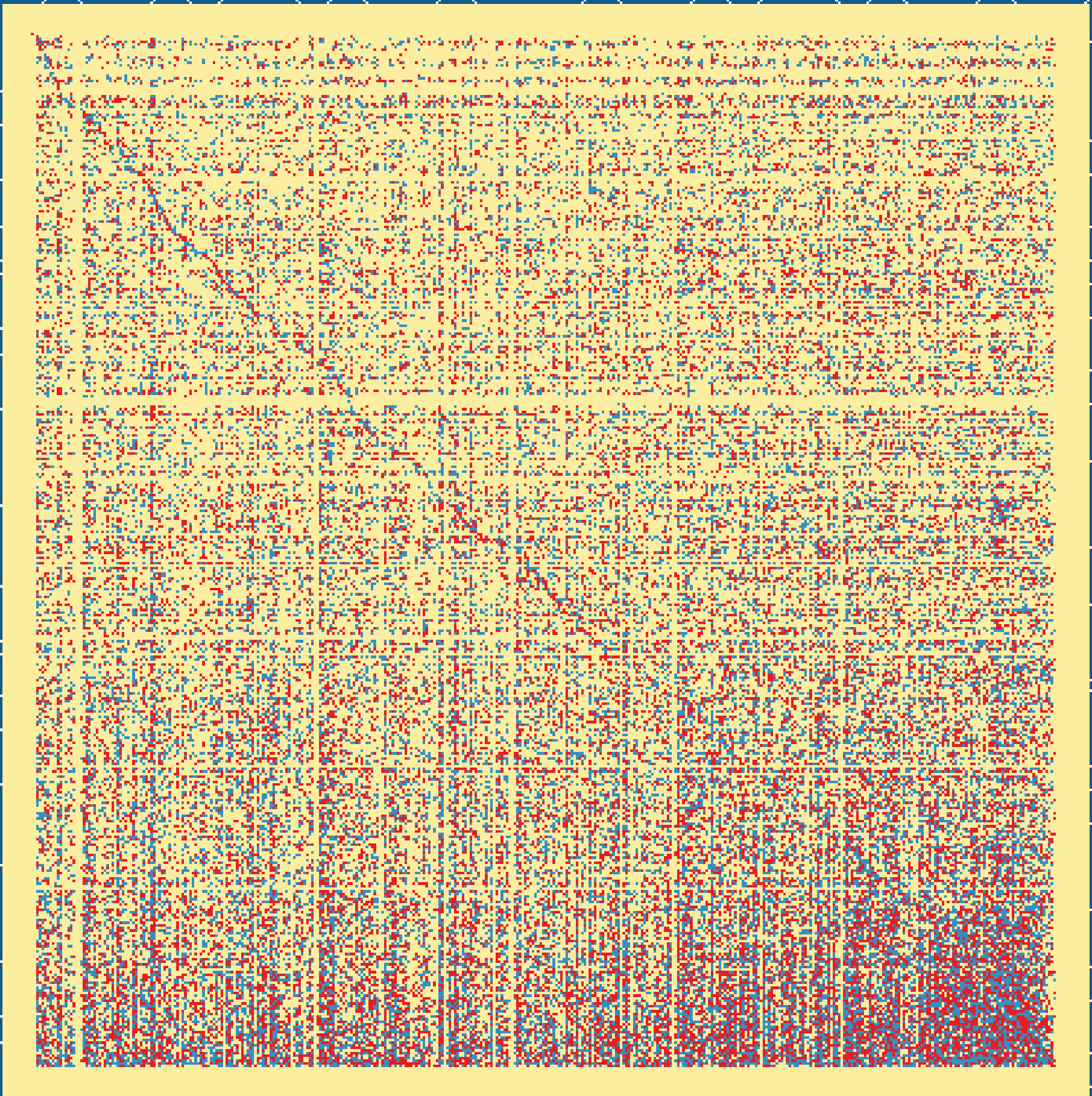
Comparative and Functional Genomics of Nitrogen-Fixing Rhizobium Symbiosis in Plants

Rens Holmer



Propositions

1. The drawbacks of nitrogen-fixing root-nodule symbiosis do not outweigh the benefits.
   (this thesis)

2. Inferring transcriptional networks from transcriptome data in plants is science fiction.
   (this thesis)

3. Given the omnipresence of big datasets in modern biological research, scientific programming is an underrepresented aspect of the educational curriculum.

4. Whereas interdisciplinary research is often celebrated for its achievements, the risk that it can result in insufficient representation of its constituent disciplines is equally often neglected.

5. Cycling to work is the most time-efficient way of commuting.

6. COVID-19 vaccination can only be effective when implemented at a global scale.

Propositions belonging to the thesis, entitled

Comparative and functional genomics of nitrogen-fixing rhizobium symbiosis in plants

Rens Holmer
Wageningen, 14 June 2021

Comparative and Functional Genomics of Nitrogen-Fixing Rhizobium Symbiosis in Plants

Rens Holmer

Thesis committee

Promotor
Prof. Dr A.H.J. Bisseling
Professor of Molecular Biology
Wageningen University & Research

Co-promotors
Dr R. Geurts
Associate professor, Laboratory of Molecular Biology
Wageningen University & Research

Dr S. Smit
Assistant professor, Bioinformatics Group
Wageningen University & Research

Other members
Dr P.-M. Delaux, Université de Toulouse, France
Prof. Dr L. Bentsink, Wageningen University & Research
Dr S. Simon, Wageningen University & Research
Dr T.E.P.M.F. Abeel, Delft University of Technology, the Netherlands

This research was conducted under the auspices of the Graduate School Experimental Plant Sciences

Comparative and Functional Genomics of Nitrogen-Fixing Rhizobium Symbiosis in Plants

Rens Holmer

Thesis
submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Monday 14 June 2021
at 11 a.m. in the Aula.

Rens Holmer
Comparative and Functional Genomics of Nitrogen-Fixing Rhizobium Symbiosis in Plants

PhD thesis, Wageningen University, Wageningen, The Netherlands (2021)
With references, with summary in English

ISBN 978-94-6395-793-9
DOI `https://doi.org/10.18174/545827`

Contents

Chapter 1

General Introduction

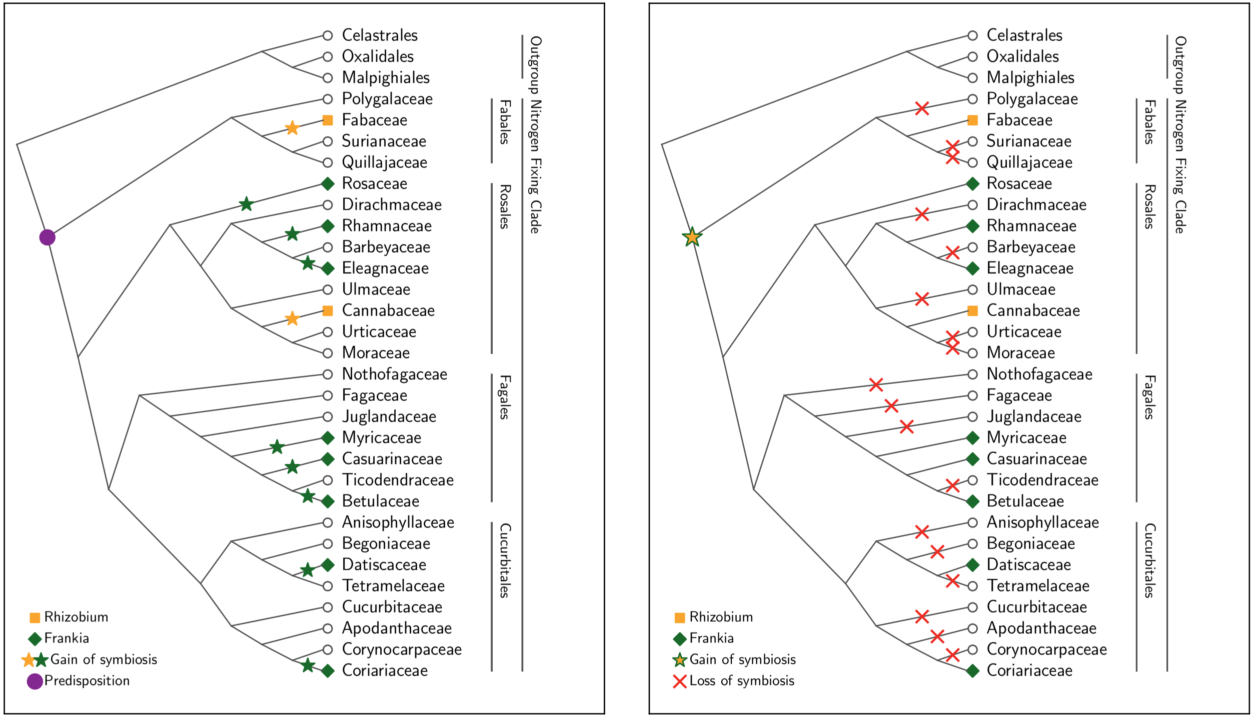
In a world where every living organism is continuously challenged by pathogenic microorganisms, one striking natural phenomenon is the symbiosis between nitrogen-fixing bacteria and plants. This mutualistic relationship takes place inside the cells of the root nodule: a dedicated plant root organ that can house millions of bacteria. By the definition of mutualism, both partners must benefit from their interaction [1]. In case of the mutualistic interaction between plants and nitrogen-fixing bacteria, the plant receives ammonia – a compound that plants cannot produce themselves – from the bacteria. In turn, the bacteria receive carbohydrates as energy source from the plant. A well-studied example of this nodulation is the relationship between legumes and rhizobium bacteria, where much of the bacteria-supplied nitrogen is used in making for instance protein-rich beans [2] or leaves [3]. More importantly, the intracellular symbiosis between plants and nitrogen-fixing bacteria makes these plants independent of exogenous nitrogen sources.

The independence of nodulating plant species from external nitrogen makes nitrogen-fixing symbiosis an inherently interesting trait for agriculture. With 68% of the worlds total nitrogen usage spent on chemical fertilizer in 2020, and demand increasing faster than supply, reducing fertilizer requirement is a key component of future sustainable agriculture [4]. With only a limited number of species capable of forming nitrogen-fixing symbiosis, it has been a longstanding goal of synthetic biology to engineer nodulation in crops other than legumes [5–8].

For some time it is believed that requirements for engineering nodulation can be learned by observing the natural world [9]. Since not all plant species can form nodules, it should be possible to identify what it takes to form nodules by comparing nodulating with non-nodulating species. More specifically, understanding how nodulation evolved naturally should lead to an engineering strategy that can turn a non-nodulating plant into a nodulating one. Therefore, this thesis studies the evolution of nodulation.

1.1 Evolution of nodulation

Based on their type of microsymbiont, nodulating plants species can be categorized into two groups: those that form nodules with *Frankia* bacteria (known as actinorhizal plants), and those that form nodules with rhizobium bacteria. Both groups of nodulating plants are confined to the same limited phylogenetic distribution: be it with *Frankia* or rhizobium bacteria, nodulating plant species only occur in the Fagales, Fabales, Rosales, and Cucurbitales orders (figure 1.1) [9]. Although the clade formed by these



(a) Multiple independent gains

(b) Single gain, massive loss

Figure 1.1: Two competing hypotheses on the evolutionary origin of nodulation. All plant species that form nitrogen-fixing root nodules are found in the nitrogen-fixing clade. Species that form nodules with *Frankia* and rhizobium are found in respectively eight and two families. Due to the absence of nodulation in many lineages within the nitrogen-fixing clade two competing hypothesis on how nodulation evolved exist. Either nodulation evolved several times, possibly preceded by a genetic predisposition limited to the nitrogen-fixing clade, or nodulation originated once and was subsequently lost on a large scale. Phylogenetic tree based on Sun et al. [10].

four orders is often referred to as the nitrogen-fixing clade, there are many lineages of plant species within these four orders that cannot form such a symbiotic relationship. The ones that do can roughly be grouped into ten clades, eight of which form a symbiosis with *Frankia* and two that form a symbiosis with rhizobia [11].

The scattered phylogenetic distribution of nodulation within the nitrogen-fixing clade has previously resulted in the adoption of an evolutionary hypothesis that assumes multiple independent gains of nodulation, possibly preceded by an unobserved gain of predisposition at the root of the nitrogen-fixing clade (figure 1.1a) [9, 12–15]. The alternative notion of a single gain of symbiosis at the root of the nitrogen-fixing clade followed by widespread loss (figure 1.1b) has been briefly discussed in Soltis et al. [9] and Vessey et al. [16]. Since this alternative single-gain hypothesis requires many more evolutionary events it has generally been discarded for not being parsimonious and therefore unlikely. However, none of the previous studies have directly investigated how the molecular mechanisms required for nodulation have evolved. Instead, nodulation has been treated as a single character, albeit with a variety of morphologies, and the molecular characteristics of the supposed predisposition have remained abstract. As such, existing hypotheses on the evolution of nodulation have remained detached from the involved molecular mechanisms. Nevertheless, the ability of a plant to form nodules is ultimately encoded in its genome, and nodulation evolved through genetic change. Therefore, to mechanistically understand how nodulation evolved it is imperative to understand what genetic changes enabled the ancestor(s) of current nodulating plant species to first form nodules. To achieve this molecular perspective on the evolution of nodulation, a comparative genome-wide analysis is required.

Historically, the legume-rhizobium symbiosis has received most attention for two main reasons: all commercially interesting plant species that can form nodules with rhizobium bacteria are legumes, and the legume-rhizobium symbiosis has proven to be more amenable to various kinds of experimental work than the actinorhizal symbiosis. Additionally, genome sequencing initiatives of model legumes such as *Lotus japonicus* [17], *Medicago truncatula* [18], and *Glycine max* [19] have accelerated research on molecular aspects of rhizobium symbiosis. As a result, the majority of what is known about molecular mechanisms in nitrogen-fixing plant-microbe symbiosis is derived from work on legume-rhizobium symbiosis.

Apart from legumes, there is one other clade of plant species within the nitrogen-fixing clade that can form nodules with rhizobium bacteria (figure 1.1). All species in the genus *Parasponia* (family Cannabaceae) engage in nitrogen-fixing symbiotis with rhizobium bacteria, similar to legumes [20]. Given that the most recent common ancestor of legumes and Cannabaceae existed $\sim$100 million years ago, *Parasponia* is a key system for studying the evolution of nodulation in the nitrogen-fixing clade [21]. Additionally, *Trema*, the most closely related sister genus to *Parasponia*, does not form nodules with rhizobium bacteria. This close relationship between nodulating and non-nodulating plant species reinforces the question whether nodulation evolved through a single gain and multiple losses, or whether there were multiple independent gains. Identifying the molecular mechanisms with which *Parasponia* engages in rhizobium symbiosis is crucial for broadening our understanding of how nodulation evolved. This makes *Parasponia* and *Trema* ideal candidates for investigating what differentiates nodulating from non-nodulating plants, from both evolutionary and engineering perspectives.

1.2 Molecular mechanisms involved in nodulation

Whether or not a plant can form a symbiotic nitrogen-fixing relationship with bacteria is determined by a range of intricate interactions at the molecular level. In case of the legume-rhizobium symbiosis it is generally understood that several steps are required to form the intracellular mutualism. The following section provides a concise summary of the molecular interactions involved in establishing the legume-rhizobium symbiosis[1].

Initially, plant-secreted flavonoids activate bacterial transcription of several genes, encoding enzymes, that together produce lipochitooligosaccharides (LCOs) [22]. In turn, these bacterial LCOs are recognized by the plant through a heteromeric protein complex containing plant lysin motif (LysM) containing receptor-like kinases. In *Medicago truncatula* this receptor complex includes *LYSM DOMAIN CONTAINING RECEPTOR KINASE3* (*MtLYK3*) and *NOD FACTOR PERCEPTION* (*MtNFP*) [23, 24]. Likewise, in *Lotus japonicus*, *NOD FACTOR RECECTOR1* (*LjNFR1*) and *LjNFR5* fulfill a similar role [25–27]. After LCO perception, various downstream targets such as other kinases are phosphorylated [28]. Subsequently, a key stage in the symbiotic signalling is the initiation of nuclear calcium spiking, a process where the concentration of Ca^{2+} ions in the nucleus rapidly fluctuates [29]. Calcium spiking is perceived by the *CALCIUM AND CALMODULIN DEPENDENT KINASE* (*CCamK*) [30], which in turn

[1] Chapter 2 of this thesis provides an extensive review of known molecular mechanisms in rhizobium symbiosis, along with known similar mechanisms in other plant-microbe mutualisms.

can activate a cascade of transcriptional regulators that initiate cell division leading to root nodule formation [31–34]. Ultimately, the symbiotic bacteria almost fully colonize the root nodules, where the plant produces haemoglobin to facilitate oxygen homeostasis [35]. An oxygen-poor environment is needed so that the bacterial nitrogenase enzyme can turn environmental N_2 into ammonia [36]. After diffusion out of the bacterial cells, ammonia is protonated into ammonium which acts as a nitrogen source for the plant. In turn, the plant provides sugars as a carbon source for the bacteria to maintain the symbiotic interaction (see Udvardi and Poole [37] for a comprehensive review on metabolism and transport in the rhizobium-legume symbiosis).

It should be noted that symbiotic signalling is not a linear process, as it includes feedback mechanisms [38] and can be modulated by environmental conditions such as nitrate availability [39]. Furthermore, from the variety of indispensable transcriptional regulators it has become clear that one or more transcriptional networks underlie rhizobium symbiosis [40]. Unfortunately, such systems involving feedback mechanisms are notoriously difficult to study and engineer [41]. Indicative of current understanding of the genetic differences between symbiotic and non-symbiotic plants is a recent engineering attempt [42]. There, the transfer of eight genes crucial for nodule formation in legumes to non-nodulating plants did not result in any symbiotic phenotype. From this unsuccessful engineering attempt it has become clear that it is not sufficient to only know which genes are involved in nodulation, but that it is equally important to know when and where these nodulation genes must be expressed. This notion is supported by recent work on the symbiotic transcriptional regulator *NODULE INCEPTION* (*NIN*), where elements essential for the symbiotic functioning of NIN are conserved in at least nine legume species [43]. As such, evolutionary studies aimed at understanding nitrogen-fixing symbiosis will have to include a regulatory perspective.

1.3 Common symbiosis signalling pathway

Previously, it has been noted that large parts of the molecular signalling mechanisms in rhizobium symbiosis are co-opted from the arbuscular mycorrhizal (AM) symbiosis [44]. For instance, both bacterial and fungal LCOs and other chitinous compounds secreted by AM fungi are recognized by members of the same LysM receptor family [45, 46]. In addition, orthologous[2] genes are responsible for recognizing bacterial LCOs in several legumes and *Parasponia* [47]. This suggests some shared evolutionary origin between symbiotic LCO perception in rhizobium and AM symbiosis, although in extant lineages specific receptors differ between rhizobium and AM symbiosis. Downstream of LCO perception, *SYMBIOTIC RECEPTOR KINASE* (*SYMRK*) is required for symbiotic interactions with arbusculary mycorrhizae, rhizobia, and *Frankia* [48]. Likewise, CCamK is required for the readout of calcium spiking in both rhizobium and AM symbiosis [30]. Furthermore, *CYCLOPS*, one of the downstream transciptional regulators that is required for rhizobium symbiosis, is also required for AM symbiosis [31]. Taken together, these conserved molecular mechanisms raise the question whether multiple indendent evolutionary gains of nodulation are still an accurate hypothesis.

From an evolutionary perspective, there is less uncertainty about the origin of AM symbiosis than there is about the origin of nodulation. Whereas only a limited number of species can form a nitrogen-fixing symbiosis with bacteria, roughly 80% of land plants can form a symbiotic relationship with arbuscular mycorrhizal fungi [49]. Given this ubiquitous presence of AM symbiosis in land plants, it is generally believed that non-AM-symbiotic species have lost the ability to form the symbiosis [50]. In line with this, recent phylogenomic studies identified multiple genes that are consistently lost in non-AM-symbiotic species [51, 52]. Whereas some of the identified genes were known to be involved in AM symbiosis, knockout mutants of several new candidates were found to be impaired in AM symbiosis [52]. This highlights the notion that mechanistic insight into a trait can be gained from understanding its molecular evolution. To resolve the molecular evolutionary trajectory of nodulation, the success of phylogenomic approaches in the arbuscular mycorrhizal symbiosis can serve as inspiration.

1.4 Techniques and technology

Identifying genetic differences between nodulating and non-nodulating plant species is impossible without genome sequence information. Whereas this seems obvious, it has taken decades of research to develop the technologies that generated the data that is used in this thesis [53]. In

[2] Two or more homologous genes from distinct species are orthologous when they are derived from a single gene in the most recent common ancestor and are therefore the result of a speciation event. In contrast, two or more genes are paralogous when they arose after a gene-duplication event.

fact, our understanding of the molecular mechanisms underlying the natural world has always been driven by the use of new technology. Without Franklin and Goslings X-ray diffraction images [54], Watson and Crick would not have known about the structure of DNA [55]. Without significant technological innovation, Sanger would not have been able to determine the amino acid sequence of insulin [56, 57]. Given that the molecular world is unobservable by eye, we rely exclusively on technology to study it.

The ability to determine the DNA sequence of an organism is like reading its molecular blueprint. As such, DNA sequencing technologies unlock the capacity to study the structure, organization, and evolution of an organism's genome. Whereas DNA sequencing has been possible for nearly four decades [58], initially the process was limited to processing the equivalent of a single gene per day [53]. Notwithstanding this low throughput, sequencing technologies have been intricately linked with computational methods from the start. Examples include processing fragmented raw sequencing data into an assembled genome [59], or the comparative analysis of gene copies in multiple species [60]. With the price per sequenced nucleotide dropping faster than Moore's law due to the development of next-generation sequencing methods, high-throughput DNA sequencing has become the workhorse of molecular biology in the 21st century. Since the completion of the first human genome in 2003 [61] (duration 13 years, estimated cost $3 billion), sequencing entire eukaryotic genomes has continuously become cheaper. Where sequencing of the first – relatively small – plant genome (*Arabidopsis thaliana*) took 10 years and cost $100 million [62, 63], the effort and money required for a similar effort today are much smaller. As a result, many question in molecular biology are currently being answered with DNA sequencing technologies: e.g. genome-wide gene expression can be measured by sequencing cDNA generated from mRNA, ribosome activity can be determined by ribosome profiling, and transcription factor binding sites can be identified with chromatin immunoprecipitation sequencing. This widespread adoption of high-throughput DNA sequencing has resulted in a steady demand for novel and/or efficient algorithmic approaches.

Although currently data generation is cheap, getting actionable knowledge from the ever increasing amount of sequence information is challenging. Originally defined as "the study of informatic processes in biotic systems" [64], the term bioinformatics quickly became synonymous with any computational approach to studying biology and working with biological data [65]. Undeniably, the increasing amount of sequence data has further cemented the link between molecular techniques and computational methods, up to a point where increasingly large portions of research budgets have to be allocated to bioinformatics [66]. Consequently, due to advances in computational tools for assembling, annotating, browsing and querying genomic information we are now better able to turn data into novel hypotheses than ever before [67].

In recent years, the technological advances in high-throughput DNA sequencing have been adopted for molecular research on nodulation. As a result, genome sequences of several nodulating legumes are available (e.g. the model legumes *Medicago truncatula* [18] and *Lotus japonicus* [17], but also *Trifolium pratense* [68] and *Cicer arietinum* [69]). Additionally, genome sequence information for several non-nodulating plant species in the nitrogen-fixing clade is currently available (e.g. *Malus* × *domestica* [70], *Prunus persica* [71] and *Cucumis sativus* [72]). Crucially, at the start of this thesis, genome sequence information for *Parasponia*, the only non-leguminous plant species that can form a symbiosis with rhizobium bacteria is not available. With the current cost of sequencing it has become timely to sequence genomes and transcriptomes of multiple *Parasponia* species and their closely related non-symbiotic *Trema* species. Using phylogenomic profiling of protein-coding genes and their transcriptional regulation, this thesis explores the evolution of molecular mechanisms involved in rhizobium symbiosis. The resulting knowledge on the evolution of nodulation will provide novel leads for engineering nitrogen-fixing crops.

1.5 This thesis

This thesis studies why some plants engage in nitrogen-fixing symbiosis with rhizobium bacteria, and others do not. It does so by reconstructing how rhizobium symbiosis naturally evolved, using a range of bioinformatics techniques. As a result, this thesis offers several new insights into the molecular evolution of rhizobium symbiosis. Additionally, it presents a variety of new hypotheses on molecular mechanisms used in the mutualistic interaction between rhizobium bacteria and plants. Whereas this thesis mainly applies existing technology to answer a biological question, it also describes a new framework to study the evolution of transcriptional networks and a browser that can be used for comparative studies in any organism. As such, the interdisciplinary nature of this thesis is reflected in the nature of the four chapters.

Chapter 2 deals exclusively with biological aspects of symbiotic plant-microbe interactions. It serves as a more elaborate review of current knowledge on molecular mechanisms in both rhizobium symbiosis and other forms of symbiotic interactions. As such, it sets the stage for a targeted comparative approach by listing candidate genes from other species.

Chapter 3 provides a new perspective on the evolution of rhizobium symbiosis by analyzing the newly sequenced genomes of several nodulating *Parasponia* and non-nodulating *Trema* species. By phylogenomic footprinting of protein-coding genes using new and existing data we describe a correlation between absence of nodulation and the loss of several known symbiosis genes. This finding potentially places the origin of nodulation at the root of the nitrogen-fixing clade, much earlier than previously thought.

