Q, Zhang G, Zhang Y, Xu S, Zhao R, Zhan Z, Li X, Ding Y, Yang S, Wang W. 2008. On the origin of new genes in *Drosophila*. *Genome Research* **18**, 1446-1455.

Zimin A, Stevens KA, Crepeau M, Holtz-Morris A, Koriabine M, Marcais G, Puiu D, Roberts M, Wegrzyn JL, de Jong PJ, Neale DB,

Salzberg SL, Yorke JA, Langley CH. 2014. Sequencing and assembly of the 22-Gb loblolly pine genome. *Genetics* **196**, 875-890.

Zulawski M, Schulze G, Braginets R, Hartmann S, Schulze WX. 2014. The Arabidopsis Kinome: phylogeny and evolutionary insights into functional diversification. *BMC Genomics* **15**, 548.