Chapter 4 extends the approach of chapter 3 from individual genes to transcriptional networks inferred from public RNA sequencing datasets. It describes a new comparative framework that uses phylogenomic profiling of transcriptional interactions. We find that the ability to identify conserved transcriptional networks depends mainly on the accuracy of the inferred transcriptional networks. With a false-positive rate of 90%-99% we find that current computational methods for predicting transcriptional networks are unsuitable for comparative analysis. If transcriptional interactions can be reliably identified, the comparative approach developed in this chapter is a powerful tool for studying evolution of transcriptional networks involved in nodulation.

Chapter 5 describes a deployable web based system for browsing and querying the variety of data and analysis results that are used in modern molecular biological research. Whereas genome browsers have previously been limited to individual species, GeneNoteBook generalizes this concept to any comparative genomics project. A GeneNoteBook with protein-coding gene sequences, expression levels, sequence domains and phylogenetic trees of all species used in this thesis is available at `www.parasponia.org`. Making the newly generated genomic and transcriptomic data of *Parasponia* and *Trema* accessible on the web has established a powerful bioinformatics platform for the comparative analysis of rhizobium symbiosis.

Chapter 2

Commonalities in symbiotic plant-microbe signalling

Rens Holmer*, Luuk Rutten*, Wouter Kohlen, Robin van Velzen, and René Geurts

* authors contributed equally

This chapter was published as

Holmer R et al. “Commonalities in Symbiotic Plant-Microbe Signalling”. *Advances in Botanical Research*. Vol. 82. Elsevier, 2017, pp. 187–221

Abstract

Plants face the problem that they have to discriminate symbionts from a diverse pool of soil microbes, including pathogens. Studies on different symbiotic systems revealed commonalities in plant-microbe signalling. In this chapter we focus on four intimate symbiotic interactions: two mycorrhizal ones, with arbuscular- and ectomycorrhizal fungi, and two nitrogen-fixing ones, with rhizobium and *Frankia* bacteria. Comparing these systems uncovered commonalties in the way plants attract their symbiotic partners. Especially flavonoids, and in a lesser extent strigolactones, are pivotal plant signals that are perceived by the microsymbiont. In response, signal molecules are exuded by the microbes to trigger symbiotic responses in their host plant. Strikingly, microbes that establish an endosymbiotic relation with their host plant, namely arbuscular mycorrhizal fungi, rhizobium and *Frankia* bacteria, make use of a symbiotic signalling network that is highly conserved in plants. The use of flavonoids as attractants for symbiotic microbes, in combination with the use of a common plant signalling network to establish endosymbioses, raises questions about how plants manage to discriminate their microbial partners.

2.1 Introduction

High throughput sequencing approaches have uncovered an overwhelming diversity of soil microbes. Plants affect this microbial community – directly or indirectly – with their root systems. For example, roots exude substantial amounts of organic and amino acids, polymerized sugars (e.g. mucilage) as well as release border cells and dead root cap cells, which all form a nutrient source for many microbes [74]. On top of that more specific secondary metabolites are exuded that manipulate the microbial community by acting as antimicrobial agent or as attractant. Conversely, soil microbes can affect plant growth. For example, microbes can promote plant growth by improving nutrient availability, or inducing resistance against biotic and abiotic stresses [75]. On the other hand, pathogenic microbes can induce resource loss and disease. In this complex plant root microbiome network the plant must therefore discriminate between bacteria and fungi that provide an advantage and those that act as commensals or even pathogens. In this chapter we will focus on the molecular communication in a symbiotic context which occurs in plant roots and the rhizosphere. Plants establish several intricate long-term mutualistic relationships with microbes that are hosted intercellularly (ecto) or intracellularly (endo). Here, we will discuss the commonalities of four intimate symbiotic interactions. Thereby we will focus on two key stages of the interaction: attraction of the microbial partner, and subsequent microbe-induced signalling to establish a symbiosis.

2.2 Intimate plant root-microbe symbioses

Plant root symbioses occur at different levels of engagement, ranging from loosely attached microbes that provide a certain advantage to the plant to bacteria that are intracellularly accommodated as organelle-like structures [76]. The best studied plant root symbioses are those with arbuscular mycorrhizal and ectomycorrhizal fungi, and those with rhizobium and *Frankia* nitrogen fixing bacteria, together encompassing a diverse range of plant and microbial species.

2.2.1 Mycorrhizal symbioses

Mycorrhizal symbioses – the symbiotic interactions between some soil fungi and plant roots – can occur in several forms. Of these the ancient arbuscular (endo-) mycorrhiza and the much younger forms of ectomycorrhiza are best studied.

Based on fossil records arbuscular **mycorrhizal symbiosis** is estimated to be at least ~400-460 million years old, and evolved in a period that coincides with colonization of terrestrial habitats by plants [77–79]. Still today the vast majority of land plant species establish an arbuscular mycorrhizal symbiosis, underlining the ecological importance of this interaction [80]. The fungi that establish an arbuscular mycorrhizal symbiosis belong to a distinct taxonomic phylum, the Glomeromycota. This phylum possibly represents more than 1,000 species, though only less than 300 have been characterized to a certain level of detail [77]. Arbuscular mycorrhizal fungi are obligate biotrophs. Their hyphae penetrate plant roots intercellularly and form intracellular feeding structures – called arbuscules – in root cortical cells (figure 2.1A). Arbuscules are surrounded by a plant-derived membrane, but are largely deprived of plant cell wall material [81]. At this symbiotic interface nutrients are exchanged. Minerals – especially phosphates and nitrates – taken up by the fungal extraradical mycelium are delivered to the plant in return for carbohydrates. Arbuscules remain functional for several days, after which they collapse and disappear, leading to a reversion of the plant cell to its asymbiotic cortical fate.

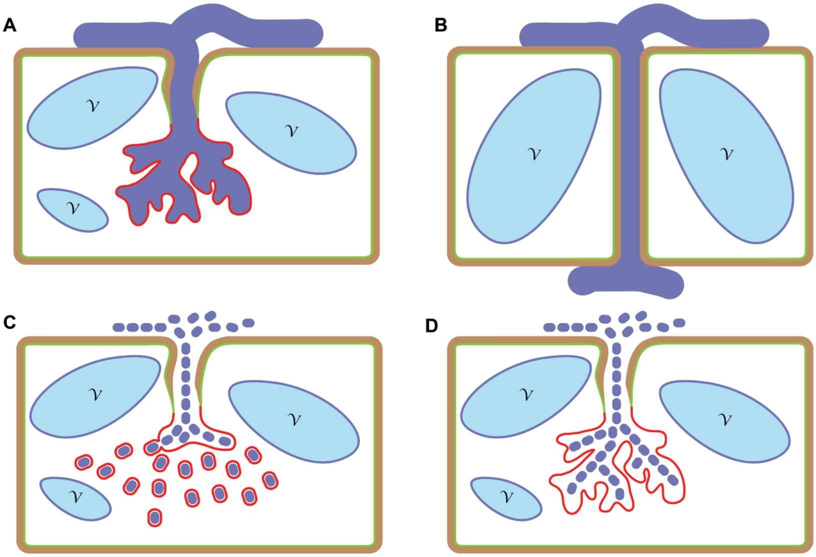


Figure 2.1: schematic representation of the cellular mode of infection of four symbioses discussed in this chapter. Green lines indicate plant cell membrane, red lines indicate the plant cell membrane derived symbiotic interface in the form of the periarbuscular membrane for AM, symbiosomes for rhizobium/legume and fixation threads for *Frankia* and rhizobium/*Parasponia*. **(A)** Hyphae of an endomycorrhizal fungus penetrate the cell and forms a feeding structure named arbuscule. Arbuscules are not surrounded by a plant derived cell wall. **(B)** Hyphae of an ectomycorrhizal fungus grow intracellularly. **(C)** Rhizobium bacteria are released as transient organelle-like structures – named symbiosomes – in nodule cells of most legumes. **(D)** *Frankia* and rhizobium infect cells of nodules of *Parasponia* and of some basal legumes through fixation threads. Fixation threads are largely deprived of plant cell wall. The bacteria in fixation threads remain in contact with the apoplast. Blue: vacuole (*V*), purple: microbe, brown: cell wall, green: plasmamembrane, red: plant-derived endosymbiotic membrane.

Ectomycorrhizal symbiosis can occur between diverse groups of plant and fungal species, as a result of several independent evolutionary events [82]. Overall this type of symbiosis can occur in about 2% of all land plants, including all dominant tree species in temperate forests, such as pines (*Pinus*), Douglas firs (*Pseudotsuga*), oaks (*Quercus*), willows (*Salix*), beeches (*Fagus*) and birches (*Betula*) [83, 84]. Ectomycorrhizal fungi belong to several taxonomic phyla including Basidiomycota, Ascomycota, and Zygomycota [84] which are all closely related to species with a saprotrophic lifestyle. Saprophytic fungi have an extensive repertoire of genes encoding degrading enzymes that can effectively mobilize resources, in particular nitrogen and phosphorus, from a variety of organic substrates [85]. However, compared to their saprophytic sister clades, ectomycorrhizal fungi only have a limited set of genes encoding plant cell walldegrading enzymes (e.g., pectin lyases and pectinases) [86]. Nevertheless, expression studies indicate that some of the plant cell wall-degrading enzymes that have been maintained may function during symbiosis [81].

Despite the diverse and paraphyletic groups of plant and fungal species that establish ectomycorrhizal symbioses, there is a remarkable resemblance in symbiotic phenotypes. The fungi preferentially colonise newly formed lateral roots. Upon hyphal attachment, they modulate root growth which allows them to colonise the root apoplast, forming a so-called Hartig net. The penetration depth of the hyphae is variable, but it typically comprises several layers of cortical cells, excluding the endodermis. In contrast to the endomycorrhizal symbiosis root cells are not invaded intracellularly (figure 2.1B). Ultimately, many fungal hyphae cover the root surface forming a thick, multi-layered mantle, insulating the infected lateral root.The molecular mechanism underlying Hartig net development has only recently been partially elucidated using genome sequencing data in combination with reverse genetic studies. For example, using reverse genetics in the fungus *Laccaria bicolor* an aquaporin (LbAQP1) was shown to be essential for Hartig net development [87]. Additionally, it was found that *L. bicolor* produces auxin (IAA) in its mycelium that triggers auxin-related responses in the plant root [88]. This finding is in line with pioneer work that showed that increased mycorrhizal activity is associated with increased auxin biosynthesis in by the fungus [89]. Together, these studies make clear that fungal signaling intertwines with plant auxin homeostasis to establish a symbiotic interaction.

2.2.2 Nitrogen fixing endosymbioses

A selective, though diverse, group of plant species is able to establish an endosymbiosis with nitrogen-fixing (diazotrophic) bacteria. These bacteria belong either to the genus *Frankia* or to the paraphyletic group of bacteria known as rhizobia. *Frankia* and rhizobia strains gained the symbiotic trait by horizontal gene transfer.

The genus *Frankia* is a diverse assemblage of filamentous sporangia-forming actinobacterial species that can be saprophytic, facultative symbiotic, or obligatory symbiotic. The *Frankia* genus can be separated in four separate bacterial clusters based on phylogenetic analysis, with only three of them that can establish symbiosis [90]. Plant species that can form a nitrogen-fixing endosymbiosis with *Frankia* bacteria ($\sim$230 species known as actinorhizal plants) are dispersed over 25 genera and 8 taxonomic families, suggesting multiple evolutionary origins of this symbiosis [91].

Relative to *Frankia*, rhizobia are even more diverse, representing 15 genera in eight families of α-, β- and γ-Proteobacteria [92]. Nitrogen-fixing symbiosis with rhizobia is prominent in the legume family (Fabaceae), but can also occur in *Parasponia*, a genus in the Cannabis family (Cannabaceae). Based on the phylogenetic distance between Fabaceae and Cannabaceae it is most probable that - similarly to the actinorhizal symbiosis - there are multiple origins for rhizobial symbiosis [20, 21]. Both types of endosymbioses with diazotrophic bacteria have in common the formation by the host of specific nodule-like structures (root nodules) in which the bacteria proliferate and fix nitrogen.

The reason for this may be that rhizobia and *Frankia* bacteria are generally not able to infect differentiated cells of the plant root. Only cells of the future nodule that are mitotically activated by the microsymbiont can be infected, suggesting that these cells are developmentally reprogrammed [93]. The nodules are optimized to facilitate growth of the microbial partner, which, once inside nodule cells, differentiates in its symbiotic form and fixes atmospheric nitrogen into ammonia in exchange for carbohydrates.

Variation exists in the way the nitrogen fixing bacterial partner is hosted. In most legume nodules, rhizobia are hosted in transient organelle-like structures, called symbiosomes. Symbiosomes are released from intracellular infection threads that have guided the rhizobium bacteria from the epidermis towards the newly formed nodule. Hundreds of symbiosomes surrounded by a plant-derived membrane, often containing only one bacterium, can be present in a single nodule cell. This membrane forms a symbiotic interface where nutrient exchanges take place between the bacteria and the cytoplasm of the host cell (figure 2.1D). In *Parasponia* and actinorhizal plants symbiosomes are not formed. Instead, the bacteria remain in thread-like structures, known as fixation threads (figure 2.1C). Fixation threads differ from the penetrating infection thread by a reduction of plant cell wall material. Fixation threads occur also in a few legume species, and may represent a more ancestral form of bacterial endosymbiosis than symbiosomes [20].

Of the four intimate symbiotic interactions that are central here, three have evolved more than once: the symbioses with rhizobia, *Frankia* and ectomycorrhizal fungi. This suggests an evolutionary advantage of root symbiosis for both partners. Interestingly, several studies indicate that similar mechanisms have been coopted in all four symbiotic interactions. Below we will discuss the commonalities in signalling mechanisms between the four symbioses central in this chapter.

2.3 Recognition and attraction of symbiotic partners

As outlined earlier, not all plants are able to form an intricate root microbe symbiosis; nor are all soil microbes symbiotic. Consequently, symbiotic partners need to recognize each other. Microbes recognize potential host plants by root exudates. Indeed, plants can exude signalling molecules to attract their symbiotic partner. Common signals in symbiotic partner recognition are exuded flavonoids, which play a role in all four symbioses. In addition, it was noted that exuded strigolactones can act as signal molecules, especially in arbuscular mycorrhizal symbiosis. Strikingly, both types of molecules function also as endogenous plant signals.

2.3.1 Flavonoids induce microbial responses

Flavonoids are a subclass of plant polyphenolic compounds and are a major class of secondary metabolites. As is typical for plant secondary metabolites, flavonoids are diverse: $\sim$9000 chemical structures have so far been reported [95]. Flavonoids are synthesized through the phenylpropanoid pathway. A chalcone synthase produces the chalcone scaffolds from which all other flavonoids are derived and is the first enzyme specific for flavonoid production [96]. A series of enzymatic reactions can alter the chalcone scaffold into a huge diversity of compounds, and flavonoids are typically categorized in subclasses based on these enzymatic reactions. The major subclasses of flavonoids include phlobaphenes, flavones, flavanones, flavonols, aurones, isoflavonoids, anthocyanins and condensed

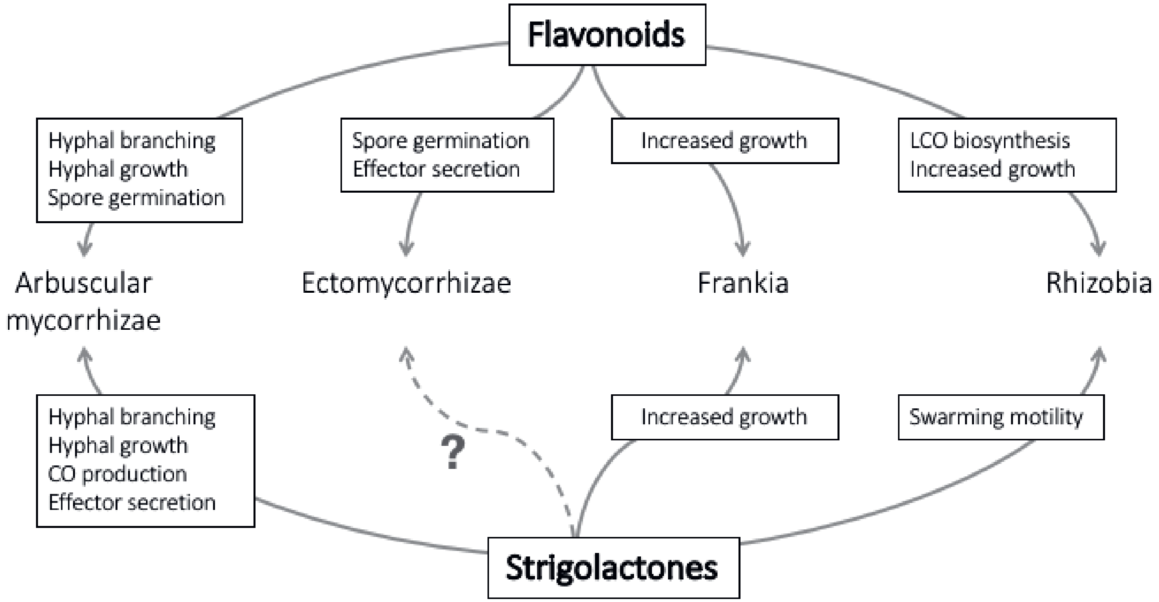


Figure 2.2: Flavonoids and Strigolactones are generic attractants for microsymbionts. COs = Chitooligosaccharides, LCOs = Lipochitooligosaccharides. Increased growth in *Frankia* can likely be attributed to either flavonoids, strigolactones, or both, as total root exudates were used to demonstrate this [94].

tannins. Many flavonoids are known to have glycosidated forms, i.e. quercitrin is formed by the addition of the deoxy sugar rhamnose to the flavonol quercetin, whereas rutin is formed by the addition of the disaccharide rutinose. Such small modifications can have drastic consequences for the observed effects in symbioses.

The involvement of flavonoids in symbioses has been described for all four types of symbiosis discussed in this review (arbuscular mycorrhizal, ectomycorrhizal, rhizobial and actinorhizal symbiosis). For two compounds positive effects in all four symbioses have been described (figure 2.2). Naringenin positively influences arbuscular mycorrhizal colonization [97] and rhizobium symbiosis [98], enhances spore germination of the ectomycorrhizal fungus *Suillis bovinus* [99] and restores *Frankia* nodulation in a chalcone synthase mutant of the actinorhizal plant *Casuarina glauca* [100]. Quercetin has been reported to stimulate spore germination, hyphal branching and growth of arbuscular mycorrhizal fungi [101], the growth rate of rhizobium bacteria [102], the actinorhizal nodulation [103] and it also stimulates the production of the symbiotic effector protein MiSSP7 in the ectomycorrhizal fungus *Laccaria bicolor* [85].

The molecular mode of action of naringenin and quercetin is not always known. Best studied is the effect of naringenin – and other flavonoids – in rhizobia, where flavonoids target NodD proteins. Rhizobial NodD proteins belong to the class of LysR-type transcriptional regulators that are activated upon the binding of external signals [22, 104]. Binding of a flavonoid molecule causes a conformational change which results in an increased binding affinity for specific cis regulatory elements. In case of NodD this element is known as the nod box [105]. Rhizobia generally have several operons that contain a nod box in their promoter region. Most prominent are the genes encoding an ABC transporter (NodI and NodJ) and 3 genes encoding the enzymes N-acetylglucosaminyltransferase (NodC), a chitooligosaccharide deacetylase (NodB) and a N-acyltransferase (NodA). These proteins are essential for biosynthesis and secretion of lipochitooligosaccharide molecules (LCOs),which act as potent symbiotic signal molecules(see section 2.4) [44, 106].

In case of arbuscular mycorrhizae and *Frankia* it remains elusive whether flavonoids trigger biosynthesis of similar symbiotic signalling molecules, despite the fact that flavonoids have a positive effect on both symbioses [97, 107]. LCOs and short chain chitin oligomers (tetra and pentameric COs) have been shown to be produced by the mycorrhizal fungus *Rhizophagus irregularis*, but their biosynthetic pathways have not yet been uncovered [108–111]. In case of symbiotic *Frankia* species, LCO biosynthesis genes are not common, and only found in a representative of a relatively isolated taxonomic lineage (cluster 2): namely *(candidatus) Frankia datiscae strain DG1* [112]. For this strain it was found that nodA, nodB, nodC, nodI and nodJ are expressed when the bacteria occupy Datisca glomerata root nodules [94]. Therefore, it is most probable that *F. datiscae* LCO signals play a symbiotic role.

Other flavonoids have been described to be involved in one or a few of the discussed symbioses, but were never tested in the other types of symbioses. Nevertheless, these observations can shed an interesting light on the symbiotic role of flavonoids. Especially interesting is the described host specificity in the legume rhizobia interaction [113], which in part is determined by recognition of specific flavonoids. Whereas a specific flavonoid can induce expression of the LCO biosynthetic nodABC operon in one bacterium, the same compound can have a negative effect in another bacterium. For example, the flavonoid coumestrol positively influences the symbiosis between Glycine max and Sinorhizobium fredii USDA191 [114] but negatively influences the symbiosis between Medicago sativa and Sinorhizobium meliloti 1021 [115]. In this context, it is also relevant to note that the composition of root exudates may vary depending on the developmental stage of the root. For example, studies in Medicago sativa indicate that flavonoids with a positive effect on the symbiosis are exuded in the elongation and differentiation zone that is susceptible to rhizobium infection, whereasrepelling flavonoids are exudedin the adjacent regions of the root (i.e. the root tip and the more mature part of the root) [115, 116]. However, such studies have not been further extended.

In addition to direct application of flavonoids to microbial cultures, reverse genetic studies in plants have been conducted. In most studies chalcone synthase genes were targeted. Chalcone synthase knockdown in actinorhizal *Casuarina glauca*, or in the legume model *Medicago truncatula* results in impaired nodulation [100, 117]. In both plant systems this phenotype can be restored by the application of naringenin. In contrast, no effect was reported on the arbuscular mycorrhizal symbiosis when using a chalcone-synthase double-mutant in maize [118]. The fact that a chalcone synthase maize mutant can be normally mycorrhizal with different fungal species, demonstrate that flavonoids are not an essential signal for this symbiosis. Nevertheless, flavonoids may act as facultative signals, and may play a role in host selection, by activating certain fungus over others [119].

The importance of flavonoids in root nodule symbiosis may be the result of the fact that flavonoids also act as an endogenous plant signal that controls auxin transport [117, 120, 121]. Based on quantitative modeling and experimental studies it is hypothesized that a transient decrease in auxin efflux can lead to formation of a local auxin maximum, which is the onset of nodule development [122, 123]. Such a function of flavonoids in nodulation is supported by the finding that naringenin can restore nodulation in the *Medicago truncatula* cytokinin signalling mutant Mtcre1 [124]. This study demonstrates that naringenin not only acts as an attractant of symbiotic microbes, but also functions as an endogenous plant signal, which - in a symbiotic context - acts downstream, or in parallel, to rhizobium-induced cytokinin signalling.

2.3.2 Dual role of strigolactones

Strigolactones are carotenoid-derived terpenoid lactones, often composed of four rings. Three rings form a tricyclic lactone, which is connected to the fourth butenolide ring via an enol ether bridge [125]. Strigolactones are known as endogenous plant hormones that control several steps in plant development [126]. Over the last decade major advances have been made on the identification of the strigolactone biosynthesis and perception pathway. A carotenoid isomerase (named DWARF27 or D27 in most species), two carotenoid cleaving dioxygenases (named CCD7 and CCD8), and a cytochrome P450 (possibly MAX1 in *Arabidopsis thaliana*) are sequentially required to produce the strigolactone backbones: either 4-deoxyorobanchol or 5-deoxystrigol [127, 128]. It is postulated that this backbone can be further decorated to produce the wealth of different strigolactone metabolites [129]. In plants the strigolactone receptor was identified as a / hydrolase (named OsDWARF14 or D14 in rice (*Oryza sativa*) and AtDAD2 in *Arabidopsis thaliana)*. Together with a specific F-box protein named OsDWARF3/AtMAX2 it forms the SCF E3 ubiquitin ligase complex required for strigolactone signalling [130, 131].

The discovery that strigolactones stimulate hyphal branching in the arbuscular mycorrhizal fungus *Gigaspora margarita* [132] has launched an interest in the involvement of these compounds in symbiotic signalling. The observation that strigolactones induce hyphal branching at very low concentrations in *Gigaspora margarita* has led to the hypothesis that the induction of hyphal branching must be receptor-mediated [132]. Furthermore, it was found that the synthetic strigolactone analog GR24 triggers mitochondrial activity in the arbuscular mycorrhizal fungi *Gigaspora intraradices* and *Gigaspora rosea* [133]. However, it should be noted that in order to induce hyphal branching in *G. rosea* besides GR24, also the flavonoid quercetin is needed in the fungal growth medium [134]. As quercetin is known to stimulate arbuscular mycorrhizal growth, hyphal branching and spore germination [135], this suggests that with this specific fungus strigolactones alone might not be sufficient to induce hyphal branching. In an independent experiment increased production of short chain COs upon application of GR24 was reported for *Rhizophagus irregularis* [110] (figure 2.2). In addition, a putative effector protein (RiSIS1) was identified in a screening of upregulated genes in GR24-treated *Rhizophagus irregularis* [136]. Using host-induced gene silencing the RiSIS1 gene was knocked down during infection, which resulted in significant suppression of colonization and in stunted arbuscules. This suggests that RiSIS1 is a strigolactone induced effector protein.

Application of GR24 to four ectomycorrhizal species revealed no effect on hyphal branching [137] (figure 2.2). This suggests strigolactones play a less important or different role in this type of symbiosis. In contrast, a negative effect of GR24 was observed on growth and branching of a range of phytopathogenic fungi [138], including species previously found not to be affected by GR24 [137]. It should be noted that these effects were only observed when relatively high concentrations of GR24 were used [138] and as such it remains unclear whether these concentrations were biologically relevant.

Apart from the beneficial effects in arbuscular mycorrhizal symbiosis, several studies revealed effects of strigolactones in the rhizobium/legume symbiosis (figure 2.2). Exogenous application of GR24 increases Medicago sativa nodule number when inoculated with Sinorhizobium meliloti (Soto et al., 2010). Interestingly, the same study reports that the bacterial growth and nodC expression are not affected by GR24, leading the authors to hypothesize that the effect of GR24 is on the plant. However, more recently it was suggested that GR24 might affect Sinorhizobium meliloti by promoting bacterial swarming motility [139]. In an independent experiment in *Medicago truncatula* low concentrations (0.1 M) of GR24 also resulted in increased nodule numbers, but higher concentrations (2-5 M) resulted in reduced nodule numbers and lateral root density [140]. Taken together this suggests that strigolactones act mainly as plant hormones involved in developmental programs during rhizobial symbiosis. In line with this, the strigolactone biosynthesis gene MtD27 was shown to be inducible by rhizobium LCOs three hours after application and that this induction is regulated by the common symbiotic signalling pathway [141].

Mutants and knockdown experiments of strigolactone biosynthesis genes in several species shed light on the dual role of strigolactones in symbioses. Whereas often symbiotic phenotypes are observed, it is not trivial to decide whether these phenotypes are an effect of a change in direct signalling between host and symbiont, or whether a change in hormonal balance causes a difference in plant developmental program. Carotenoid cleavage dioxygenases (CCD) are among the most studied strigolactone biosynthetic enzymes in a symbiotic context. For both ccd7 and ccd8 knockout mutant phenotypes in arbuscular mycorrhizal symbiosis and rhizobia symbiosis have been reported. Mutants and/or knockdown of ccd7/8 in several plant species display reduced mycorrhizal colonization [142–146]. The importance of strigolactones in mycorrhizal colonization is further supported by the identification and knockout of a strigolactone transporter in *Petunia hybrida* (PhPDR1) clearly demonstrates the effect of strigolactones on *Gigaspora margarita* and *Rhizophagus irregularis* mycorrhization efficiency: soil levels of strigolactones are lower, plant show reduced colonization and fungi display reduced growth, branching and spore germination [144]. In addition, nodulation was reported to be impaired in the *Lotus japonicus* CCD7 knockdown and both ccd7 and ccd8 mutants of pea [145, 147]. The GRAS transcriptional regulators NSP1 and NSP2 were identified as regulators of strigolactone biosynthesis in rice and *Medicago truncatula* by regulating D27 expression [148]. Medicago truncatula nsp1 and nsp2 mutants are not capable of forming nodules [149, 150]. The nsp1 mutant and the nsp1/nsp2 double mutant produce no detectable amounts of strigolactones, whereas the nsp2 mutant has a reduced and different strigolactone composition. Interestingly, mycorrhizal colonization of the nsp1/nsp2 double mutant by *Rhizophagus irregularis* was only mildly reduced [148].In addition, the *Lotus japonicus* nsp1 mutant is unable to form nodules, however infection by the arbuscular mycorrhizal fungus *Rhizophagus irregularis* was unaffected [151].

The rice and pea Osd3/Psrms4 are markedly reduced in mycorrhizal colonization [147, 152]. This suggests that strigolactone perception in planta plays a role in AM colonization. Strikingly, in the pea Psrms4 mutant nodule numbers are increased [147], indicating that the effect of strigolactones in nodulation is regulated differently compared to mycorrhization.

Interestingly, a severe mycorrhization phenotype in rice could be complemented by introducing a copy of OsD14-LIKE gene [153]. OsD14-Like is paralogous to OsD14 and has strong similarities with the *Arabidopsis thaliana* karrikin receptor AtKAI2. OsD14 and OsD14-LIKE have been reported to have partially overlapping, but also distinct, functions for strigolactone and karrikin responses, as the Atkai2 mutants are insensitive to karrikins but weakly responsive to strigolactones [154]. In addition, it was recently demonstrated that in *Arabidopsis thaliana* AtD14 and AtD14-

like have different affinities for specific strigolactone stereoisomers [154]. This could indicate that the perception of specific strigolactones is regulated by multiple receptor complexes.

Taken together, the involvement of strigolactones in arbuscular mycorrhiza symbiosis is relatively well described, although several details remain unclear. A possible involvement in rhizobial symbiosis is just starting to be discovered. However, given the distinct nature of both symbioses, the mechanisms by which strigolactones function are likely different between the two. For ectomycorrhizal and actinorhizal symbioses no clear data on the involvement of strigolactones is available yet. As strigolactones are plant hormones involved in key developmental processes it is not surprisingly that ectomycorrhizal hosts were found to possess the strigolactone biosynthetic genes [155].

2.4 A Conserved signalling pathway for endosymbioses

As mentioned above, arbuscular mycorrhizal fungi, rhizobia and some basal *Frankia* species produce LCO signals in a symbiotic context, whereas no evidence has been found that LCO signals are playing a role in ectomycorrhizal symbiosis. This suggests that LCO signalling is a feature of microbes that establish an endo- rather than an ectosymbiosis.

LCOs are prominent signal molecules that are perceived by the host plant and set in motion symbiotic responses. Genetic studies in legumes, rice, *Parasponia* and the actinorhizal plant species *Datisca glutinosa* (nodulated by *Frankia* sp. harboring LCO biosynthesis genes), but also in *Casuarina glauca*, a species that is nodulated by *Frankia* sp. that lack LCO biosynthesis genes uncovered a common symbiotic signalling network. This conserved symbiotic network stretches from transmembrane receptor kinases to a network of transcription factors that control the readout of symbiotic signalling [44]. A hallmark of endosymbiotic signalling is the induction of regular oscillations of the nuclear calcium concentration. To achieve this a complex of nuclear envelope-localized proteins are essential, including a potassium-permeable channel (encoded by MtDMI1, LjCASTOR, LjPOLLUX), a cyclic nucleotidegated calcium channel, and a calcium ATPase [30, 156–158]. The induced calcium oscillations are decoded by a calcium-/calmodulin-dependent kinase (CCaMK), which is the onset of a transcriptional network [40]. Besides some common elements, like the CCaMK interacting transcription factor LjCYCLOPS, the activated network varies between arbuscular mycorrhizal and root nodule symbioses. For example, activation of the NIN transcription factor is essential for root

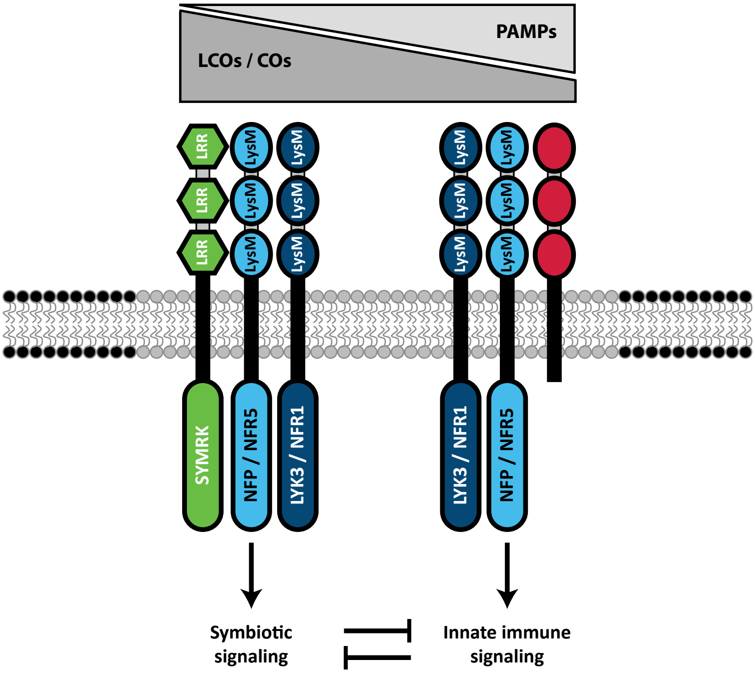


Figure 2.3: Hypothetical model explaining the dual functioning of LCO receptors in symbiotic and innate immune signalling. Symbiotic signals and pathogen associated molecular patterns (PAMPs) are perceived by NFP/NFR5-NFR1/LYK3 receptor complexes. To commit either symbiotic or innate immune signalling a third receptor is essential. For symbiotic signalling this receptor may be SYMRK, as it interacts with NFP/NFR5. To induce innate immune responses this receptor has not been identified yet, but may have similarities to CeBIP in rice. OsCEBiP binds chitin oligomers and forms a heteromeric complex with the rice homolog of NFR1/LYK3 (OsCERK1) to activate innate immune signalling.

nodule formation in legumes and *Casuarina glauca*, whereas it is not for arbuscular mycorrhizal symbiosis [159–161]. Conversely, arbuscular mycorrhizal symbiosis requires activation of GRAS transcription regulators such as MtRAM1 in medicago, which is not essential for root nodule formation [162]. Despite this divergence in transcriptional responses, the common symbiotic signalling genes are conserved in angiosperm and gymnosperm species that form an arbuscular mycorrhizal symbiosis. By contrast, plants that exclusively establish an ectomycorrhizal symbiosis - e.g. Pinaceae species - have lost several of these genes [155]. This supports the idea that ectomycorrhizal symbioses are founded on different signalling cues than arbuscular mycorrhizal and root nodule endosymbioses.

2.4.1 LCO signalling

Most comprehensive studies on symbiotic signalling have been done in the legume model systems *Lotus japonicus* and *Medicago truncatula*. Both species have evolved a strain specificity (*Mesorhizobium loti* for *Lotus japonicus* and *Sinorhizobium meliloti* for *Medicago truncatula*). In these systems it was revealed that rhizobium LCOs are specifically recognized by a heteromeric complex of receptor-like kinases (LysM-RLKs) containing Lysin motif (LysM) domains: named LjNFR1 and LjNFR5 in *Lotus japonicus*, and MtLYK3 and MtNFP in *Medicago truncatula* [23–26]. The LysM domain is a ubiquitous molecular structure of 42-48 amino acids with a symmetrical folding. LysM domain-containing proteins were first described in bacteria to bind peptidoglycan [163]. In legumes LjNFR1/MtLYK3 and LjNFR5/MtNFP harbor 3 LysM domains in the receptor region which are essential to recognize specific rhizobium LCOs [27]. In addition, it was found in *Lotus japonicus* that LjNFR5 interacts also withLjSYMRK, a LRR-type receptor that commits an essential function in symbiotic signalling [164]. Interestingly, SYMRK is also essential for arbuscular mycorrhizal symbiosis, whereasboth LjNFR1/MtLYK3 and LjNFR5/MtNFP only play an additive role in arbuscular mycorrhizal symbiosis [44]. Arbuscular mycorrhizal LCOs are known to trigger lateral root formation in *Medicago truncatula*, a response that is abolished in the Mtnfp knockout mutant [109]. Mtlyk3 and Ljnfr1 mutants display only a reduced level of infection when inoculated with a low dose of arbuscular mycorrhizal spores [165]. Nevertheless, transcriptome studies in *Medicago truncatula* revealed that MtNFP is playing a prominent role in arbuscular mycorrhizal LCO-induced transcriptional changes [166]. Two reasons may explain this discrepancy between knockout phenotype and function. Firstly, the weak arbuscular mycorrhizal symbiosis phenotype of the Ljnr1/Mtlyk3 and Ljnfr5/Mtnfp knockout mutants may be the result of gene redundancy in *Lotus japonicus* and *Medicago truncatula*. Both rhizobium LCO receptors evolved upon gene duplication events, giving rise to closely related homologs [18, 47, 167]. Expression studies of these homologous genes show that they may also function in rhizobium and/or arbuscular mycorrhizal symbiosis [18, 168]. Secondly, it was found that arbuscular mycorrhizal fungi not only produce LCOs, but also short chain chitooligosaccharides (tetra and pentameric COs) as symbiotic signals [110]. Such COs trigger in part similar symbiotic responses as reported for arbuscular mycorrhizal LCOs, though lack the capacity to promote lateral root formation [109]. LCO and CO signals may be perceived by different (symbiotic) receptor complexes.

Non legume systems provided additional support for a function of

NFR1/LYK3 and NFR5/NFP homologous genes in arbuscular mycorrhizal symbiosis. Reverse genetic studies in *Parasponia andersonii* and tomato (Solanum lycopersicum) revealed an essential role for putative NFR5/NFP orthologs in arbuscular mycorrhizal symbiosis [47, 169]. In rice (Oryza sativa) it was demonstrated that the putative ortholog of NFR1/LYK3 - CHITIN-ELICITOR RECEPTOR KINASE 1 (OsCERK1)- plays such role [165, 170]. In *Frankia* no reverse genetic studies on LysM-RKs have been published yet. However, it is tempting to speculate that in actinorhizal plant species that can be nodulated by cluster 2 *Frankia* species, close homologs of NFR1/LYK3 and/or NFR5/NFP play a symbiotic role in LCO perception.

The finding that COs and the chitin innate immune receptor OsCERK1 commit symbiotic functions uncovered a functional overlap between pathogenicity and symbiosis. Subsequent studies in *Lotus japonicus* and *Medicago truncatula* revealed four lines of supportive evidence for such dual function of LCO receptors. (I.) Rhizobium LCOs transiently trigger defense-related gene expression in an LjNFR1-dependent manner [171]. (II.) Mt-NFP has a function in defense against fungal and oomycete pathogens [172–174]. (III.) Ectopic expression of both receptors - LjNFR5-LjNFR1 or MtNFP-MtLYK3 - in Nicotiana benthamiana leaves triggers a hypersensitive response (HR) [27, 175]. (IV.) Ectopic expression of MtNFP in *Medicago truncatula* triggers a premature cell death in nodules [176]. This, and other studies, made also clear that LCO receptors are under tight post-translational control in legumes, probably to prevent pathogenic responses. For example, in *Medicago truncatula* nodules MtNFP and MtLYK3 accumulate only in nodule cells where infection takes place, but both receptors are rapidly removed from the membrane surrounding the rhizobium infection thread [176]. Furthermore, it was found that LCO receptors are located in lipid-raft-like micro-domains in the plasma membrane, which play an important role in complex formation and receptor turnover [177, 178]. Taken together, these data suggest that dual functioning of LCO receptors in defense and symbiosis is a conserved feature in legumes and non-legume species.

The biological function of the overlap of LCO receptors in symbiotic and innate immune signalling remains unclear. However, a challenging model can be postulated [106]. In this model competition between receptors occur to form multimeric complexes that differ in their functioning. Presence of LCOs (and/or short-chain COs) results in preferential formation of symbiotic receptor complexes at the expense of the formation of complexes that act in innate immunity (figure 2.3). In legumes, such innate immune receptor complex has not yet been characterized. However, studies in rice revealed that perception of chitin oligomers requires an additional LysM-domain-containing receptor, which lacks an intracellular kinase domain [179]. This CHITIN ELICITOR BINDING PROTEIN (OsCEBiP) binds chitin oligomers and forms a heteromeric complex with OsCERK1 to activate chitin-triggered defense responses [180, 181]. Such innate immune receptor complex may also have a function in symbiosis. It is known that several typical innate immune responses, such as calcium influx, production of reactive oxygen (ROS) species, and focal exocytosis are associated with rhizobial and arbuscular mycorrhizalinfection [182]. Rhizobium triggers formation of infection threads, which are tip-growing structures. ROS production is thought to facilitate the oxidative crosslinking of the infection thread matrix to allow the formation of a tube-like infection thread [182]. In a scenario that innate immune responses play a symbiotic role, the spatiotemporal regulation of receptor complexes becomes crucial to prevent HR.

2.4.2 Bypassing LCO signaling

Besides LCO-mediated signalling, alternative routes occur to mediate symbiotic responses. For example, many *Frankia* species (clusters 1 and 3) do not possess the machinery to produce LCOs [183]. Furthermore, there are some legume lineages - e.g. several *Aeschynomene* species - that are nodulated by *Bradyrhizobium* strains that lack the highly conserved nodABC operon [184, 185] necessary for LCO synthesis. Nevertheless, studies in actinorhizal plant species *Casuarina glauca* and *Alnus glutinosa* using the non-LCO producing *Frankia* strain Cci3i, revealed that both SYMRK and CCaMK are essential to establish a symbiotic interaction, and activation of symbiotic signalling induces calcium oscillations [48, 186–188]. This strongly suggests that the underlying signalling pathway to establish an endosymbiosis is highly conserved, but can be activated by different signalling inputs.

The way non-LCO-producing rhizobia and *Frankia* achieve activation of the common symbiosis signalling pathway may vary. One way is by producing effector-like molecules that are secreted via the type III secretion system (T3SS). This mechanism is used by several rhizobium strains [189], and studies in soybean revealed that such effectors can bypass NFR1-NFR5 based signalling [190]. However, additional mechanisms may also occur. For example, in case of *Aeschynomene* legumes the common symbiosis signalling pathway can also be activated in a T3SS-independent way [184, 189].

The current hypothesis is that cluster 1 and 3 *Frankia* strains produce a signalling molecule upon host recognition, of which the chemical nature is still poorly understood, but most probable different from LCOs. A first characterization of such signals came from studies on *Frankia* sp. strain CcI3i that nodulates *Casuarina glauca*. The signalling molecules produced by this strain are of low molecular weight, lying within the range 5005,000 Dalton. Moreover, rhizobium and arbuscular mycorrhizal LCOs typically accumulate in the organic fraction upon 1-butanol extraction, whereas, in case of *Frankia* CcI3i exudates, only water fractions could induce symbiotic responses (i.e. calcium oscillation). Furthermore, a chitinase treatment on the active water fractions did not affect its symbiosis signalling capacity [186]. This makes it highly unlikely this strain produces LCO-type symbiotic signal molecules.

Studies with other *Frankia* strains revealed that at least within a taxonomic cluster the symbiotic signals are to a certain level conserved. For example, *Alnus glutinosa* and *Casuarina glauca* are nodulated by different *Frankia* strains, though both belonging to taxonomic cluster 1. Despite this strain specificity, *Frankia* sp. strain AC14a that nodulates *Alnus glutinosa*, induces also calcium oscillation responses in *Casuarina glauca*. However, the more distant cluster 3 strain BCU110501 was unable to induce such response [186]. This suggests that the symbiotic signals produced by *Frankia* species partially is conserved within a taxonomic cluster, but may differ in a broader phylogentic context.

2.5 Repressing immunity

Although innate immune responses may be an integral part of the symbiotic infection process, it is essential that severe immune responses are avoided. Immune responses are controlled by two antagonistic hormones jasmonic acid and salicylic acid. The latter hormone is a major signal in resistance to biotrophic pathogens, whereas defense against necrotrophic mainly relies on jasmonic acid [191]. Both hormones act antagonistically, such that activation of jasmonic acid signalling compromises salicylic acid-dependent innate immune responses, and vice versa.

Studies in legumes suggest that repression of innate immunity is in part controlled by LCO signalling. In alfalfa (*Medicago sativa*) evidence was found that LCO signalling suppresses salicylic acid-dependent responses. LCO-deficient or incompatible rhizobia induce accumulation of salicylic acid, whereas compatible strains trigger a decrease of this defence hormone [192]. Similarly, studies in pea (*Pisum sativum*) showed that endomycorrhizal fungi only trigger a transient increase in salicylic acid levels, which is repressed during prolonged colonization. In contrast, in a symbiosis deficient ccamk knockout mutant salicylic acid levels remain high upon inoculation with endomycorrhiza, suggesting that this suppression is based on activation of the symbiosis signalling network [193]. Interestingly, defence responses in non-legumes (*Zea mays*, *Setaria viridis*), and even non-AM plants (*Arabidopsis thaliana*) seem to be downregulated upon LCO perception, however it is currently unclear how this is linked to JA and SA signalling [194, 195].

The Jasmonic acid - salicylic acid balance is in part controlled by DELLA GRAS-type transcriptional regulators [196]. DELLAs promote jasmonic acid signalling by binding JAZ (JASMONATE ZIM-DOMAIN) repressor proteins [197]. JAZ proteins repress jasmonic acid signalling upon binding with the MYC2 transcriptional activator [197, 198]. As MYC2 activity promotes DELLA accumulation, this results in a feedforward loop in jasmonic acid signalling [199, 200]. Several experiments indicate that endomycorrhizal fungi and rhizobium exploit this pathway, thereby indirectly reducing salicylic acid responses. Della knockout mutants in *Medicago truncatula* and rice are impaired in nodulation and/or arbuscule formation [201–204]. These phenotypes can be mimicked by application of gibberellins, whereas ectopic expression of a dominant active DELLA allele (MtDELLA118) promotes symbtioc responses [201, 203, 205]. Interestingly, the dominant active allele can also complement the cyclops symbiotic signalling mutant [201]. This is likely due to the fact that in *M. truncatula* the DELLA1 protein was found to be able to form a complex with CYCLOPS and CCaMK, together activating the RAM1 GRAS-type transcriptional regulator [203]. Taken together this suggests that MtDELLA1 plays an important role in the LCO signalling network, amongst others by promoting endomycorrhizal symbiosis through the modulation of jasmonic acid- salicylic acid balance by interacting with JAZ proteins.

Besides LCO triggered repression of immunity, plant immunity can also be manipulated by microbe secreted effector proteins. Studies in arbuscular mycorrhiza and ectomycorrhiza uncovered several small secreted effector proteins that are produced by the arbuscular mycorrhizal fungus *R. irregularis* and the ectomycorrhizal fungus *L. bicolor* [108, 111, 206]. The mode of action of two such effector proteins has been characterized.

The *Rhizophagus irregularis* effector protein RiSP7 is secreted into *Medicago truncatula* root cells, where it localizes in the nucleus and interacts with a defense controlling ethylene-responsive transcription factor (MtERF19) [207]. In *Medicago truncatula* roots this gene is highly expressed upon pathogenic interaction, but only transiently during arbuscular mycorrhizal colonization. Ectopic expression of RiSP7 in *Medicago truncatula* roots positively affects mycorrhizal colonization, while reducing defense responses. Intriguingly, RiSP7 has some similarity to the secreted NodO protein of *Rhizobium leguminosarum*, which enhances LCO signalling in the host plant. However, localization studies suggest that NodO localizes in the plant membrane, rather than acting as a nuclear effector [208, 209].

The ectomycorrhizal fungus *Laccaria bicolor* expresses the LbMiSSP7 gene encoding a secreted effector protein in response to plant exuded flavonoids [85]. In black cottonwood poplar (*Populus trichocarpa*) it was shown that LbMiSSP7 is secreted in root cells where it localizes in the nucleus. There it stabilizes a JAZ protein (PtJAZ6) by direct interaction [210]. As outlined above, JAZ proteins are repressors of jasmonic acid triggered immunity. Generally, JAZ proteins are degraded upon interaction with the F-box protein COI1 (CORONATINE-INSENSITIVE 1). This degradation is triggered by jasmonic acid signalling. LbMiSSP7 interaction to PtJAZ6 affects formation of the JAZCOI1 complex. This prevents the jasmonic acid-dependent degradation of JAZ, resulting in reduced plant immune responses. Given that jasmonic acid is a negative regulator of ectomycorrhizal symbiosis, counteracting this plant innate immune response promotes the plant-fungus interaction.

2.6 Perspectives in symbiotic signalling

Central questions for future research will be on specificity of symbiotic signalling. How can a single symbiotic network that is conserved in most land plants trigger distinct root phenotypes? Since the symbiotic signalling network is basically conserved in most plant species the differences in the readout may be determined by yet unknown factors, such as the hormonal balance and/or the nutrient status of the root. For example, recently it was shown that *Medicago truncatula* lateral roots have an increased sensitivity to rhizobium LCOs compared with the main root. This indicates that susceptibility of a plant root varies, depending on the developmental and/or nutrient status [211].

Additional questions concerning specificity can also be addressed concerning the plant exuded flavonoids that act as attractants for symbiotic microbes. As shown for naringenin, these compounds are perceived by a diverging range of symbionts. Most probably this range extends to other soil borne microbes, most of which will not be symbiotic. Therefore, perhaps exuded flavonoids do not act as specific signals, but rather are more generic signals to which any root microbe can respond. For example, it was reported that exuded flavonoids may play a role also in phosphate and iron acquisition [212]. In addition, the finding that flavonoids -similar to strigolactones- have a dual function, not only act as an attractant, but also function as endogenous plant signal interfering with auxin homeostasis, provides novel leads in symbiosis research.

Extending the range of model systems that are amenable for molecular genetic studies provided novel insights in symbiotic signalling. Establishment of new protocols for culturing *Frankia* and arbuscular mycorrhizal fungi, transformation of the ectomycorrhizal fungus *Laccaria bicolor*, the actinorhizal plants *Datisca*, *Casuarina*, and the non-legume rhizobia host *Parasponia* in combination with microbial genome sequencing has opened new avenues. Although unraveling symbiotic signalling in these systems is still in its infancy, the recent findings that have been achieved are already groundbreaking. As mentioned above, it was demonstrated that especially in the endosymbioses (*Frankia*, rhizobium and arbuscular mycorrhiza) commonalities occur in symbiotic signalling [47, 48]. One such commonality is that symbionts recognize plant secreted flavonoids and strigolactones. Another common theme is the use of LCO or CO signals of microbial origin of which biosynthesis is activated upon recognition of plant exuded molecules like flavonoids and/or strigolactones. LCOs/COs activate a conserved symbiotic network in plants that controls the diverse signalling output of the different symbiotic interactions [45]. Furthermore, it became apparent that LCO induced signalling can be bypassed. Especially in *Frankia* this appears to be a common strategy. Nevertheless, first studies indicate that LCO-independent signalling relies on the same symbiotic signalling network as identified in LCO dependent systems. Uncovering the nature of the non-LCO signal molecules in *Frankia* and rhizobia will add a new building brick in the symbiotic signalling network.

In the last decade new insights in the molecular aspects of root symbiosis were mainly generated by studying legume models *Medicago truncatula* and *Lotus japonicus*. With new model species in place in combination with next generation sequence technologies, this field will be revolutionized in the years to come.

Acknowledgements

Rik Huijsman for providing assistance with figures

Funding

R.G. is supported by NWO-VICI 865.13.001 and W.K by NWO-VENI 863.15.010

Chapter 3

Comparative genomics of the nonlegume *Parasponia* reveals insights into evolution of nitrogen-fixing rhizobium symbioses

Robin van Velzen*, Rens Holmer*, Fengjiao Bu+, Luuk Rutten+, Arjan van Zeijl, Wei Liu, Luca Santuari, Qingqin Cao, Trupti Sharma, Defeng Shen, Yuda Roswanjaya, Titis A.K. Wardhani, Maryam Seifi Kalhor, Joelle Jansen, Johan van den Hoogen, Berivan Güngör, Marijke Hartog, Jan Hontelez, Jan Verver, Wei-Cai Yang, Elio Schijlen, Rimi Repin, Menno Schilthuizen, M. Eric Schranz, Renze Heidstra, Kana Miyata, Elena Fedorova, Wouter Kohlen, Ton Bisseling, Sandra Smit, and René Geurts

*,+ authors contributed equally

This chapter was published as

Velzen R van et al. “Comparative genomics of the nonlegume Parasponia reveals insights into evolution of nitrogen-fixing rhizobium symbioses”. *Proceedings of the National Academy of Sciences of the United States of America* 115.20 (2018), E4700–E4709

Supplementary information is available at PNAS online

Abstract

Nodules harboring nitrogen-fixing rhizobia are a well-known trait of legumes, but nodules also occur in other plant lineages, with rhizobia or the actinomycete *Frankia* as microsymbiont. It is generally assumed that nodulation evolved independently multiple times. However, molecular-genetic support for this hypothesis is lacking, as the genetic changes underlying nodule evolution remain elusive. We conducted genetic and comparative genomics studies by using *Parasponia* species (Cannabaceae), the only nonlegumes that can establish nitrogen-fixing nodules with rhizobium. Intergeneric crosses between *Parasponia andersonii* and its non nodulating relative *Trema tomentosa* demonstrated that nodule organogenesis, but not intracellular infection, is a dominant genetic trait. Comparative transcriptomics of *P. andersonii* and the legume Medicago truncatula revealed utilization of at least 290 orthologous symbiosis genes in nodules. Among these are key genes that, in legumes, are essential for nodulation, including *NODULE INCEPTION* (*NIN*) and *RHIZOBIUM-DIRECTED POLAR GROWTH* (*RPG*). Comparative analysis of genomes from three *Parasponia* species and related non nodulating plant species show evidence of parallel loss in non nodulating species of putative orthologs of *NIN*, *RPG*, and *NOD FACTOR PERCEPTION*. Parallel loss of these symbiosis genes indicates that these non nodulating lineages lost the potential to nodulate. Taken together, our results challenge the view that nodulation evolved in parallel and raises the possibility that nodulation originated 100 Mya in a common ancestor of all nodulating plant species, but was subsequently lost in many descendant lineages. This will have profound implications for translational approaches aimed at engineering nitrogen-fixing nodules in crop plants.

3.1 Introduction

Nitrogen sources such as nitrate or ammonia are key nutrients for plant growth, but their availability is frequently limited. Some plant species in the related orders Fabales, Fagales, Rosales, and Cucurbitales – collectively known as the nitrogen-fixing clade – can overcome this limitation by establishing a nitrogen-fixing endosymbiosis with *Frankia* or rhizobium bacteria [9]. These symbioses require specialized root organs, known as nodules, that provide optimal physiological conditions for nitrogen fixation [37]. For example, nodules of legumes (Fabaceae, order Fabales) contain a high concentration of hemoglobin that is essential to control oxygen homeostasis and protect the rhizobial nitrogenase enzyme complex from oxidation [37, 214]. Legumes, such as soybean (*Glycine max*), common bean (*Phaseolus vulgaris*), and peanut (*Arachis hypogaea*), represent the only crops that possess nitrogen-fixing nodules, and engineering this trait in other crop plants is a long-term vision in sustainable agriculture [6, 215].

Nodulating plants represent $\sim$10 related clades that diverged $>$100Mya, supporting a shared evolutionary origin of the underlying capacity for this trait [9]. Nevertheless, these nodulating clades are interspersed with many non nodulating lineages. This has led to two hypotheses explaining the evolution of nodulation [9]. The first is that nodulation has a single origin in the root of the nitrogen-fixation clade, followed by multiple independent losses. The second is that nodulation originated independently multiple times, preceded by a single hypothetical predisposition event in acommon ancestor of the nitrogen-fixing fixation clade. The latter of these hypotheses is more widely accepted [11–15, 216, 217].

Genetic dissection of rhizobium symbiosis in two legume models – *Medicago truncatula* (medicago) and *Lotus japonicus* (lotus) – has uncovered symbiosis genes that are essential for nodule organogenesis, bacterial infection, and nitrogen fixation (Dataset S1). These include genes encoding LysM-type receptors that perceive rhizobial lipochitooligosaccharides (LCOs; also known as Nod factors) and transcriptionally activate the *NODULE INCEPTION* (*NIN*) transcription factor [23–27, 160]. Expression of *NIN* is essential and sufficient to set in motion nodule organogenesis [33, 160, 161, 218]. Some symbiosis genes have been coopted from the more ancient and widespread arbuscular mycorrhizal symbiosis [44, 45]. However, causal genetic differences between nodulating and non nodulating species have not been identified [93].

To obtain insight into the molecular-genetic changes underlying evolution of nitrogen-fixing root nodules, we conducted comparative studies by using *Parasponia* (Cannabaceae, order Rosales). The genus *Parasponia* is the only lineage outside the legume family establishing a nodule symbiosis with rhizobium [219–222]. Similarly as shown for legumes, nodule formation in *Parasponia* is initiated by rhizobium-secreted LCOs [47, 188, 223]. This suggests that *Parasponia* and legumes use a similar set of genes to control nodulation, but the extent of common gene use between distantly related nodulating species remains unknown. The genus *Parasponia* represents a clade of five species that is phylogenetically embedded in the closely related *Trema* genus [224]. Like *Parasponia* and most other land plants, *Trema* species can establish an arbuscular mycorrhizal symbiosis (SI Appendix, Fig. S1). However, they are non responsive to rhizobium

LCOs and do not form nodules [188, 222]. Taken together, *Parasponia* is an excellent system for comparative studies with legumes and non nodulating *Trema* species to provide insights into the molecular-genetic changes underlying evolution of nitrogen-fixing root nodules.

3.2 Results

3.2.1 Nodule organogenesis is a genetically dominant trait

First, we took a genetics approach to understanding the rhizobium symbiosis trait of *Parasponia* by making intergeneric crosses (SI Appendix, Table S1). Viable F_1 hybrid plants were obtained only from the cross *Parasponia andersonii* ($2n = 20$) × *Trema tomentosa* ($2n = 4x = 40$; Figure 3.1A and SI Appendix, Fig. S2). These triploid hybrids ($2n = 3x = 30$) were infertile, but could be propagated clonally. We noted that F_1 hybrid plants formed root nodules when grown in potting soil, similar to earlier observations for *P. andersonii* [225]. To further investigate the nodulation phenotype of these hybrid plants, clonally propagated plants were inoculated with two different strains, *Bradyrhizobium elkanii* strain WUR3 [225] or *Mesorhizobium plurifarium* strain BOR2. The latter strain was isolated from the rhizosphere of *Trema orientalis* in Malaysian Borneo and showed to be an effective nodulator of *P. andersonii* (SI Appendix, Fig. S3). Both strains induced nodules on F_1 hybrid plants (Figure 3.1B,D, and E and SI Appendix, Fig. S4) but, as expected, not on *T. tomentosa*, nor on any other *Trema* species investigated. By using an acetylene reduction assay, we noted that, in contrast to *P. andersonii* nodules, in F_1 hybrid nodules of plant H9 infected with *M. plurifarium* BOR2 there is no nitrogenase activity (Figure 3.1C). To further examine this discrepancy, we studied the cytoarchitecture of these nodules. In *P. andersonii* nodules, apoplastic *M. plurifarium* BOR2 colonies infect cells to form so-called fixation threads (Figure 3.1F and H-J), whereas, in F_1 hybrid nodules, these colonies remain apoplastic and fail to establish intracellular infections (Figure 3.1G and K). To exclude the possibility that the lack of intracellular infection is caused by heterozygosity of *P. andersonii* whereby only a non-functional allele was transmitted to the F_1 hybrid genotype, or by the particular rhizobium strain used for this experiment, we examined five independent F_1 hybrid plants inoculated with *M. plurifarium* BOR2 or *B. elkanii* WUR3. This revealed a lack of intracellular infection structures in nodules of all F_1 hybrid plants tested, irrespective which of the two rhizobium strains was used (Figure 3.1G and K and SI Appendix, Fig. S4), confirming that heterozygosity of *P. andersonii* does not play a role in the F_1 hybrid infection phenotype. These results suggest, at least partly, independent genetic control of nodule organogenesis and rhizobium infection. Because F_1 hybrids are nodulated with similar efficiency as *P. andersonii* (Figure 3.1B), we conclude that the network controlling nodule organogenesis is genetically dominant.

3.2.2 *Parasponia* and *Trema* genomes are highly similar

Based on preliminary genome size estimates made by using FACS measurements, three *Parasponia* and five *Trema* species were selected for comparative genome analysis (SI Appendix, Table S2). K-mer analysis of medium-coverage genome sequence data ($\sim$30$\times$) revealed that all genomes had low levels of heterozygosity, except those of *Trema levigata* and *T. orientalis* accession RG16 (SI Appendix, Fig. S5). Based on these k-mer data, we also generated more accurate estimates of genome sizes. Additionally, we used these data to assemble chloroplast genomes, based on which we obtained additional phylogenetic evidence that *T. levigata* is sister to *Parasponia* (Figure 3.1A and SI Appendix, Figs. S6-S8). Graph-based clustering of repetitive elements in the genomes (calibrated with the genome size estimates based on k-mers) revealed that all selected species contain approximately 300Mb of non repetitive sequence and a variable repeat content that correlates with the estimated genome size that ranges from 375 to 625 Mb (SI Appendix, Fig. S9 and Table S3). Notably, we found a *Parasponia*-specific expansion of ogre/tat LTR retro-transposons comprising 65-85Mb (SI Appendix, Fig. S9B). We then generated annotated reference genomes by using high-coverage ($\sim$125$\times$) sequencing of *P.andersonii* accession WU1 [47] and *T. orientalis* accession RG33 (SI Appendix, Tables S4 and S5). These species were selected based on their low heterozygosity levels in combination with relatively small genomes. *T. tomentosa* was not used for a high-quality genome assembly because it is an allotetraploid (SI Appendix, Fig. S5 and Tables S2 and S3).

We generated orthogroups for *P. andersonii* and *T. orientalis* genes and six other Eurosid species, including arabidopsis (*Arabidopsis thaliana*) and the legumes medicago and soybean. From both *P. andersonii* and *T. orientalis*, $\sim$35,000 genes could be clustered into $>$20,000 orthogroups (SI Appendix, Table S6 and DatasetS2; note that there can be multiple orthologous gene pairs per orthogroup). Within these orthogroups, we identified 25,605 *P. andersonii* - *T. orientalis* orthologous gene pairs based on phylogenetic analysis as well as whole-genome alignments (SI Appendix, Table

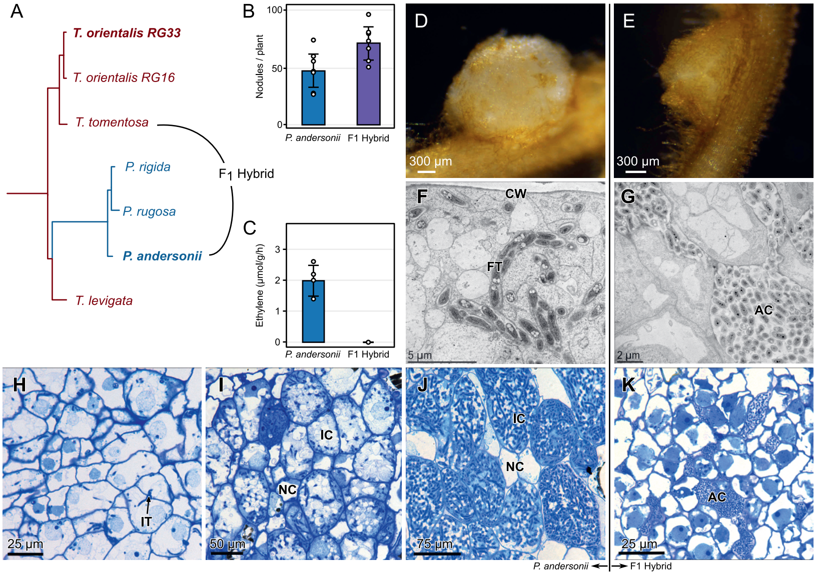


Figure 3.1: Nodulation phenotype of *P. andersonii* and interspecific *P. andersonii* × *T. tomentosa* F_1 hybrid plants. **(A)** Phylogenetic reconstruction based on whole chloroplast of *Parasponia* and *Trema*. The *Parasponia* lineage (blue) is embedded in the *Trema* genus (red). Species selected for interspecific crosses are indicated and species used for reference genome assembly are in bold. All nodes had a posterior probability of 1. **(B)** Mean number of nodules on roots of *P. andersonii* and F_1 hybrid plants ($n = 7$). **(C)** Mean nitrogenase activity in acetylene reductase assay of *P. andersonii* and F_1 hybrid nodules ($n = 4$). Bar-plot error bars indicate SDs; dots represent individual measurements. **(D)** *P. andersonii* nodule. **(E)** F_1 hybrid nodule. **(F&G)** Ultrastructure of nodule tissue of *P. andersonii* (F) and F_1 hybrid (G). Note the intracellular fixation thread (FT) in the cell of *P. andersonii* in comparison with the extracellular, apoplastic colonies of rhizobia (AC) in the F_1 hybrid nodule. **(H-J)** Light-microscopy images of *P. andersonii* nodules in three subsequent developmental stages. **(H)** Stage 1: initial infection threads (IT) enter the host cells. **(I)** Stage 2: progression of rhizobium infection in nodule host cell. **(J)** Stage 3: nodule cells completely filled with fixation threads. Note difference in size between the infected (IC) and noninfected cells (NC). **(K)** Light-microscopy image of F_1 hybrid nodule cells. Note rhizobium colonies in apoplast, surrounding the host cells (AC). Nodules have been analyzed 6 wk post inoculation with *M. plurifarium* BOR2. CW, cell wall.

S6). These orthologous gene pairs had a median percentage nucleotide identity of 97% for coding regions (SI Appendix, Figs. S10 and S11). This further supports the recent divergence of the two species and facilitates their genomic comparison.

3.2.3 Common utilization of symbiosis genes in *Parasponia* and *Medicago*

To assess commonalities in the utilization of symbiosis genes in *Parasponia* species and legumes, we employed two strategies. First, we performed phylogenetic analyses of close homologs of genes that were characterized to function in legume-rhizobium symbiosis. This revealed that *P. andersonii* contains putative orthologs of the vast majority of these legume symbiosis genes (96 of 126; Datasets S1 and S3). Second, we compared the sets of genes with enhanced expression in nodules of *P. andersonii* and medicago. RNA sequencing of *P. andersonii* nodules revealed 1,719 genes that are functionally annotated and have a significantly enhanced expression level ($foldchange \geq 2$, $p \leq 0.05$, DESeq2 Waldtest) in any of three nodule developmental stages compared with uninoculated roots (SI Appendix, Fig. S12 and Dataset S4). For medicago, we generated a comparable data set of 2,753 nodule-enhanced genes based on published RNA sequencing data [226]. We then determined the overlap of these two gene sets based on orthogroup membership and found that 382 orthogroups comprise both *P. andersonii* and medicago nodule-enhanced genes. This number is significantly greater than is to be expected by chance (permutation test, $p < 0.00001$; SI Appendix, Fig. S13 and Dataset S5). Based on phylogenetic analysis of these orthogroups, we found that in 290 cases putative orthologs have been utilized in *P. andersonii* and medicago root nodules (Datasets S5 and S6). Among these 290 commonly utilized genes are 26 putative orthologs of legume symbiosis genes, e.g. the LCO-responsive transcription factor NIN and its downstream target *NUCLEAR TRANSCRIPTIONFACTOR-YA1* (*NFYA1*) that are essential for nodule organogenesis [40, 161, 227, 228] and *RHIZOBIUM-DIRECTED POLAR GROWTH* (*RPG*) involved in intracellular infection [229]. Of these 26, five are known to function also in arbuscular mycorrhizal symbiosis (namely *VAPYRIN*, *SYMBIOTIC REMORIN*, the transcription factors *CYCLOPS* and *SAT1*, and a cysteine proteinase gene) [31, 230–236]. To further assess whether commonly utilized genes may be coopted from the ancient and widespread arbuscular mycorrhizal symbiosis, we determined which fraction is also induced upon mycorrhization in medicago based on published RNA sequencing data [237]. This revealed that only 8% of the commonly utilized genes have such induction in both symbioses (Dataset S5).

By exploiting the insight that nodule organogenesis and rhizobial infection can be genetically dissected using hybrid plants, we classified these commonly utilized genes into two categories based on their expression profiles in roots and nodules of both *P. andersonii* and F_1 hybrids (Figure 3.2). The first category comprises 126 genes that are up-regulated in both *P. andersonii* and hybrid nodules and that we associate with nodule organogenesis. The second category comprises 164 genes that are up-regulated in only the *P. andersonii* nodule and that we therefore associate with infection and/or fixation (Dataset S5). Based on these results, we conclude that *Parasponia* and medicago utilize orthologous genes that commit various functions in at least two different developmental stages of the root nodule.

3.2.4 Lineage-specific adaptation in *Parasponia* HEMOGLOBIN 1

Notable exceptions to the pattern of common utilization in root nodules are the oxygen-binding hemoglobins. Earlier studies showed that *Parasponia* and legumes have recruited different hemoglobin genes [238]. Whereas legumes use class II *LEGHEMOGLOBIN* to control oxygen homeostasis, *Parasponia* recruited the paralogous class I *HEMOGLOBIN 1* (*HB1*) for this function (Figure 3.3A and B). Biochemical studies have revealed that *P. andersonii* PanHB1 has oxygen affinities and kinetics that are adapted to their symbiotic function, whereas this is not the case for *T. tomentosa* TtoHB1 [238, 239]. We therefore examined *HB1* from *Parasponia* species, *Trema* species, and other nonsymbiotic Rosales species to see if these differences are caused by a gain of function in *Parasponia* or a loss of function in the non symbiotic species. Based on protein alignment, we identified *Parasponia*-specific adaptations in 7 amino acids (Figure 3.3C and D). Among these is Ile(101), for which it is speculated to be causal for a functional change in *P. andersonii* HB1 [239]. Hemoglobin-controlled oxygen homeostasis is crucial to protect the rhizobial nitrogen-fixing enzyme complex Nitrogenase in legume rhizobium-infected nodule cells [37, 214]. Therefore, *Parasponia*-specific gain of function adaptations in HB1 may have comprised an essential evolutionary step toward functional nitrogen-fixing root nodules with rhizobium endosymbionts.

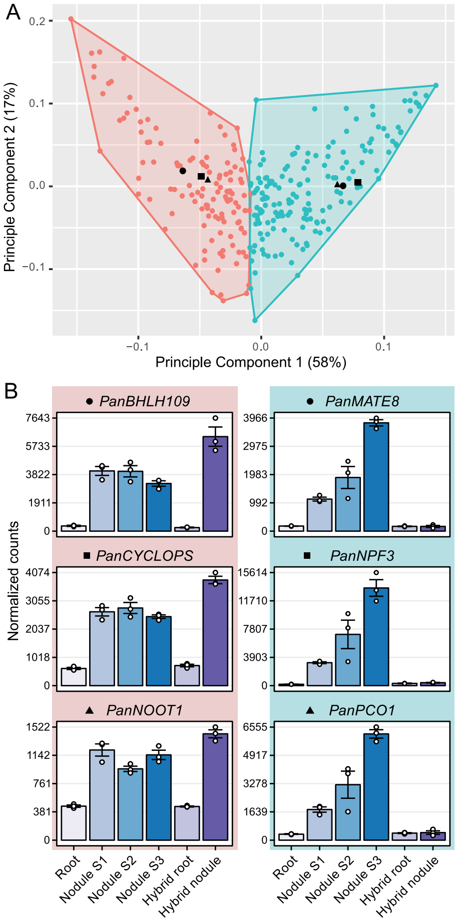


Figure 3.2: Clustering of commonly utilized symbiosis genes based on expression profile in *P. andersonii*. **(A)** Principal component analysis plot of the expression profile of 290 commonly utilized symbiosis genes in 18 transcriptome samples: *P. andersonii* roots and nodules (stage 13) and hybrid roots and nodules (line H9). All samples have three biological replicates. The first two components are shown, representing 75% of the variation in all samples. Colors indicate clusters (k-means clustering using Pearson correlation as distance measure, k = 2) of genes with similar expression patterns. The three genes with the highest Pearson correlation to the cluster centroids are indicated as black dots, triangles, and squares, and their expression profiles are given in **B**. Cluster 1 (pink) represents genes related to nodule organogenesis: these genes are up-regulated in *P. andersonii* and hybrid nodules. Cluster 2 (green) represents genes related to infection and fixation: these genes are highly up-regulated in *P. andersonii* nodules but do not respond in the hybrid nodule. PanBHLH109, BASIC HELIXLOOPHELIX DOMAIN CONTAINING PROTEIN 109; PanMATE8, MULTI ANTIMICROBIAL EXTRUSION PROTEIN 8; PanNOOT1, NODULE ROOT 1; PanNPF3, NITRATE/PEPTIDE TRANSPORTER FAMILY 3; PanPCO1, PLANT CYSTEINE OXIDASE 1.

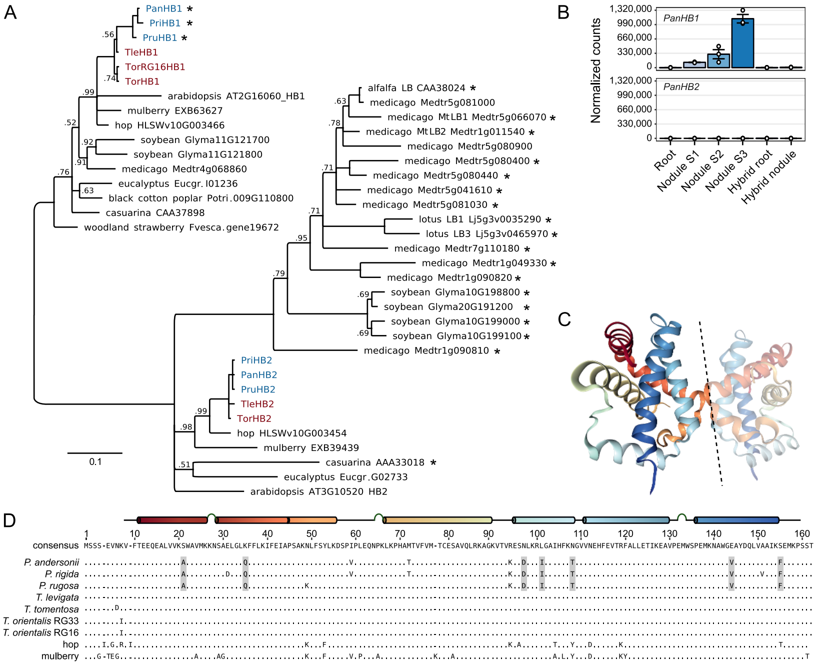


Figure 3.3: *Parasponia*-specific adaptations in class 1 hemoglobin protein HB1. **(A)** Phylogenetic reconstruction of class 1 (OG0010523) and class 2 hemoglobins (OG0002188). Symbiotic hemoglobins are marked with an asterisk; legumes and the actinorhizal plant casuarina have recruited class 2 hemoglobins for balancing oxygen levels in their nodules. Conversely, *Parasponia* has recruited a class 1 hemoglobin PanHB1, confirming parallel evolution of symbiotic oxygen transport in this lineage: *M. truncatula* (Medtr), *G. max* (Glyma), *P. trichocarpa* (Potri), *F. vesca* (Fvesca), *E. grandis* (Eugr), *A. thaliana* (AT). Node values indicate posterior probabilities below 1; scale bar represents substitutions per site. *Parasponia* marked in blue, *Trema* in red. **(B)** Expression profile of PanHB1 and PanHB2 in *P. andersonii* roots, stage 13 nodules, and *P. andersonii* × *T. tomentosa* F_1 hybrid roots and nodules (line H9). Expression is given in DESeq2-normalized read counts; error bars represent SE of three biological replicates and dots represent individual expression levels. **(C)** Crystal structure of the asymmetric dimer of PanHB1 as deduced by Kakar et al. Dashed line separates the two units. **(D)** Protein sequence alignment of class 1 hemoglobins from *Parasponia spp.*, *Trema spp.*, hop (*H. lupulus*), and mulberry (*M. notabilis*). Only amino acids that differ from the consensus are drawn. A linear model of the crystal structure showing α-helices and turns is depicted above the consensus sequence. There are seven amino acids (marked gray) that consistently differ between all *Parasponia* and all other sampled species: Ala(21), Gln(35), Asp(97), Ile(101), Thr(108), Val(144), and Phe(155). These differences therefore correlate with the functional divergence between *P. andersonii* PanHB1 and *T. tomentosa* TtoHB1 [238] & [239].

3.2.5 Parallel loss of symbiosis genes in *Trema* and other relatives of *Parasponia*

Evolution of complex genetic traits is often associated with gene copy number variations (CNVs) [240]. To test if CNVs were associated with the generally assumed independent evolution of nodulation in *Parasponia*, we focused on two genesets: (1) close homologs and putative orthologs of the genes that were characterized to function in legume-rhizobium symbiosis and (2) genes with a nodule-enhanced expression and functional annotation in *P. andersonii* (these sets partially overlap and together comprise 1,813 genes; SI Appendix, Fig. S14). We discarded *Trema*-specific duplications as we considered them irrelevant for the nodulation phenotype. To ensure that our findings are consistent between the *Parasponia* and *Trema* genera and not the result of species-specific events, we analyzed the additional draft genome assemblies of two *Parasponia* and two *Trema* species (SI Appendix, Table S5). As these additional draft genomes were relatively fragmented, we sought additional support for presence and absence of genes by mapping sequence reads to the *P. andersonii* and *T. orientalis* reference genomes and by genomic alignments. This procedure revealed only 11 consistent CNVs in the 1,813 symbiosis genes examined, further supporting the recent divergence between *Parasponia* and *Trema* (SI Appendix, Fig. S15). Because of the dominant inheritance of nodule organogenesis in F_1 hybrid plants, we anticipated finding Parasponia-specific gene duplications that could be uniquely associated with nodulation. Surprisingly, we found only one consistent Parasponia-specific duplication in symbiosis genes, namely, for a *HYDROXYCINNAMOYL-COASHIKIMATE TRANSFERASE* (*HCT*; SI Appendix, Figs. S16 and S17). This gene has been investigated in the legume forage crop alfalfa (*Medicago sativa*), in which it was shown that *HCT* expression correlates negatively with nodule organogenesis [241, 242]. Therefore, we do not consider this duplication relevant for the nodulation capacity of *Parasponia*. Additionally, we identified three consistent gene losses in *Parasponia*, among which is the ortholog of *EXOPOLYSACCHARIDE RECEPTOR 3* that, in lotus, inhibits infection of rhizobia with incompatible exopolysaccharides [243, 244] (SI Appendix, Figs. S18-S20 and Table S7). Such gene losses may have contributed to effective rhizobium infection in *Parasponia*, and their presence in *T. tomentosa* could explain the lack of intracellular infection in the F_1 hybrid nodules. However, they can not explain the dominance of nodule organogenesis in the F_1 hybrid.

Contrary to our initial expectations, we discovered consistent loss or pseudogenization of seven symbiosis genes in *Trema* (SI Appendix, Figs.

S21-S23 and Table S7). Based on our current sampling, these genes have a nodule-specific expression profile in *P. andersonii*, suggesting that they function exclusively in symbiosis (Figure 3.4). Three of these are orthologs of genes that are essential for establishment of nitrogen-fixing nodules in legumes: *NIN*, *RPG*, and the LysM-type LCO receptor *NFP/NFR5*. In the case of *NFP/NFR5*, we found two close homologs of this gene, *NFP1* and *NFP2*, a duplication that predates the divergence of legumes and *Parasponia* (Figure 3.5). In contrast to *NFP1*, *NFP2* is consistently pseudogenized in *Trema* species (Figure 3.5 and SI Appendix, Figs. S22 and S23). In an earlier study, we used RNAi to target *PanNFP1* (previously named *PaNFP*), which led to reduced nodule numbers and a block of intracellular infection by rhizobia as well as arbuscular mycorrhiza [47]. However, we can not rule out that the RNAi construct unintentionally also targeted PanNFP2, as both genes are 70% identical in the 422bp RNAi target region. Therefore, the precise functioning of both receptors in rhizobium and mycorrhizal symbiosis remains to be elucidated. Based on phylogenetic analysis, the newly discovered *PanNFP2* is the ortholog of the legume *MtNFP/LjNFR5* genes encoding rhizobium LCO receptors required for nodulation, whereas *PanNFP1* is most likely a paralog (Figure 3.5). Also, *PanNFP2* is significantly more highly expressed in nodules than PanNFP1 (SI Appendix, Fig. S25). Taken together, this indicates that *PanNFP2* may represent a key LCO receptor required for nodulation in *Parasponia*.

Based on expression profiles and phylogenetic relationships, we also postulate that *Parasponia NIN* and *RPG* commit essential symbiotic functions similarly as in other nodulating species (Figure 3.3 and SI Appendix, Figs. S25-S28) [159–161, 229, 245]. Compared with uninoculated roots, expression of *PanRPG* is >300-fold higher in *P. andersonii* nodules that become intracellularly infected (nodule stage 2), whereas, in F_1 hybrid nodules, which are devoid of intracellular rhizobium infection, this difference is less than 20-fold (Figure 3.4). This suggests that *PanRPG* commits a function in rhizobium infection, similarly as found in medicago [229]. The transcription factor *NIN* has been studied in several legume species as well as in the actinorhizal plant casuarina (*Casuarina glauca*) and, in all cases, shown to be essential for nodule organogenesis [159–161, 245]. Loss of *NIN* and possibly *NFP2* in *Trema* species can explain the genetic dominance of nodule organogenesis in the *Parasponia* × *Trema* F_1 hybrid plants.

Next, we assessed whether loss of these symbiosis genes also occurred in more distant relatives of *Parasponia*. We analyzed non nodulating species representing six additional lineages of the Rosales clade,

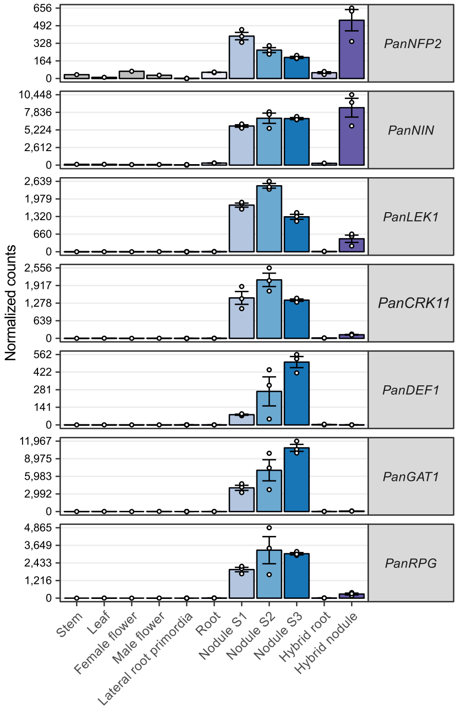


Figure 3.4: Expression profile of *P. andersonii* symbiosis genes that are lost in *Trema* species. Expression of symbiosis genes in *P. andersonii* stem, leaf, female and male flowers, lateral root primordia, roots, and three nodule stages (S1S3), and in F_1 hybrid roots and nodules (line H9). Expression is given in DESeq2-normalized read counts; error bars represent SE of three biological replicates for lateral root primordia, root, and nodule samples. Dots represent individual expression levels. *PanCRK11*, *CYSTEINE-RICH RECEPTOR KINASE 11*; *PanDEF1*, *DEFENSIN 1*; *PanLEK1*, *LECTIN RECEPTOR KINASE 1*; *PanNFP2*, *NOD FACTOR PERCEPTION 2*; *PanNIN*, *NODULE INCEPTION*; *PanRPG*, *RHIZOBIUM DIRECTED POLAR GROWTH*.

namely hop (*Humulus lupulus*, Cannabaceae) [246], mulberry (*Morus notabilis*, Moraceae) [247], jujube (*Ziziphus jujuba*, Rhamnaceae) [248], peach (*Prunus persica*, Rosaceae) [71], woodland strawberry (*Fragaria vesca*, Rosaceae) [249], and apple (*Malus* × *domestica*, Rosaceae) [70]. This revealed a consistent pattern of pseudogenization or loss of *NFP2*, *NIN*, and *RPG* orthologs, the intact jujube *ZjNIN* being the only exception (Figure 3.6). We note that, for peach, *NIN* was previously annotated as a protein-coding gene [71]. However, based on comparative analysis of conserved exon structures, we found two out-of-frame mutations (SI Appendix, Fig. S28). We therefore conclude that the *NIN* gene is also pseudogenized in peach. Because the pseudogenized symbiosis genes are largely intact in most of these species and differ in their deleterious mutations, the loss of function of these essential symbiosis genes should have occurred relatively recently and in parallel in at least seven Rosales lineages.

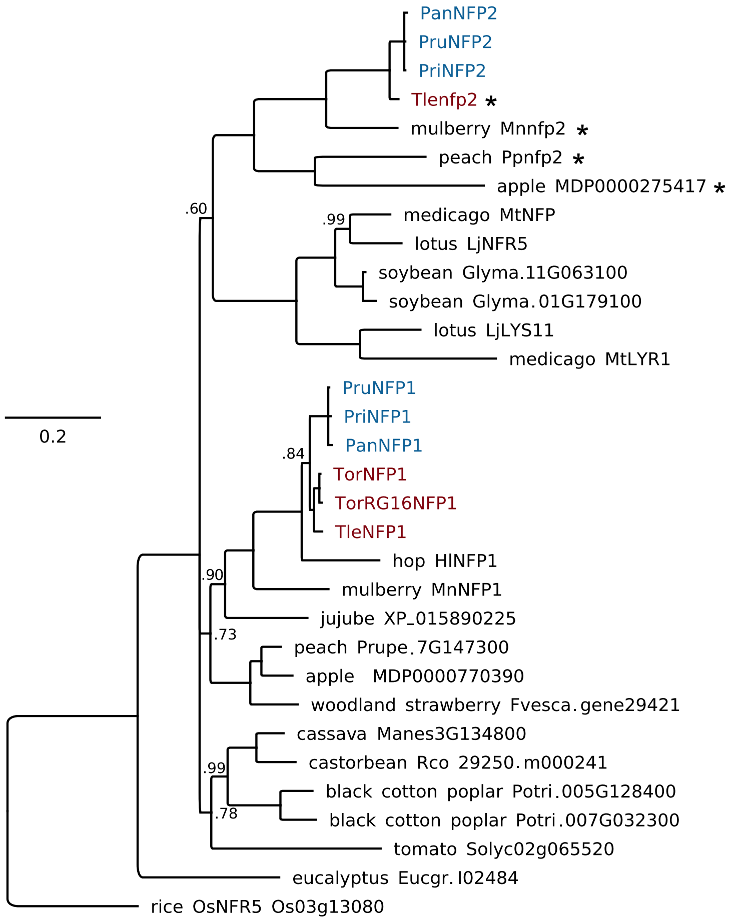


Figure 3.5: *Parasponia NFP2* are putative orthologs of legume LCO receptors *MtNFP/LjNFR5*. Phylogenetic reconstruction of the NFP/NFR5 orthogroup based on kinase domain. Protein sequences deduced from pseudogenes are marked with an asterisk. Included species are *Parasponia andersonii* (Pan), *Parasponia rigida* (Pri), *Parasponia rugosa* (Pru), *Trema orientalis* RG33 (Tor), *Trema orientalis* RG16 (TorRG16), *Trema levigata* (Tle), medicago (*Medicago truncatula*, Mt), lotus (*Lotus japonicus*, Lj), soybean (*Glycine max*, Glyma), peach (*Prunus persica*, Ppe), woodland strawberry (*Fragaria vesca*, Fvesca), black cotton poplar (*Populus trichocarpa*, Potri), eucalyptus (*Eucalyptus grandis*, Eugr), jujube (*Ziziphus jujuba*), apple (*Malus* × *domestica*), mulberry (*Morus notabilis*), hops (*Humulus lupulus*), cassava (*Manihot esculenta*), rice (*Oryza sativa*), tomato (*Solanum lycopersicum*), and castor bean (*Ricinus communis*). Node numbers indicate posterior probabilities below 1; scale bar represents substitutions per site. *Parasponia* proteins are marked in blue, *Trema* in red.

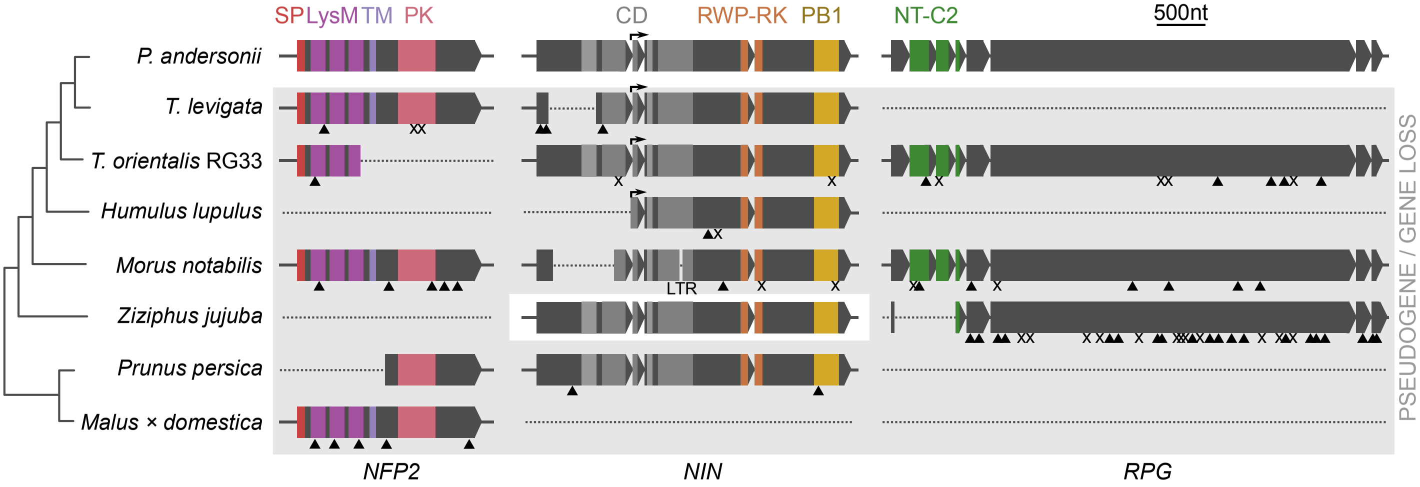


Figure 3.6: Parallel loss of symbiosis genes in nonnodulating Rosales species. Pseudogenization or loss of NFP2, NIN, and RPG in two phylogenetically independent *Trema* lineages, *Humulus lupulus* (hop), *Morus notabilis* (mulberry), *Ziziphus jujuba* (jujube), *Prunus persica* (peach), and *Malus* × *domestica* (apple). In *H. lupulus*, NIN is pseudogenized, whereas NFP2 and RPG were not found (this may be because of the low N50 of the publicly available assembly). In *Z. jujuba*, NFP2 is lost and RPG is pseudogenized, but NIN is intact. In *F. vesca*, all three genes are lost (not shown). Introns are indicated but not scaled. Triangles indicate frame shifts; "X" indicates premature stop codons; "LTR" indicates LTR retrotransposon insertion (not scaled); arrows indicate alternative transcriptional start site in NIN. CD, 4 conserved domains (gray); LysM, 3 Lysin motif domains (magenta); NT-C2, N-terminal C2 domain (green); PB1, Phox and Bem1 domain (yellow); PK, protein kinase (pink); RWP-RK, conserved amino acid domain (orange); SP, signal peptide (red); TM, transmembrane domain (lilac).

3.3 Discussion

Here we present the nodulating non legume *Parasponia* as a comparative system to obtain insights in molecular genetic changes underlying evolution of nitrogen-fixing root nodules. We show that nodulation is a genetically dominant trait and that *P. andersonii* and the legume medicago share a set of 290 genes that have a nodule-enhanced expression profile. Among these are NIN and RPG, two genes that, in legumes, are essential for nitrogen-fixing root nodulation [159–161, 245]. Both of these genes, as well as a putative ortholog of the NFP/NFR5-type LysM receptor for rhizobium LCO signal molecules – named NFP2 in *Parasponia* – are consistently pseudogenized or lost in *Trema* and other non nodulating species of the Rosales order. This challenges the current view on the evolution of nitrogen-fixing plant-microbe symbioses.

Evolution of nodulation is generally viewed as a two-step process: first an unspecified predisposition event in the ancestor of all nodulating species, bringing species in the nitrogen-fixing clade to a precursor state for nodulation; and subsequently, nodulation originated in parallel: eight times with *Frankia* and twice with rhizobium [9, 11–15, 216, 217]. This hypothesis is most parsimonious and suggests a minimum number of independent gains and losses of symbiosis. Based on this hypothesis, it is currently assumed that non host relatives of nodulating species are generally in a precursor state for nodulation [15].

Our results are difficult to explain under the hypothesis of parallel origins of nodulation. The functions of *NFP2*, *NIN*, and *RPG* currently can not be linked to any non symbiotic processes. Therefore, it remains obscure why these symbiosis genes were maintained over an extended period of time in non nodulating plant species and were subsequently independently lost. Additionally, the hypothesis of parallel origins of nodulation would imply convergent recruitmentof at least 290 genes to commit symbiotic functions in *Parasponia* and legumes. Because these 290 genes encode proteins with various predicted functions (i.e. from extracellular signaling receptors to sugar transporters; Dataset S5), as well as comprise at least two different developmental expression patterns (nodule organogenesis and intracellular infection and/or fixation; Figure 3.2 and Dataset S5), this would imply parallel evolution of a genetically complex trait.

Alternatively, the parallel loss of symbiosis genes in non nodulating plants can be interpreted as parallel loss of nodulation [9]. Under this hypothesis, nodulation possibly evolved only once in an ancestor of the nitrogen-fixing clade. Subsequently, nodulation was lost in most descendant lineages. This single-gain/massive-loss hypothesis fits our data better in two ways. First, a single gain explains the origin of the conserved set of at least 290 symbiosis genes utilized by *Parasponia* and medicago because they then result from the same ancestral recruitment event. Second, it more convincingly explains the parallel loss of symbiosis genes in non nodulating plants because then gene loss correlates directly with loss of nodulation. Additionally, the single-gain/massive-loss model eliminates the predisposition event, a theoretical concept that currently cannot be addressed experimentally. We therefore favor this alternative hypothesis over the currently most widely held assumption of parallel origins of nodulation.

Loss of nodulation is not controversial, as it is generally considered to have occurred at least 20 times in the legume family [11, 15]. Nevertheless, the single-gain/massive-loss hypothesis implies many more evolutionary events than the current hypothesis of parallel gains. On the contrary, it is conceptually easier to lose a complex trait, such as nodulation, than to gain it [216]. Genetic studies in legumes demonstrated that nitrogen-fixing symbioses can be abolished by a single knockout mutation in tens of different genes, among which are *NFP/NFR5*, *NIN*, and *RPG* (Dataset S1). Because parsimony implies equal weights for gains and losses, it therefore may not be the best way to model the evolution of nodulation.

Preliminary support for the single-gain/massive-loss hypothesis can be found in fossil records. Putative root nodule fossils have been discovered from the late Cretaceous (∼84Mya), which corroborates our hypothesis that nodulation is much older than is generally assumed [250]. Legumes are the oldest and most diverse nodulating lineage, but the earliest fossils that can be definitively assigned to the legume family appeared in the late Paleocene (∼65Mya) [251]. Notably, the age of the nodule fossils coincides with the early diversification of the nitrogen-fixing clade that has given rise to the four orders Fabales, Rosales, Cucurbitales, and Fagales [11]. As it is generally agreed that individual fossil ages provide minimum bounds for dates of origins, it is therefore not unlikely that the last common ancestor of the nitrogen-fixing clade was a nodulator.

Clearly, the single-gain/massive-loss hypothesis that is supported by our comparative studies with *Parasponia* requires further substantiation. First, the hypothesis implies that many ancestral species in the nitrogen-fixing clade were able to nodulate. This should be further supported by fossil evidence. Second, the hypothesis implies that actinorhizal plant species maintained *NIN*, *RPG*, and possibly *NFP2* (the latter only in case LCOs are used as symbiotic signal) [252]. Third, these genes should be essential for nodulation in these actinorhizal plants as well as in *Parasponia*.

This can be shown experimentally, as was done for *NIN* in casuarina [159].

Loss of symbiosis genes in non nodulating plant species is not absolute, as we observed a functional copy of *NIN* in jujube. This pattern is similar to the pattern of gene loss in species that lost endomycorrhizal symbiosis in which, occasionally, endomycorrhizal symbiosis genes have been maintained in non mycorrhizal plants [253, 254]. Conservation of *NIN* in jujube suggests that this gene has a non symbiotic function. Contrary to *NFP2*, which is the result of a gene duplication near the origin of the nitrogen-fixing clade, functional copies of *NIN* are also present in species outside the nitrogen-fixing clade (SI Appendix, Fig. S26). This suggests that these genes may have retained – at least in part – an unknown ancestral non symbiotic function in some lineages within the nitrogen-fixing clade. Alternatively, *NIN* may have acquired a new non-symbiotic function within some lineages in the nitrogen-fixing clade.

As hemoglobin is crucial for rhizobium symbiosis in legumes [214], it is striking that *Parasponia* and legumes do not use orthologous copies of hemoglobin genes in their nodules [238]. Superficially, this seems inconsistent with a single gain of nodulation. However, hemoglobin is not crucial for all nitrogen-fixing nodule symbioses because several *Frankia* microsymbionts possess intrinsic physical characteristics to protect the Nitrogenase enzyme for oxidation [255–257]. In line with this, *Ceanothus spp.* (Rhamnaceae, Rosales) – which represent actinorhizal nodulating relatives of *Parasponia* – do not express a hemoglobin gene in their *Frankia*-infected nodules [256–258]. Consequently, hemoglobins may have been recruited in parallel after the initial gain of nodulation as parallel adaptations to rhizobium microsymbionts. Based on the fact that *Parasponia* acquired lineage-specific adaptations in HB1 that are considered to be essential for controlling oxygen homeostasis in rhizobium root nodules [238, 239], a symbiont switch from *Frankia* to rhizobium may have occurred recently in an ancestor of the *Parasponia* lineage.

Our study provides leads for attempts to engineer nitrogen-fixing root nodules in agricultural crop plants. Such a translational approach is anticipated to be challenging [8], and the only published attempt so far, describing transfer of eight LCO signaling genes, was unsuccessful [42]. Our results suggest that transfer of symbiosis genes may not be sufficient to obtain functional nodules. Eventhough F_1 hybrid plants contain a full haploid genome complement of *P. andersonii*, they lack intracellular infection. This may be the result of haploinsufficiency of *P. andersonii* genes in the F_1 hybrid or because of an inhibitory factor in *T. tomentosa*. For example, inhibition of intracellular infection may be the result of a dominant-negative factor or the result of heterozygosity negatively affecting the formation of, e.g., LysM receptor complexes required for appropriate perception of microsymbionts. Such factors may also be present in other non host species. Consequently, engineering nitrogen-fixing nodules may require gene knockouts in non nodulating plants to overcome inhibition of intracellular infection. *Trema* may be the best candidate species for such a (re)engineering approach because of its high genetic similarity with *Parasponia* and the availability of transformation protocols [259]. Therefore, the *Parasponia*-*Trema* comparative system may not only be suited for evolutionary studies, but also can form an experimental platform to obtain essential insights for engineering nitrogen-fixing root nodules.

3.4 Materials and methods

3.4.1 *Parasponia*-*Trema* intergeneric crossing and hybrid genotyping

Parasponia and *Trema* are wind-pollinated species. A female-flowering *P. andersonii* individual WU1.14 was placed in a plastic shed together with a flowering *T. tomentosa* WU10 plant. Putative F_1 hybrid seeds were germinated (SI Appendix, Supplementary Methods) and transferred to potting soil. To confirm the hybrid genotype, a PCR marker was used that visualizes a length difference in the promoter region of LIKE-AUXIN 1 (LAX1; primers, LAX1-forward, ACATGATAATTTGGGCATGCAACA; LAX1-reverse, TCCCGAATTTTCTACGAATTGAAA; amplicon size, *P. andersonii* 974bp; *T. tomentosa* 483 bp). Hybrid plant H9 was propagated in vitro [47, 260]. The karyotype of the selected plants was determined according to [261].

3.4.2 Assembly of reference genomes

Cleaned DNA sequencing reads were de novo assembled by using ALLPATHS-LG (release 48961) [262]. After filtering of any remaining adapters and contamination, contigs were scaffolded with two rounds of SSPACE-standard (v3.0) [263] with the mate-pair libraries using default settings. We used the output of the second run of SSPACE scaffolding as the final assembly (full details and parameter choices are providedin SI Appendix, Supplementary Methods). Validation of the final assemblies showed that 90-100% of the genomic reads mapped back to the assemblies (SI Appendix, Table S4), and 94-98% of CEGMA [264] and BUSCO [265] genes were detected (SI Appendix, Table S5).

3.4.3 Annotation of reference genomes

Repetitive elements were identified following the standard Maker-P recipe (`http://weatherby.genetics.utah.edu/MAKER/wiki/index.php/Repeat_Library_Construction-Advanced`, accessed October 2015) as described on the GMOD site: (1) RepeatModeler with Repeatscout v1.0.5, Reconv1.08, RepeatMasker version open4.0.5, using RepBase version 2014-0131 [266] and TandemRepeatFinder; (2) GenomeTools LTRharvest and LTRdigest [267]; (3) MITEhunter with default parameters [268]. We generated species-specific repeat libraries for *P. andersonii* and *T. orientalis* separately and combined these into a single repeat library, filtering out sequences that are $>$98% similar. We masked both genomes by using RepeatMasker with this shared repeat library.

To aid the structural annotation, we used 11 *P. andersonii* and 6 *T. orientalis* RNA-sequencing (RNA-seq) datasets (SI Appendix, Table S8). All RNA-seq samples were assembled de novo by using genome-guided Trinity [269], resulting in one combined transcriptome assembly per species. In addition, all samples were mapped to their respective reference genomes by using BWA-MEM and processed into putative transcripts by using cufflinks [270] and transdecoder [271]. As protein homology evidence, only UniProt [272] entries filtered for plant proteins were used. This way we included only manually verified protein sequences and prevented the incorporation of erroneous predictions. Finally, four gene predictor tracks were used: (1) SNAP [273] trained on *P. andersonii* transdecoder transcript annotations; (2) SNAP trained on *T. orientalis* transdecoder transcript annotations; (3) Augustus [274], as used in the BRAKER pipeline, trained on RNA-seq alignments [275]; and (4) GeneMark-ET, as used in the BRAKER pipeline, trained on RNA-seq alignments [276].

First, all evidence tracks were processed by Maker-P [277]. The results were refined with EVidenceModeler (EVM) [278], which was used with all of the same tracks as Maker-P, except for the Maker-P blast tracks and with the addition of the Maker-P consensus track as additional evidence. Ultimately, EVM gene models were preferred over Maker-P gene models except when there was no overlapping EVM gene model. Where possible, evidence of both species was used to annotate each genome (e.g. de novo RNA-seq assemblies of both species were aligned to both genomes).

To take maximum advantage of annotating two highly similar genomes simultaneously, we developed a custom reconciliation procedure involving whole-genome alignments. The consensus annotations from merging the EVM and Maker-P annotations were transferred to their respective partner genome by using nucmer [279] and RATT revision 18 [280] (i.e., the *P. an-*

dersonii annotation was transferred to *T. orientalis* and vice versa) based on nucmer whole-genome alignments (SI Appendix, Fig. S10). Through this reciprocal transfer, both genomes had two candidate annotation tracks. This allowed for validation of annotation differences between *P. andersonii* and *T. orientalis*, reduced technical variation, and consequently improved all downstream analyses. After automatic annotation and reconciliation, 1,693 *P. andersonii* genes and 1,788 *T. orientalis* genes were manually curated. These were mainly homologs of legume symbiosis genes and genes that were selected based on initial data exploration.

To assign putative product names to the predicted genes, we combined BLAST [281] results against UniProt, TrEMBL, and nr with InterProScan [282] results (custom script). To annotate Gene Ontology (GO) [283] terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) [284] enzyme codes we used Blast2GO based on the nr BLAST results and InterProScan results. Finally, we filtered all gene models with hits to InterPro domains that are specific to repetitive elements.

3.4.4 Orthogroup inference

To determine relationships between *P. andersonii* and *T. orientalis* genes, as well as with other plant species, we inferred orthogroups with OrthoFinder version 0.4.0 [285]. As orthogroups are defined as the set of genes that are descended from a single gene in the last common ancestor of all of the species being considered, they can comprise orthologous as well as paralogous genes. Our analysis included proteomes of selected species from the Eurosid clade: *A. thaliana* TAIR10 (Brassicaceae,Brassicales) [286] and *Eucalyptus grandis* v2.0 (Myrtaceae, Myrtales) from the Malvid clade [287]; *Populus trichocarpa* v3.0 (Salicaeae, Malpighiales) [288], legumes *Medicago truncatula* Mt4.0v1 [18] and *Glycine max* Wm82.a2.v1 (Fabaceae, Fabales) [19], *Fragaria vesca* v1.1 (Rosaceae, Rosales) [249], and *Parasponia andersonii* and *Trema orientalis* (Cannabaceae, Rosales) from the Fabid clade (Dataset S2). Sequences were retrieved from phytozome [289].

3.4.5 Gene copy number variation detection

To assess orthologous and paralogous relationships between *Parasponia* and *Trema* genes, we inferred phylogenetic gene trees for all 21,959 orthogroups comprising *Parasponia* and/or *Trema* genes by using the neighbor-joining clustering algorithm [290]. Based on these gene trees, for each *Parasponia* gene, its relationship to other *Parasponia* and *Trema* genes was defined as follows: (1) orthologous pair indicates that the sister lineage is a single gene from the *Trema* genome, suggesting that they are the result of a speciation event; (2) inparalog indicates that the sister lineage is a gene from the *Parasponia* genome, suggesting that they are the result of a gene duplication event; (3) singleton indicates that the sister lineage is a gene from a species other than *Trema*, suggesting that the *Trema* gene was lost; and (4) multi-ortholog indicates that the sister lineage comprises multiple genes from the *Trema* genome, suggesting that the latter are in paralogs. For each *Trema* gene, the relationship was defined in the same way but with respect to the *Parasponia* genome (SI Appendix, TableS6). Because phylogenetic analysis relies on homology, we assessed the level of conservation in the multiple sequence alignments by calculating the trident score using MstatX [291]. Orthogroups with $trident\_score \leq 0.1$ were excluded from the analysis. Examination of orthogroups comprising $>$20 inparalogs revealed that some represented repetitive elements; these were also excluded. Finally, orthologous pairs were validated based on the whole-genome alignments used in the annotation reconciliation.

3.4.6 Nodule-enhanced genes

To assess gene expression in *Parasponia* nodules, RNA was sequenced from the three nodule stages described earlier as well as uninoculated roots (SI Appendix, TableS8). RNA-seq reads were mapped to the *Parasponia* reference genome with HISAT2 version 2.02 [292] using an index that includes exon and splice site information in the RNA-seq alignments. Mapped reads were assigned to transcripts with featureCounts version1.5.0 [293]. Normalization and differential gene expression were performed with DESeq2. Nodule enhanced genes were selected based on $foldchange \geq 2$ and $p \leq 0.05$ in any nodule stage compared with uninoculated root controls. Genes without functional annotation or orthogroup membership or from orthogroups with low alignment scores ($trident\_score \leq 0.1$, as detailed earlier) or representing repetitive elements were excluded from further analysis. To assess expression of *Parasponia* genes in the hybrid nodules, RNA was sequenced from nodules and uninoculated roots. Here, RNA-seq reads were mapped to a combined reference comprising two parent genomes from *P. andersonii* and *T. tomentosa*. To assess which genes are nodule-enhanced in medicago, we reanalyzed published RNA-seq read data from Roux et al. [226] [archived at the National Center for Biotechnology Information (NCBI) under sequence read archive (SRA) study ID code SRP028599]. To assess which of these genes may be coopted from the ancient and widespread arbuscular mycorrhizal symbiosis, we generated a set of 575 medicago genes induced upon mycorrhization in medicago by reanalyzing published RNA-seq read data from Afkhami and Stinchcombe [237] [archived at the NCBI under SRA study ID code SRP078249]. Both medicago data sets were analyzed as described earlier for *Parasponia* but by using the medicago genome and annotation version 4.0v2 as reference [18].

To assess common recruitment of genes in nodules from *Parasponia* and medicago, we counted orthogroups comprising *P. andersonii* and medicago nodule-enhanced genes. To assess whether this number is higher than expected by chance, we performed the hypergeometric test as well as three different permutation tests in which we randomized the *Parasponia* gene set, the medicago gene set, or both sets with 10,000 permutations. We then determined putative orthology between the *Parasponia* and medicago genes within the common orthogroups based on phylogenetic analysis. *Parasponia* and medicago genes were considered putative orthogroups if they occurred in the same subclade with more than 50% bootstrap support; otherwise, they were considered close homologs.

Availability of data and materials

The data reported in this study are tabulated in Datasets S1-S7 and SI Appendix; sequence data are archived at NCBI under BioProject numbers PRJNA272473 and PRJNA272482; draft genome assemblies, phylogenetic datasets, and orthogroup data are archived at the Dryad Digital Repository (`https://doi.org/10.5061/dryad.fq7gv88`). Analyzed data can also be browsed or downloaded through a Web portal at `www.parasponia.org`. All custom scripts and code are available online at `https://github.com/holmrenser/parasponia_code`.

Acknowledgements

We thank Shelley James, Thomas Marler, Giles Oldroyd, and Johan van Valkenburg for providing germplasm and Ries de Visser (IsoLife) for supporting acetylene reduction assays.

Chapter 4

Comparative analysis of symbiotic transcriptional networks

Rens Holmer, Wouter Kohlen, Dick de Ridder, René Geurts, and Sandra Smit

Supplementary information is available at
https://github.com/holmrenser/transcriptional_networks_nodulation

Abstract

Nitrogen-fixing plant-microbe symbioses are mutualistic interactions that provide nitrogen sources in the form of ammonium to the plant. Among the most studied forms of this symbiosis are the interactions of rhizobium bacteria with legumes, and with *Parasponia* (Cannabaceae). To provide a protective environment, bacteria are housed intracellularly nodules – dedicated root organs that are formed upon perception of the bacteria. Phylogenomic studies have uncovered that legumes and *Parasponia* likely represent a single evolutionary origin of nodulation. Additionally, the transcription factor *NODULE INCEPTION* (*NIN*) has been implied to be involved in the molecular evolution of nodulation. This raises the hypothesis that transcriptional rewiring has been an important process in evolving nitrogen-fixing symbiosis. However, current knowledge on symbiotic transcriptional networks is sparse, and it is unknown whether these interactions are unique to symbiotic species. To overcome these hurdles, we attempted to reconstruct transcriptional networks at a genome-wide scale for three symbiotic and four non-symbiotic plant species using publicly available RNA sequencing data. Additionally, we present a data integration scheme for the comparative analysis of transcriptional networks in multiple species using groups of orthologous genes. To our disappointment, we were unable to identify conserved transcriptional interactions in our predicted networks. This led us to rigorously benchmark the predictive performance our transcriptional networks, resulting in an estimated false positive rate of 90%-99%. From this we conclude that current high-throughput methods for predicting transcriptional networks are unsuitable for comparative analysis. Future studies on transcriptional rewiring during the evolution of nodulation will have to focus on using RNA sequencing data with a higher resolution, or direct measurements of transcroptional interaction such as ChIP-seq, DAP-seq or DNAse-seq.

4.1 Introduction

Most legumes, the cannabis-relative *Parasponia*, and actinorhizal plants form a root nodule endosymbiosis with nitrogen-fixing bacteria [21]. Through this symbiosis, plants obtain ammonium as essential nitrogen nutrients from rhizobium or *Frankia* bacteria that are housed intracellularly in root nodules. To better understand the molecular mechanisms behind nitrogen-fixing nodule symbiosis, it is necessary to understand its evolution [294]. Phylogenomic studies have revealed that this mutualistic relationship in plant species from ten distinct clades likely has a single evolutionary origin [213, 295]. As a consequence, it is hypothesized that the molecular mechanisms required for rhizobium symbiosis in legumes and *Parasponia* are very similar. The perception of bacterial lipochitooligosaccharide (LCO) signalling molecules by orthologous copies of NFP in *Medicago*, *Lotus*, and *Parasponia* supports this notion of conserved signalling mechanisms [25, 26, 47, 296]. To achieve a fully mutualistic symbiosis, the interaction between plants and rhizobium bacteria is further characterized by a cascade of transcriptional regulators, most of which are essential [148, 161, 297]. Previous work in legumes has revealed that a variety of transcription factors act in concert during different stages of nodule organogenesis and bacterial infection [40]. Upon rhizobium induced LCO signalling, CYCLOPS binds to the proximal promoter region of NODULE INCEPTION (NIN) [32], which in turn can activate expression of NUCLEAR FACTOR GAMMA A1 (NFYA1) [33], which itself is a transcription factor [227]. Additionally, the NIN promoter is likely targeted by cytokinin responsive RESPONSE REGULATOR (RR) transcription factors in more inner cell layers of the root [43]. Several transcriptional coregulators in the form of GRAS proteins are likely involved in the above processes, such as *NODULATION SIGNALLING PATHWAY 1* and *2* (*NSP1*, *NSP2*), *SCARECROW13-LIKE INVOLVED IN NODULATION* (*SIN1*), and several *DELLA* proteins [203, 298, 299]. In summary, multiple transcriptional regulators involved in symbiotic transcriptional networks have been identified in legumes.

Recently it was found that within the so-called nitrogen-fixing clade that exists of four orders Fabales, Rosales, Cucurbitales and Fagales, occurrence of nitrogen-fixing nodule symbiosis strongly correlates with the presence of the transcription factor *NIN* [213, 295]. Whereas this supports the notion that conserved signalling networks underlie rhizobium symbiosis, there are two exceptions to this pattern. The first exception is *Ziziphus jujuba*, a shrub that has an ortholog of *NIN*, but does not engage in rhizobium symbiosis [213]. The second exception is all plant species outside of the nitrogen-fixing clade clade, where every plant species has one or more orthologs of the known nodulation genes, but does not engage in rhizobium symbiosis [295]. These exceptions suggest that absence or presence of transcriptional regulators is not sufficient to explain their mechanistic workings. Instead, a process of rewiring transcriptional networks seems to be involved in the evolution of rhizobium symbiosis [40, 300], a concept generally referred to as deep homology [301]. To study the evolution of symbiotic transcriptional networks, instead of identifying absence and presence of individual genes, it is necessary to identify absence and presence of interactions between transcriptional regulators and their targets. To achieve a similar comparative perspective as was previously done for individual genes, regulatory interactions between orthologs of known nodulation genes should be identified in a range of symbiotic and non-symbiotic plants. However, transcriptional interactions between nodulation genes in plant species other than legumes remain largely unstudied. Additionally, it is currently not feasible to investigate multiple regulatory interactions in a range of species within reasonable time using classical mutagenesis or binding affinity studies. Instead, we used a computational strategy for comparative analysis of symbiotic transcriptional networks to better understand the evolution of rhizobium symbiosis.

Several computational approaches have been developed to infer transcriptional networks from high throughput transcriptome measurements (see [302] for a review). At their core, these methods rely on the observation that a single transcriptome measurement is a static read-out of a dynamic transcriptional network. It is then assumed that given sufficient static transcriptome measurements, it is possible to reverse engineer the transcriptional network that has generated the measured transcriptomes [303]. There are two types of approaches for predicting transcriptional networks from transcriptome measurements, one based on co-expression and one based on direct inference [304]. The co-expression based approaches rely on the assumption that genes with a similar expression profile are regulated by the same transcription factors. Co-expression analysis has been used extensively in plants (e.g. [305, 306]), but transcriptional networks inferred from co-expression alone are generally not very precise and thus often require additional processing [304]. In contrast, direct prediction of individual regulatory interactions on a genome wide scale provides a more mechanistic model of transcriptional regulation [304]. Additionally, direct prediction is advantageous in that it provides testable hypotheses, and allows for direct comparison of regulatory interactions between species. But, due to previous computational limitations these direct approaches have not found much application in plants (see [307] for an overview). Recent developments in machine learning methods have unlocked the possibility to directly predict transcriptional networks from transcriptome sequencing data with reasonable computational requirements [302, 308, 309]. Previously, the use of these methods has been limited to individual organisms under specific circumstances [310], or limited numbers of genes [311, 312]. Whereas most of the methods used for predicting transcriptional networks have been validated on unicellular organisms using the DREAM5 benchmark [302], the estimation of performance for multicellular organisms is often neglected. As a result, it is not well understood what the predictive accuracy of these tools is when applied to multicellular organisms such as plants. Currently, a large amount of public transcriptome data is available for multiple plant species, and recent computational advances make it possible to use this data for transcriptional network prediction. Therefore, it should now be possible to provide a comparative perspective on symbiotic transcriptional networks using computational methods. Here, we present a conceptual framework for the comparative analysis of transcriptional regulation. We apply this framework to public transcriptome sequencing data of seven plant species to study whether conserved transcriptional networks underlie rhizobium symbiosis. Finally, we present a validation strategy to critically assess our predicted transcriptional networks. Taken together, our approach highlights the possibilities and pitfalls of predicted transcriptional networks in studying the evolution of rhizobium symbiosis.

4.2 Results

4.2.1 Genome-wide comparative analysis of transcriptional networks

To identify conserved transcriptional networks in multiple species, we devised a comparative strategy that makes use of transcriptional networks predicted from transcriptome data and orthogroups. In summary, we identified a reduced representation of the predicted interactions by grouping together all interactions between genes of the same orthogroup in the same species into a single predicted interaction (figure 4.1). By comparing interactions between (genes from) orthogroups, we can effectively compare interactions between species. It should be noted that in doing so, we lose some resolution in the case of species-specific gene duplications. However, this effect is likely not very big since the number of orthogroups per species and the number of genes per species are similar for all species in this study (table 4.1). With this approach, we can perform a phylogenomic analysis of transcriptional regulation by scoring absence and presence of transcriptional interactions in multiple species.

Using between 111 and 2,965 publicly available RNA-seq samples per species we predicted transcriptional networks for the following seven plant species: *Medicago truncatula*, *Parasponia andersonii*, and *Glycine max* (three species that can form nodules) and *Arabidopsis thaliana*, *Cucumis sativa*, *Populus trichocarpa*, and *Ziziphus jujuba* (four species that cannot

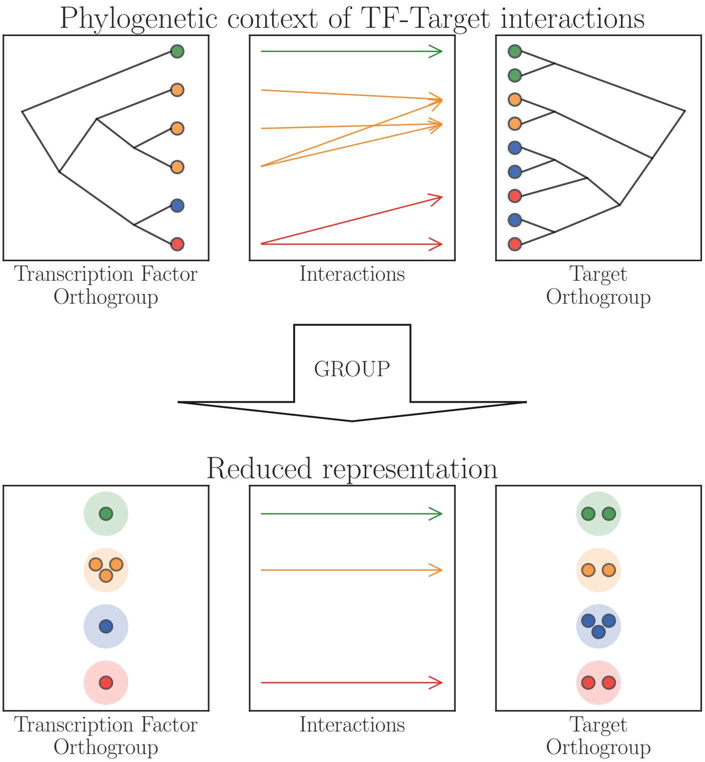

Phylogenetic context of TF-Target interactions
Transcription Factor Orthogroup
Interactions
Target Orthogroup
GROUP
Reduced representation
Transcription Factor Orthogroup
Interactions
Target Orthogroup

Figure 4.1: Schematic representation of the approach taken in this study to compare predicted interactions between transcription factors and target genes. With this approach, transcriptional networks can be placed in a phylogenetic context using orthogroups. Colours indicate species, arrows indicate predicted transcriptional interactions, and trees indicate evolutionary relationships. Grouping orthogroup genes by species allows the scoring of absence/presence patterns of transcriptional regulation.

form nodules) (table 4.1). We followed the standard approach of directly predicting transcriptional networks with tree-based regression, using the efficient implementation of GRNBoost2 [309]. The most likely regulators of a gene are selected by regressing that genes expression levels to the expression levels of multiple candidate transcription factors. This results in a set of links between all predicted transcription factors and their candidate targets, which we interpret as a transcriptional network. In addition, every predicted transcriptional interaction is assigned a weight from the regression analysis. As described in the GRNBoost2 publication [309], these weights are interpreted such that higher weights are more likely to represent true interactions. More formally, a weighted transcriptional network can be seen as an $m \times n$ matrix A, with m genes and n transcription factors, where n is a subset of m, and A_{mn} is the weight assigned to the predicted interaction between transcription factor n and gene m. By assigning a value of 1 to all interactions above a certain cut-off a weighted transcriptional network can be turned into a boolean matrix, where 1 indicates interaction and 0 indicates no interaction.

On average, most genes get between 600-800 predicted transcription factors each assigned to them with non-zero weights (figure 4.2A). Conversely, most transcription factors are predicted to have between 10,000 and 30,000 targets each (figure 4.2B). For both of these statistics we observed some variation between species. Furthermore, for the number of predicted transcription factors per gene, we found a bimodal distribution in all species. A fraction of the genes (4% - 17%) get between 0-200 predicted transcription factors, substantially lower than the 600-800 predicted transcription factors that is assigned to the rest of the genes. Overall, the predicted transcriptional networks are very densely connected, with between 25 million and 66 million predicted transcriptional interactions per species (table 4.1). This translates to GRNBoost2 assigning a non-zero probability of a transcription factor regulating a target gene for 42% - 81% of all possible interactions. These numbers are roughly 100x higher than what is observed in chromatin immunoprecipitation sequencing (ChIP-seq) experiments in *Arabidopsis thaliana* for both the number of transcription factors per gene and the number of targets per transcription factor [313].

4.2.2 Comparative analysis of transcriptional networks involved in rhizobium symbiosis

To investigate the feasibility of using the predicted transcriptional networks for studying transcriptional regulation in rhizobium symbiosis, we

Table 4.1: Overview of available data for the seven plant species used in this study

Species	Nod.[1]	Genes	Samples[2]	OGs[3]	TFs[4]	Pred. int.[5]
M. truncatula	Yes	32,814	906	32,095	3,156	43,525,698
P. andersonii	Yes	30,590	111	28,503	1,477	25,218,904
G. max	Yes	26,213	2,743	26,173	3,891	66,910,792
A. thaliana	No	19,275	2,295	19,235	1,839	27,254,697
C. sativa	No	16,720	533	16,700	1,579	21,426,012
P. trichocarpa	No	26,086	2,965	25,937	2,055	42,037,128
Z. jujuba	No	20,213	135	19,756	2,488	33,719,108

[1] Nodulating
[2] See suppl. inf. for full sample information
[3] Orthogroups
[4] Transcription factors
[5] Predicted interactions between transcription factors and targets

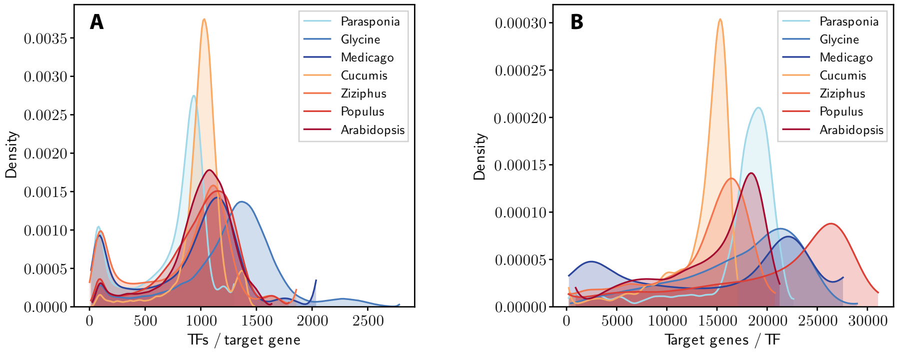


Figure 4.2: Distributions of (**A**) the number of predicted transcription factors per gene and (**B**) the number of predicted targets per transcription factor for the seven species in this study represented in a density plot

turned to a set of fourteen known interactions between symbiotic transcription factors and their targets from various legume species (table 4.2). It should be noted that most of these interactions have only been shown in one legume species, often other than *Medicago truncatula*, and that they are based on a variety of techniques. We retrieved all known interactions in *M. truncatula* from our predicted transcriptional networks. However, true interactions were never ranked as top predicted interaction. In fact, most of the known interactions received low weights from the regression analysis and were ranked between 7^{th} and 787^{th} likely regulator, where twelve of the fourteen interactions were ranked outside of the top 20 interactions (table 4.2). Furthermore, whereas the network of known interactions in *M. truncatula* is sparse, the predicted network for *M. truncatula* is highly connected (figure 4.3A&B). Additionally, the predicted networks of the same genes in other species are visually dissimilar from each other. Surprisingly, where we expected to find similar predictions between the closely related legume species *M. truncatula* and *Glycine max*, their predicted networks are quite dissimilar (figure 4.3A&D).

Upon further visual inspection, no single predicted network is similar to any of the other predicted networks. To quantify this difference, we calculated the Adjusted Rand Index (ARI) between the four nodulation genes networks. Almost all combinations of networks have an ARI close to zero, indicating little to no resemblance (figure 4.4). Two combinations have some similarity: the predicted networks of *Medicago truncatula* and *Parasponia andersonii*, and the predicted networks of *Arabidopsis thaliana* and Ziziphus jujube have and ARI of 0.45 and 0.22 respectively, indicating

Table 4.2: Overview of known transcriptional interactions in legumes and their corresponding weights and ranks in the predicted transcriptional network

Interaction[1]	**Weight**	**Rank**[2]	**Original publication**	**Original species**
NIN → NFYA1	0.1036	28	Soyano et al. [33]	*L. japonicus*
NIN → CRE1	0.1630	19	Vernié et al. [218]	*M. truncatula*
NIN → RPG	0.3375	7	Soyano et al. [33]	*L. japonicus*
CYCLOPS → NIN	0.1372	31	Singh et al. [32]	*L. japonicus*
CYCLOPS → ERN1	0.0252	88	Cerri et al. [314]	*L. japonicus*
RR1 → NIN	0.0002	664	Liu et al. [43]	*M. truncatula*
RR1 → NSP2	0.0003	787	Ariel et al. [315]	*M. truncatula*
NSP1 → D27	0.0672	42	Liu et al. [148]	*M. truncatula*
NSP1 → NIN	0.0792	47	Hirsch et al. [316]	*M. truncatula*
NSP1 → ERN1	0.0279	80	Hirsch et al. [316]	*M. truncatula*
NSP2 → D27	0.1475	28	Liu et al. [148]	*M. truncatula*
NSP2 → NIN	0.0158	108	Hirsch et al. [316]	*M. truncatula*
NSP2 → ERN1	0.1632	17	Hirsch et al. [316]	*M. truncatula*
IPN2 → NIN	0.0001	782	Kang et al. [317]	*L. japonicus*

[1] Transcription factor on the left, target on the right

[2] Rank of the known transcription factor in the list of all predicted transcription factors for the target

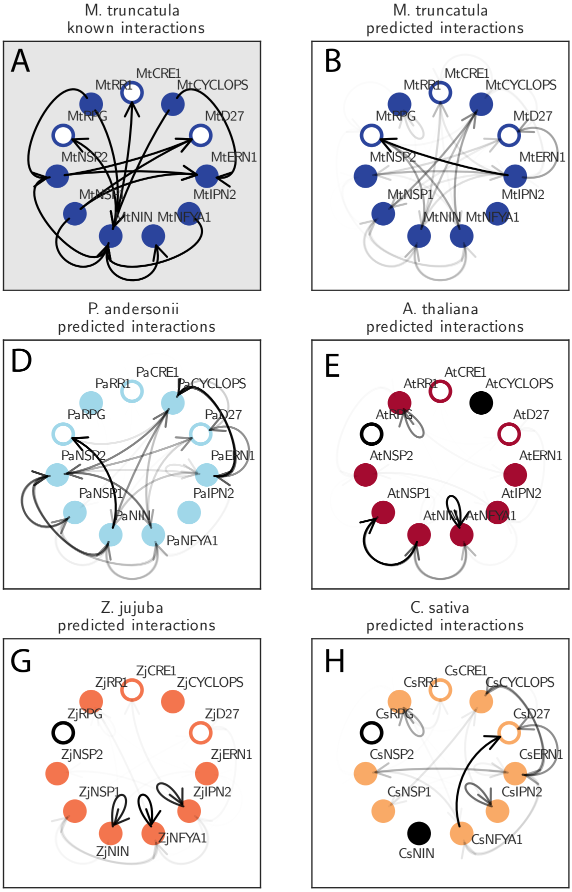


Figure 4.3: Known and predicted interactions between symbiotic transcription factors and their targets for the three symbiotic plant species in this study. (**A**) Known interactions in *Medicago truncatula*, (**B-H**) predicted interactions in *Medicago truncatula*, *Glycine max*, *Parasponia andersonii*, *Arabidopsis thaliana*, *Populus trichocarpa*, *Ziziphus jujube*, and *Cucumis sativa*. Symbiotic species are blue, non-symbiotic species are red. Filled circles represent transcription factors, open circles represent other genes. Black circles indicate absence of orthologs of the gene. Edge opacity represents the weight of the transcriptional interaction as reported by GRNBoost2: darker edges have a higher weight and should have a higher probability of representing true interactions.

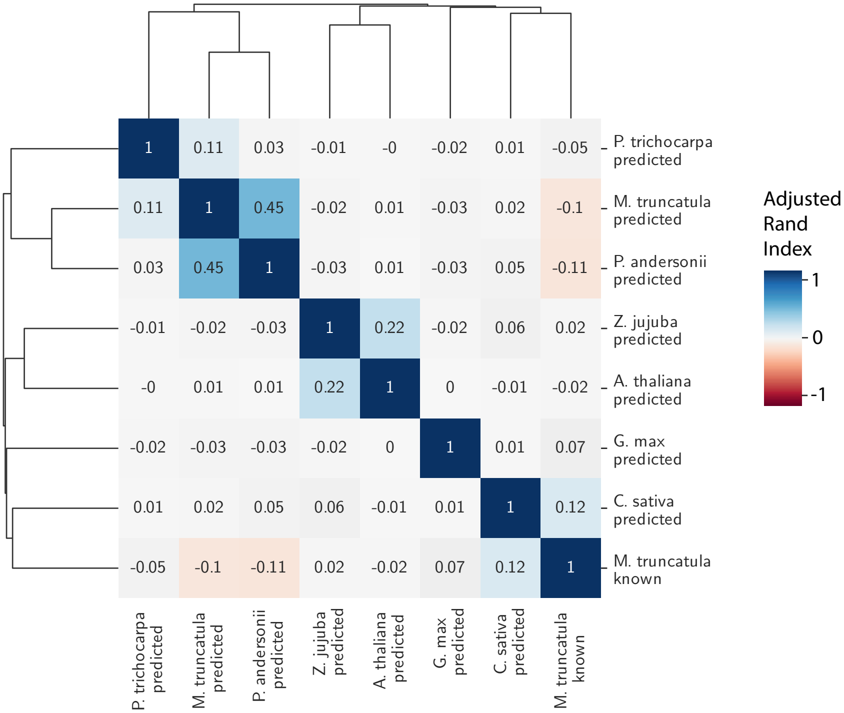


Figure 4.4: All-vs-all Adjusted Rand Index (ARI) for the known transcriptional interactions in *Medicago truncatula* and the predicted transcriptional interactions in *M. truncatula*, *Parasponia andersonii*, *Glycine max*, *Populus trichocarpa*, *Arabidopsis thaliana*, *Ziziphus jujuba*, and *Cucumis sativa*. The Adjusted Rand Index quantifies similarity between networks, where a value of one indicates perfect correspondence, and a value of zero indicates no similarity.

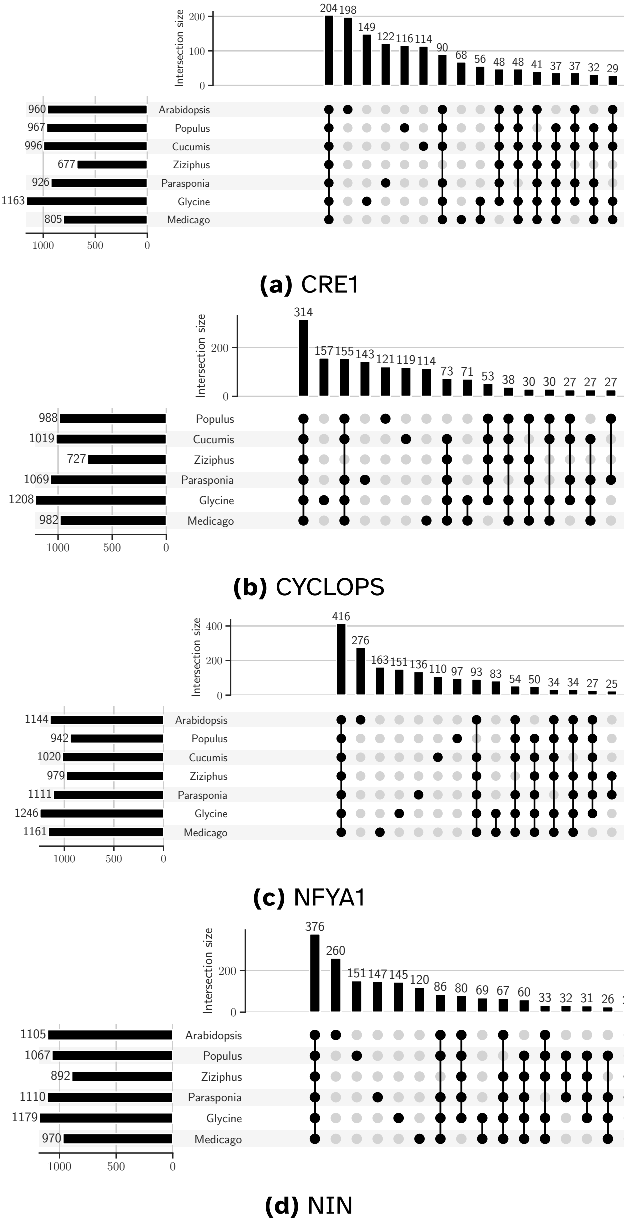
some spurious resemblance.

Moreover, given the large number of predicted transcription factors per gene (figure 4.2A), the interactions between the known nodulation genes only represent a limited number of the predicted transcription factors for each of these genes. To further investigate this, we examined the overlap between predicted transcription factors of *CYCLOPS*, *NODULE INCEPTION* (*NIN*), *NUCLEAR FACTOR GAMMA A1* (*NFYA1*) and *CYTOKININ RECPETOR1* (*CRE1*) in all species (figure 4.5). Given that these genes are known to be essential for nodulation in legumes, and that they work in concert (figure 4.3A), we expect a large degree of similarity between predicted regulators of these genes. Again, to our surprise most of the interactions are either predicted to be the same for all species, or to be uniquely present or absent for a single species (figure4.5).

To investigate whether this pattern of predicted species-specific regulatory programs exists at a genome wide scale, we turned to dimensionality reduction techniques to analyse similarity between genes in terms of their predicted regulators (figure 4.6). In brief, we used the non-linear dimensionality technique UMAP [318] on the binary interaction matrix to embed genes in a 2D space. As such, the embedding of a gene is based on its predicted regulators: if two genes are close in 2D space, their predicted regulators are similar. This allows for investigating whether the pattern of species-specific transcriptional interactions observed for a few candidate genes exists genome-wide. In the dimensionality reduction plot virtually all genes group together by species, indicating that the predicted regulators for genes from the same orthogroup vary greatly between species. Thus, the pattern we observed for a selected subset of genes translates to a genome wide pattern of species-specific predicted interactions. Following these results would indicate a strong discrepancy to our initial hypothesis of conserved transcriptional interactions. However, this interpretation relies heavily on the accuracy of our predicted transcriptional networks. Whether there is any biological relevance in our predictions cannot easily be determined from visual inspection of network plots for a few genes and our dimensionality reduction approach. This presents us with the fundamental problem that computational validation of predicted gene transcriptional interactions is impossible when there are no known interactions, as is the case for most of the plant species studied here.

4.2.3 Validation of predicted transcriptional networks in *A. thaliana*

Given the discrepancy between the known and predicted interactions in *M. truncatula*, and the high level of predicted species-specific interactions, we questioned the overall predictive accuracy of GRNBoost2. As the nine known interactions from legumes are insufficient to assess genome wide predictive performance, we therefore performed computational validation of the predicted transcriptional networks in *A. thaliana*. For these benchmarks we used two reference datasets from PlantRegMap [313]. Firstly, we used a collection of 1,428 interactions gathered using a text mining approach with manual curation on peer-reviewed literature. The advantage of this dataset is that the interactions have a high probability of being true, but that it is unlikely exhaustive for all targets of the transcription factors studied. Secondly, we used a collection of 26,274 interactions gathered using ChIP-seq experiments. The advantage of this dataset is that it is more exhaustive than the literature dataset, however ChIP-seq has been



(a) CRE1

(b) CYCLOPS

(c) NFYA1

(d) NIN

Figure 4.5: Overlap in the seven studied plant species between predicted transcription factors for the symbiotic genes (a) *CYCLOPS*, (b) *NIN*, (c) *NFYA1* and (d) *CRE1* visualized as upset plots. Set intersections are shown as variations of connected dots and are accompanied by bar graph indicating set size. For all four genes the biggest set intersection is for interactions that are found in all species, followed by species-specific interactions. Note that *NIN* has no entry for *Cucumis* and *CYCLOPS* has no entry for *Arabidopsis* as these genes are lost in these species respectively.

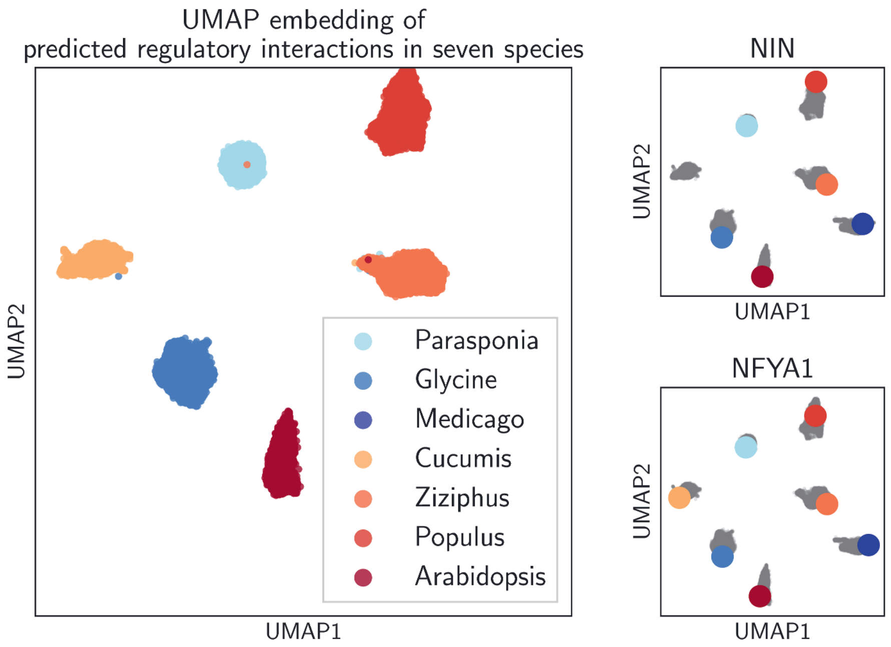


Figure 4.6: UMAP embedding of all predicted transcriptional interactions in seven species. Colours represent species. Colored clouds are the result of closely located dots. Left panel: dots represent all genes from one orthogroup in one species. *NIN* (top right) and *NFYA1* (bottom right) locations are highlighted and clearly follow the genome-wide species-specific pattern. Note that *NIN* has no dot for *Cucumis*, as *Cucumis* has lost the *NIN* gene.

reported as a noisy technique for which the biological relevance of many of the interactions are unclear. By comparing the *Arabidopsis* predicted transcriptional network to these two references we can make an estimate for the accuracy of the predictions. We found a poor predictive performance for the *Arabidopsis* transcriptional network predicted from all 2,295 samples using default parameters. When considering all interactions that get a positive weight from GRNBoost2, we recover 83% and 69% of the interactions listed in the literature and ChIP-seq references respectively. Unfortunately, the precision of the prediction is 1% and 10% for the literature and ChIP-seq respectively, translating to a false positive rate of between 90% and 99%. Furthermore, we find that ranking the predicted interactions based on the weights from GRNBoost2 yields only minor improvement when compared to the literature reference (figure 4.7A), and no improvement when compared to the CHiP-seq reference (figure 4.7B). On top of this, we find that GRNBoost2 performs only slightly better than random assignment of interactions for the literature reference, and worse than random assignment for the ChIP-seq reference (figure 4.7C).

In an attempt to improve the poor performance of GRNBoost2 on these large plant datasets, we explored a range of tool parameters, sample sizes, and filtering steps to identify their effect on the accuracy of the predictions (figure 4.8). Unfortunately, we found that tool parameters have little effect on predictive performance. If anything, the default settings are at the upper accuracy bound of the combinations explored, indicating there is little to be gained by changing parameters (figure 4.8B&C). Next, we find that sample size has no effect on recovering CHiP-seq interactions, and that up to 200 samples the accuracy of retrieving the literature interactions improves to the point where it is equal to the accuracy of using all 2,295 samples (figure 4.8D). Ultimately, of all attempts to increase predictive performance, one of the two filtering strategies shows the most profound effect (figure 4.8E). When only keeping predicted interactions where a relevant transcription factor binding site is present in the promoter region of the target gene, ChIP-seq recovery performs slightly better, but literature recovery performs worse. Strikingly, where we expected to find a higher probability of identifying a correct interaction when looking at interactions predicted in multiple species, we found the opposite. When only looking at predicted Arabidopsis interactions that are also predicted in Populus, we find a poorer performance for both the ChIP-seq and the literature comparison.

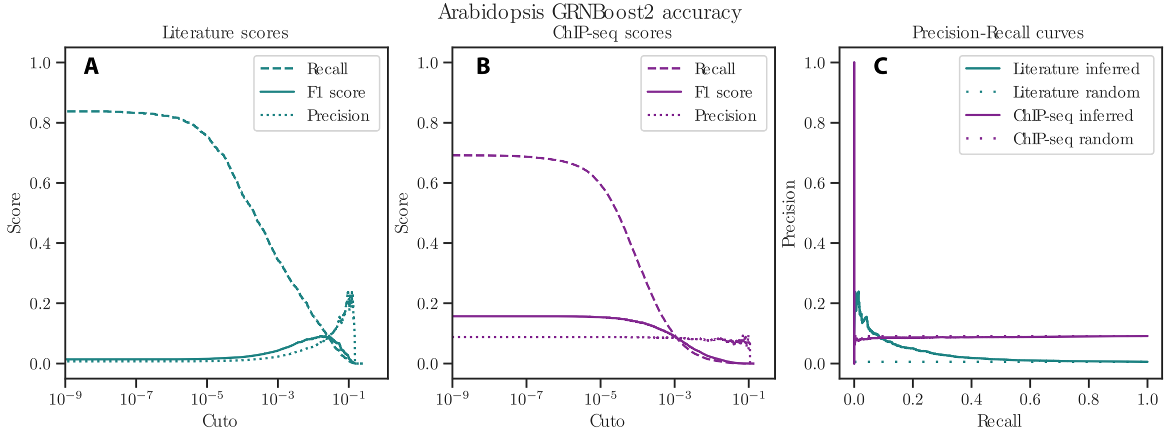


Figure 4.7: Predictive performance of the *Arabidopsis thaliana* transcriptional network inferred from all 2,295 samples. **A&B** Recall, Precision and F1-score plotted against a range of cut-offs for the predicted interaction weights using the literature dataset and the ChIP-seq dataset respectively. The F1-score represents the trade-off between recall and precision and is calculated as their geometric mean. Note that whereas a more stringent cut-off slightly increases precision in the literature dataset, this is not the case in the ChIP-seq dataset. **C** Precision recall curves using the same cut-offs as in A and B on the literature and ChIP-seq datasets, along with the theoretical precision and recall for random assignments. Curves tending to the upper right corner represent predictions that are better at dealing with the trade of between recall and precision. Calculating the area under the precision recall curve yields a single value representing the overall accuracy of the prediction. Literature AUPRC = 0.031, ChIP-seq AUPRC = 0.088.

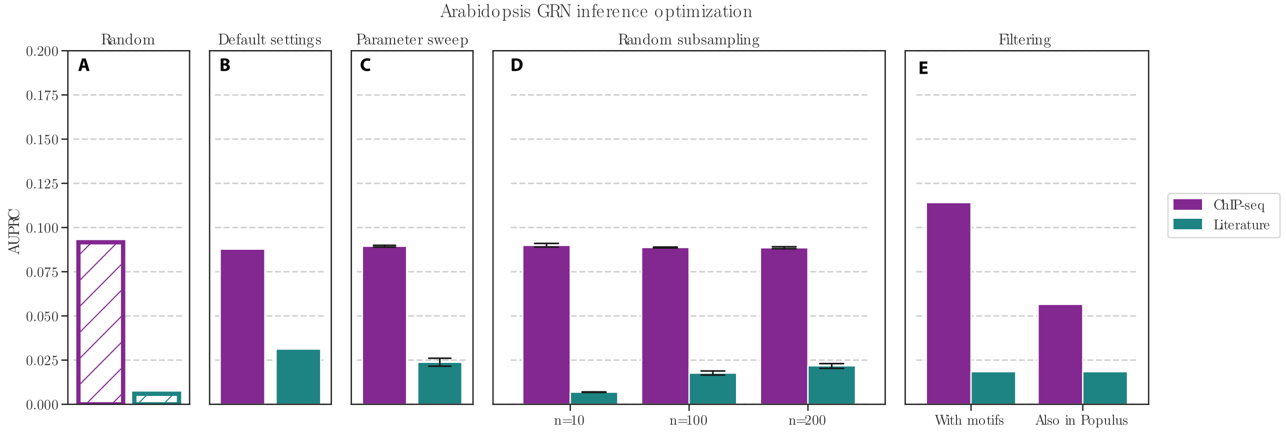


Figure 4.8: Area under the precision recall curve (AUPRC) in *Arabidopsis thaliana* when compared to literature and ChIP-seq for **A** random assignments, **B** default settings using all 2,295 RNA-seq samples, **C** 216 different parameter combinations on a random subsample of 200 samples, **D** 10 random samples of 10, 100 and 200 RNA-seq samples, **E** filtering of interactions in B based on transcription factor binding motifs or whether the same interaction was also found in *Populus*.

4.3 Discussion

Here, we attempted to study the role of transcriptional regulation in the evolution of rhizobium symbiosis at a genome wide scale. We have predicted transcriptional networks from public RNA-seq datasets for three symbiotic and four non-symbiotic plant species and devised a strategy to compare these predicted interactions between species using orthogroups. Although we have shown that it is conceptually and computationally feasible to do comparative analysis of transcriptional networks, we found that the low accuracy of current methods on plant datasets is prohibiting biological interpretation. Indicative of this limitation is the poor reconstruction of known symbiotic interactions in *Medicago truncatula* (figure 4.3), and the fact that none of the reconstructed interactions between symbiotic genes in various species resemble each other (figure 4.4). Previously, inferring transcriptional networks from RNA-seq data has been shown to be feasible and reasonably accurate in bacteria and yeast [302]. Nevertheless, benchmarking/validating transcriptional networks is difficult because there are often no gold standards, especially in plants. In absence of a gold standard, the main problem lies in defining false positives: classifying a predicted regulatory link that is not in the reference set as false positive can be incorrect because it might be the one thing the analysis is supposed to find: a previously unknown interaction. This problem increases when the scale of the predicted transcriptional networks is large and the set of known interactions is small, as is the case in plants. Given the problematic nature of validating transcriptional networks, we have opted for validating our predictions against two *Arabidopsis thaliana* reference datasets [313]: a “strict” reference dataset based on published interactions only and a “loose” reference dataset based on ChIP-seq. This strategy is in line with recent suggestions that ultimately, validation of transcriptional networks should be based on knowledge and measurement data of in-vivo interactions [319]. Using these two datasets as reference likely results in overestimating and underestimating the number of false positives respectively. Given that the true accuracy of our predictions then lies in between the accuracies for the two used datasets, we can estimate the accuracy of the predicted transcriptional network in *Arabidopsis thaliana* with reasonable precision. To our surprise we found a very poor predictive performance across a wide range of parameter settings, sample selections and filtering approaches. More importantly, we found that the poor accuracy of predicted transcriptional networks has severe implications for comparative analyses. Whereas to our knowledge there is currently no data on conserved regulatory interactions in plants, we had expected to find at least some conserved interactions, and some phylogenetic signal in the predictions in general (i.e. closely related species share conserved regulatory interactions). Instead we found that the UMAP projection groups interactions in a completely species-specific manner. Together with the poor accuracy of the predicted transcriptional networks in *A. thaliana*, we conclude that comparative analysis of predicted transcriptional networks is problematic. Whereas we cannot exclude that the species-specific patterns are due to different samples available for different species, we found that random subsampling in *A. thaliana* down to the minimum number of samples available for any of the seven species (111 samples for *Parasponia andersonii*) did not result in a significant difference in accuracy. Currently, transcriptional networks predicted from public transcriptome data cannot be used for comparative analyses in plants. If predicting a transcriptional network from RNA-seq in plants would be more accurate than randomly assigning interactions, it could still be a useful tool for designing experiments. For example, if an interaction between two genes is predicted to be present in two species, a relatively straightforward knock-out experiment of the transcription factor could corroborate the prediction. For this use-case, it is most important that the predicted interactions have a high probability of being true, i.e. that they have a high precision. However, we find an accuracy that is at best slightly better than random assignment, and a precision of 1% - 10%. Therefore, we question the use of the currently used methods for transcriptional network inference in plant science in general. An important question that remains unanswered is why the predictive performance of transcriptional networks inferred from bulk RNA-seq in plants is so much worse than that in bacteria and yeast. Similar to our work, a recent study found that even when inducing specific *A. thaliana* transcription factors and predicting interactions based on differential gene expression of potential targets, the predictive performance is barely better than randomly assigning interactions [320]. There are several likely explanations for this lack in predictive performance, mostly related to data resolution. Bacteria and yeast are unicellular organisms, so when grown under controlled conditions they will likely be very homogeneous in their expression profile. Bulk RNA-seq for multicellular organisms, like plants, unavoidably yields a mix of cells of various cell types from various tissues and therefore a complex mix of expression profiles. It is therefore perhaps not surprising that methods that inherently rely on correlation between gene expression levels do not accurately represent mechanisms of transcriptional regulation when used on these heavily mixed datasets. Herein lies one possible direction of improvement to get the predictive performance of transcriptional networks in plants at the same level it is in bacteria and yeast: these methods should ideally be used on datasets with a fine grained resolution, approaching the level of homogeneous tissues and ultimately single cells [321]. Besides the lack of resolution in the RNA-seq datasets used in this study, another explanation for the poor predictive power of the transcriptional networks in this study is that they are based on correlations between expression levels only. Using gene expression data alone has been suggested to be insufficient for measuring indirect effects [322]. Additionally, gene expression in eukaryotes can be modulated by several factors that are not measured with classical RNA-seq. Known modulators of gene expression include micro-RNAs [323] and post translational modification of transcriptional regulators [324]. A good example of the latter mechanism is *CYCLOPS*, one of the symbiotic transcription factors highlighted in this study. To become an active transcriptional regulator, *CYCLOPS* requires phosphorylation by *CALCIUM AND CALMODULIN DEPENDENT KINASE* (*CCamK*), suggesting that measuring the expression level of *CYCLOPS* alone is not sufficient to explain its activity. Ultimately, if hypotheses on transcriptional rewiring during the evolution of rhizobium symbiosis are to be tested, it might prove invaluable to determine conservation of some of the known interactions from legumes in several symbiotic and non-symbiotic plants using more conventional methods such as classic electrophoretic mobility-shift assays (EMSA) or ChIP-seq. Alternatively, it will be essential to integrate different types of data to get a full overview of the landscape of transcriptional regulation in plants. Promising recent advances in high throughput methods include DNAse-seq [325] to identify regions of the genome susceptible to protein binding or DAP-seq [326] to identify targets for a library of hundreds to thousands of transcription factors in vitro. In summary, without high-throughput and high-resolution techniques, comparative analysis of transcriptional networks will not be feasible.

4.4 Conclusions

This work highlights some of the major limitations that currently hamper comparative analysis of transcriptional networks involved in rhizobium symbiosis. Whereas the conceptual framework presented in this study provides the tools to study the evolution of transcriptional regulation, it must be concluded that there is currently insufficient data to perform a meaningful analysis in nodulating plant species. Most importantly, the false positive rate of 90% - 99% for predicted regulatory interactions renders a search for new transcription factors involved in rhizobium symbiosis using this methodology a moot point. Future efforts will have to focus on the use and integration of various molecular techniques to potentially provide the resolution required to come a little closer to the mechanisms of transcriptional regulation in multicellular organisms [327]. With the recent application of high throughput single cell RNA sequencing techniques in *Arabidopsis thaliana* [328], it is now timely to take the concepts presented in this study and apply them to such fine-grained datasets in nodulating plant species.

4.5 Materials and methods

4.5.1 Species and sample selection

Seven model plant species with at least 100 publicly available transcriptome samples were selected, three of which can form nitrogen-fixing symbiosis with rhizobium bacteria (*Medicago truncatula*, *Glycine max* and *Parasponia andersonii*) and four that can not form the symbiosis (*Arabidopsis thaliana*, *Populus trichocarpa*, *Cucumis sativus* and *Ziziphus jujuba*). In addition to transcriptome sample availability we selected these species because they span the phylogenetic range over which rhizobium symbiosis likely originated and subsequently was lost [213, 295]. We downloaded the following genome versions with their annotations from phytozome v12 [289]: *A. thaliana* TAIR10, *C. sativus* v1, *G. max* v2, *M. truncatula* v4, and *P. trichocarpa* v3. In addition, we downloaded the *P. andersonii* v1 genome and annotation from `www.parasponia.org` and the *Z. jujuba* genome and annotation from NCBI genome. For all species we only selected primary transcripts from the annotation, either by downloading the corresponding file from phytozome or by selecting the longest isoform as primary transcript. Publicly available transcriptome samples for these seven plant species were selected with the help of the online NCBI run selector. We have only included mRNA samples that were sequenced on Illumina HiSeq instruments. Supplementary table 1 contains a full list of samples used in this study, including exact download dates, metadata on the biological material, and references to original publications.

4.5.2 RNA-seq processing and quantification

Raw reads for all transcriptome samples were downloaded from the NCBI Sequence Read Archive. RNA quantification was done by pseudo-aligning the reads to the coding sequences of the relevant species with kallisto [329] using default parameters and 100 bootstrap replicates for the TPM normalization. We developed a snakemake [330] pipeline to efficiently process large numbers of samples in parallel and remove all intermediate files per sample as soon as the quantification step was finished.

4.5.3 Orthogroup inference

Predicted protein sequences of the seven plant species were used to compute orthogroups using OrthoFinder v2.2.6 [285]. We used diamond v0.9.24.125 [331] as sequence search tool and used MCL v14-137 [332] inflation parameter 2.5.

4.5.4 Selecting candidate transcription factors

For all species except *P. andersonii* we downloaded a list of candidate transcription factors from PlantTFDB 4.0 [333]. *P. andersonii* transcription factors were predicted with the PlantTFDB online prediction tool [333]. To ensure consistency between the orthogroups and the downloaded transcription factor lists, we annotated all genes from an orthogroup as transcription factor if one of the genes in the orthogroup was annotated as transcription factor by PlantTFDB.

4.5.5 Transcriptional network inference

Transcriptional networks were predicted using GRNBoost2 version 0.1.5 [309], which uses nonlinear multivariate regression in the form of a stochastic gradient boosting machine. For the comparative analysis default parameters were used. For the performance analysis we tested all 72 combinations of the following five parameters: loss function (Least Squares or Least Absolute Deviation), learning rate (0.01 or 0.001), max features (0.1, 0.5 or auto), subsample (0.7, 0.9 or 0.99), and early stop window length (25 or 100). To identify whether filtering for interactions that are consistently predicted in multiple species improved performance we only kept predicted interactions that were found in both *A. thaliana* and *P. trichocarpa*. To test the effect of filtering for predicted interactions that are supported by predicted transcription factor binding sites we downloaded binding site predictions from PlantTFDB [333].

4.5.6 Transcriptional network dimensionality reduction

To visualize the densely connected transcriptional networks we used UMAP for dimensionality reduction [318]. Briefly, a weighted transcriptional network can be seen as an $m \times n$ matrix A, with m genes and $n$transcription factors, where $n$is a subset of m, and A_{mn} is the weight assigned to the predicted interaction between transcription factor n and gene m. By using the nonlinear dimensionality reduction approach of UMAP we turn the $m \times n$ matrix into an $m \times 2$ matrix that can be visualized in 2D. We turned our transcriptional network into a boolean matrix by assigning a value of 1 to all non-zero interactions, and subsequently used UMAP with the jaccard distance metric, and we set n_neighbors to 20 and min_dist to 0.5.

4.5.7 Transcriptional interaction benchmark datasets

For benchmarking predicted transcriptional interactions in *Arabidopsis thaliana* we used a collection of interactions that was previously compiled in the PlantRegMap [313], available at `http://plantregmap.cbi.pku.edu.cn`. We used PlantRegMap interactions that are either based on literature text mining with manual curation (the 'regulations in ATRM' file available at `http://atrm.cbi.pku.edu.cn/download.php`), or from a diverse range of ChIP-seq experiments (the Arabididopsis thaliana regulation merged file available at `http://plantregmap.cbi.pku.edu.cn/download.php`).

4.5.8 Accuracy metrics for transcriptional network validation

All accuracy measures used are based on different combinations of true negatives (TN), true positives (TP), false negatives (FN) and false positives (FP). We used scikit-learn v0.20.1 (Pedregosa et al. 2011) to calculate all accuracy measures. To compare networks of known nodulation genes we calculated the Adjusted Rand Index. Briefly, the Adjusted Rand Index is a corrected-for-chance version of the Rand Index, where the Rand Index is defined as follows:

$$RI = \frac{TP + TN}{TP + FP + FN + TN}$$

To measure accuracy in the *A. thaliana* predicted transcriptional network we calculated the precision, recall and F_1-score, defined as follows:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

The Area Under the Precision Recall Curve (AUPRC) was used to define overall predictive performance by varying the weight threshold on the predicted transcriptional networks and calculating precision and recall at every cut-off. The AUPRC can then be calculated by graphing precision versus recall and calculating the area under the curve.

Chapter 5

Genenotebook: a collaborative notebook for comparative genomics

Rens Holmer, Robin van Velzen, René Geurts, Ton Bisseling, Dick de Ridder, and Sandra Smit

This chapter was published as

Holmer R et al. “GeneNoteBook, a collaborative notebook for comparative genomics”. *Bioinformatics* 35 (2019), pp. 4779–4781

Abstract

Summary: Analysis and comparison of genomic and transcriptomic data sets have become standard procedures in biological research. However, for non-model organisms no efficient tools exist to visually work with multiple genomes and their metadata, and to annotate such data in a collaborative way. Here we present GeneNoteBook: a web based collaborative notebook for comparative genomics. GeneNoteBook allows experimental and computational researchers to query, browse, visualize and curate bioinformatic analysis results for multiple genomes. GeneNoteBook is particularly suitable for the analysis of non-model organisms, as it allows for comparing newly sequenced genomes to those of model organisms.
Availability and implementation: GeneNoteBook is implemented as a node.js web application and depends on MongoDB and NCBI BLAST. Source code is available at `https://github.com/genenotebook/genenotebook`. Additionally, GeneNoteBook can be installed through Bioconda and as a Docker image.
Supplementary information: Full installation instructions and online documentation are available at `https://genenotebook.github.io`. Supplementary text is available at *Bioinformatics* online.

5.1 Introduction

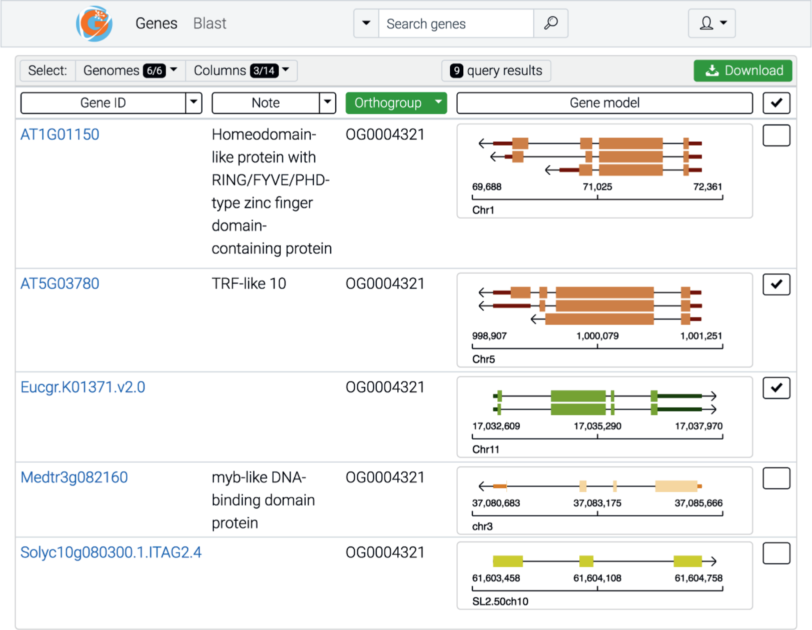
Browsing, querying, and comparing large genomic and transcriptomic datasets are indispensable aspects of genomic research. In recent years the decrease in cost of sequencing DNA or RNA has unlocked the possibility to generate eukaryotic genome assemblies with limited effort. As a result, genome analysis has become a routine exercise for research groups working on non-model organisms. Annotated genome sequences with metadata are used to identify candidate genes that can be targeted in wet lab experiments. As an example, integrating information on ortholog groups, protein domains and gene expression levels can provide valuable information on a gene s hypothetical function. For such integration, it is crucial to be able to browse, query and compare genomic data, and curate automated predictions. This should ideally be a collaborative effort between experimental and computational researchers, and should be an efficient process that requires minimal configuration.

Currently, no efficient tool exists to quickly query, browse and visualize genomic data. Whereas genome browsers such as JBrowse [335, 336] provide powerful visualizations, they are limited to positional queries and visualizations. As an extension to JBrowse, Apollo allows for the curation of gene structure models [337]. However, both JBrowse and Apollo are limited to single genomes. Additionally, genome browsers are not very suitable for the integration of various data types, such as gene expression levels and ortholog groups. Data warehouse systems, such as InterMine [338], provide more powerful query options but are relatively difficult to configure and generally do not come with data visualization options. Previously, data warehouse systems like InterMine and genome browsers like JBrowse have been combined into custom one-off data portals for model organisms, such as Araport for *Arabidopsis thaliana* [339], the Legume Information System for legumes [340] or Wormbase for *Caenorhabditis elegans* and related nematodes [341]. However, setting up a custom data portal for each new genome is inefficient and time consuming. Additionally, it is currently not possible to collaboratively curate genomic metadata, for instance by adding curator notes to genes.

To enable quick and intuitive browsing and querying of genomic data for newly sequenced organisms we have developed GeneNoteBook: a collaborative web-based notebook for comparative genomics. Our application is designed for comparative analysis of genomic data and collaborative annotation of predicted genes with expert knowledge, by integrating genome annotations, gene expression data, and gene evolutionary relationships.

5.2 Features

GeneNoteBook provides users with two views on their genomic data: a spreadsheet-like gene table with customizable fields and queries to browse and visualize information for multiple genes from multiple genomes, and a gene page with all available information for any particular gene. Additionally, gene sequences can be searched with BLAST, and an administrator section provides configuration options. For demonstration purposes, we use publicly available genomic and transcriptomic data of the model plant species *Medicago truncatula* [18] and *Arabidopsis thaliana* [342]. The use cases are based on workflows from our previous work on *Parasponia andersonii* in chapter 3, a new model for studying the symbiotic relationship between rhizobium bacteria and plants [343].



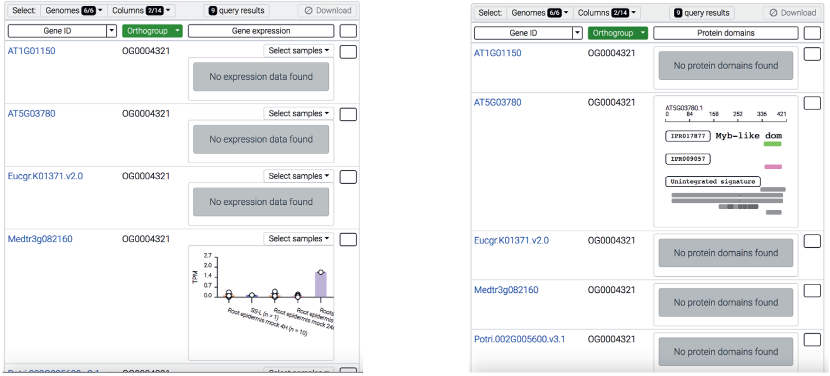
Gene ID	Note	Orthogroup	Gene model	☑
AT1G01150	Homeodomain-like protein with RING/FYVE/PHD-type zinc finger domain-containing protein	OG0004321	69,688 71,025 72,361 Chr1	□
AT5G03780	TRF-like 10	OG0004321	998,907 1,000,079 1,001,251 Chr5	☑
Eucgr.K01371.v2.0		OG0004321	17,032,609 17,035,290 17,037,970 Chr11	☑
Medtr3g082160	myb-like DNA-binding domain protein	OG0004321	37,080,683 37,083,175 37,085,666 chr3	□
Solyc10g080300.1.ITAG2.4		OG0004321	61,603,458 61,604,108 61,604,758 SL2.50ch10	□

Figure 5.1: GeneNoteBook gene table view. This view shows genes in ortholog group OG0004321. Gene model colors indicate in which species the gene is found. Several genes have notes that indicate these genes are transcription factors. UTR regions vary, but most gene models have similar exon structures. Three genes have been selected for downloading.

5.2.1 Searchable gene table with visualizations

Genome annotations loaded into a GeneNoteBook instance can be browsed through a dynamically rendered HTML table that combines queries with SVG data visualizations. Users can select genome annotation tracks for one or more genomes, and select gene attributes to be displayed as table columns. The table can be sorted and filtered on the selected attribute columns. To achieve this, GeneNoteBook automatically keeps track of all available query options: if a user adds a new attribute to a gene in the gene page, this attribute automatically becomes a query option in the gene table. Additionally, all query results can be downloaded as text files formatted according to common specifications such as FASTA, GFF3 or TSV.

Furthermore, one of the following three data visualizations can be selected to be included in the table: (1) a graphical representation of the gene model (figures 5.1 and 5.4), (2) protein domains as predicted by InterProScan [282] (figure 5.5), or (3) gene expression levels (figure 5.7). To facilitate intuitive browsing and discovery of genes of interest, the gene table always starts with a view of all available data that can subsequently be narrowed down by constructing a query. This ensures that users need only limited prior knowledge to start browsing.



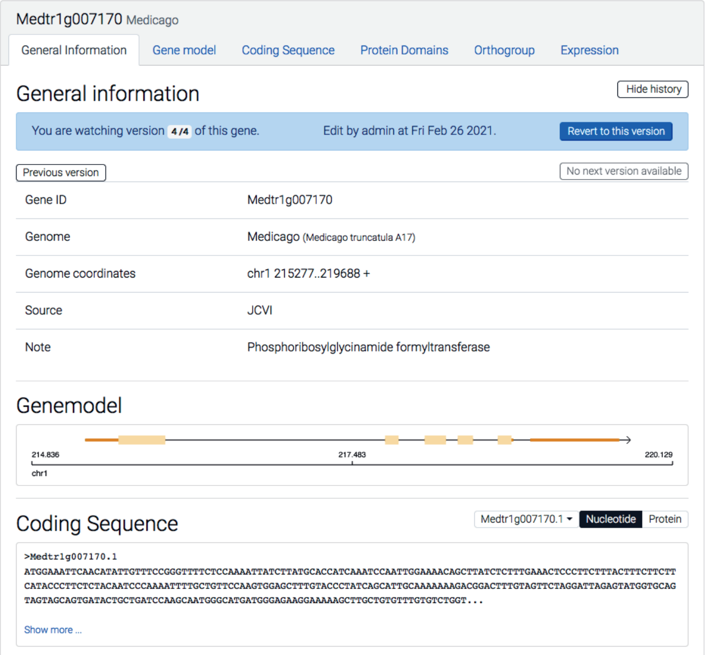
Gene ID	Orthogroup	Gene expression	□
AT1G01150	OG0004321	Select samples No expression data found	□
AT5G03780	OG0004321	Select samples No expression data found	□
Eucgr.K01371.v2.0	OG0004321	Select samples No expression data found	□
Medtr3g082160	OG0004321	Select samples	□

Gene ID	Orthogroup	Protein domains	□
AT1G01150	OG0004321	No protein domains found	□
AT5G03780	OG0004321		□
Eucgr.K01371.v2.0	OG0004321	No protein domains found	□
Medtr3g082160	OG0004321	No protein domains found	□
Potri.002G005600.v3.1	OG0004321	No protein domains found	□

Figure 5.2: GeneNoteBook's gene table with SVG visualizations. Gene tables visualizing expression levels (left panel) and InterProScan predicted protein domains (right panel) for the same ortholog group (OG0004321) as in figure 5.1. By using the gene table, it is trivial to see what types of information are available for all hits that result from a query for a variety of organisms.

5.2.2 Single gene page with version history

For every gene in a GeneNoteBook instance, a comprehensive information page is rendered. The single gene page starts with a list of all gene attributes in the form of key-value pairs. Users with curator or administrator access (see section 5.5.2 for user account types) can modify existing attributes, or add new attributes. This allows for the curation of automatically assigned protein product names, or addition of trivial names or notes to genes of interest. To prevent simultaneous conflicting edits to the same gene, once a user starts editing the attributes of a gene, the ability to edit the gene is locked for all other users. Once the edited fields are saved, the gene is unlocked and all changes are saved to a version history. To make sure a user cannot lock a gene forever the gene is automatically unlocked after 10 minutes of inactivity. A version history that is maintained in the database assures that all manual curations and additions can be tracked and reverted if needed. In addition to the listed attributes, the single gene page contains a panel with the coding sequence of all transcripts of the gene and visualizations for the gene model, predicted protein domains, gene expression levels and a phylogenetic tree for the corresponding ortholog group.



General information

You are watching version 4/4 of this gene. Edit by admin at Fri Feb 26 2021.

Gene ID	Medtr1g007170
Genome	Medicago (Medicago truncatula A17)
Genome coordinates	chr1 215277..219688 +
Source	JCVI
Note	Phosphoribosylglycinamide formyltransferase

Genemodel

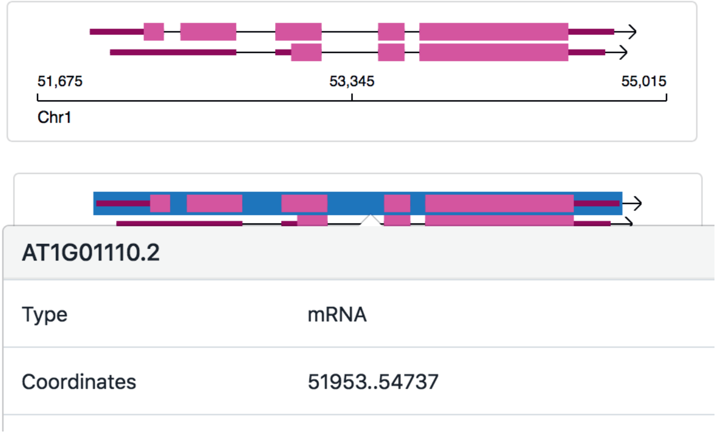
Coding Sequence

```
>Medtr1g007170.1
ATGGAAATTCAACATATTGTTTCCGGGTTTTCTCCAAAATTATCTTATGCACCATCAAATCCAATTGGAAAACAGCTTATCTCTTTGAAACTCCCTTCTTTACTTTCTTCTT
CATACCCTTCTCTACAATCCCAAAATTTTGCTGTTCCAAGTGGAGCTTTGTACCCTATCAGCATTGCAAAAAAAGACGGACTTTGTAGTTCTAGGATTAGAGTATGGTGCAG
TAGTAGCAGTGATACTGCTGATCCAAGCAATGGGCATGATGGGAGAAGGAAAAAGCTTGCTGTGTTTGTGTCTGGT...
```

Figure 5.3: Top part of the gene page highlighting the *Medicago truncatula* gene Medtr1g007170 with version history. The blue bar at the top highlights that this gene has been edited in the past and currently has two versions. Curator and administrator users can decide to revert the changes by clicking the 'Revert to this version' button.

5.2.3 Visualizations

GeneNoteBook dynamically renders SVG visualizations for various data types and analysis results. Since GeneNoteBooks minimally required input data is an annotated genome sequence, the gene structure model visualization is always available. Visualizations for predicted protein domains, ortholog group phylogenetic trees and gene expression levels become available once the appropriate data (see table 5.1) are loaded into GeneNoteBook. The gene model visualization provides a graphical representation of the transcripts and exons of a gene. Coding sequence exons are rendered larger than UTR exons to accentuate the difference. Exons or mRNAs can be clicked to display exon- or mRNA-specific information, such as coordinates and IDs, in a pop up window.



AT1G01110.2

Type	mRNA
Coordinates	51953..54737

Figure 5.4: Gene model visualization of the *Arabidopsis thaliana* gene AT1G01110 with two alternative splice forms. The general genomic location of this gene is on chr1, starting at nucleotide 51,675 and ending at nucleotide 55,015 (top panel). The exact coordinates of each exon or mRNA can be accessed through a popup window that opens when an exon or intron is clicked or hovered. In this case the full mRNA interval has been clicked, showing that it ranges from 51,953 to 54,737 (bottom panel). Black lines indicate introns. Narrow purple bars indicate UTR regions. Wider pink bars indicate coding sequence exons.

Protein domains as predicted by InterProScan [282] are displayed sorted by Interpro domain type, with their Interpro ID and short description. Protein domains can be clicked to display additional information in a pop up window.

Ortholog group membership is visualized as a phylogenetic tree. Other genes in the phylogenetic tree are automatically turned into hyperlinks if they are included in the GeneNoteBook instance, which greatly simplifies navigating between various genes within an ortholog group. By default the genes in the tree are labelled with their gene ID and, if available, their name. Genes are coloured based on the genome in which they are found.

GeneNoteBook can visualize gene expression quantified from RNA sequencing experiments as bar plots with error-bars representing the standard error. Additionally, individual data points are displayed as open circles. Transcriptome samples can be grouped in replica groups either upon loading the data, or in the Admin panel. Y-axis limits are automatically calculated from the data and a drop down menu allows users to select which experiments they want displayed.

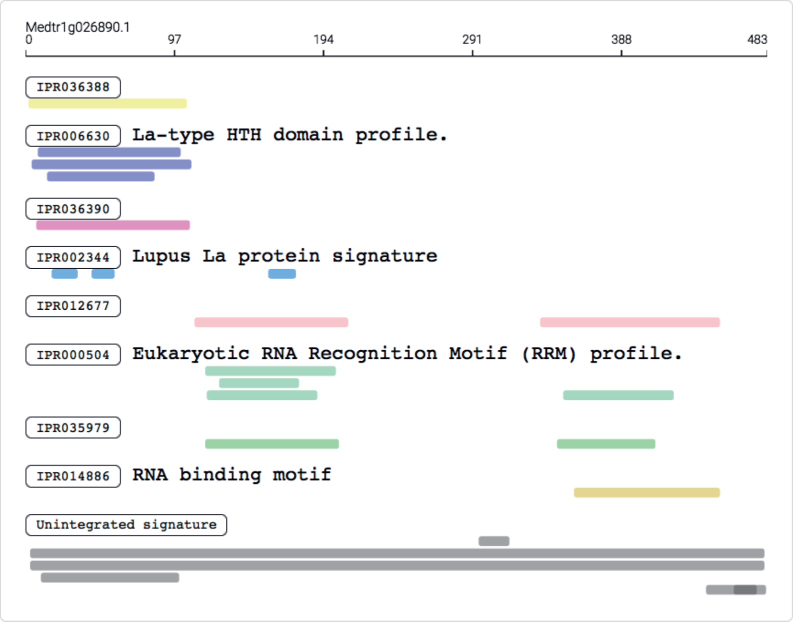


Figure 5.5: InterProScan protein domain prediction visualization for the first splice form of the *Medicago truncatula* gene Medtr1g026890. The axis at the top of the visualization represents the length of the protein in amino acids. Protein domains are sorted and colored by their Interpro annotation. More information on the source of the prediction and other information provided by InterProScan is available in a popup window that can be accessed by hovering or clicking a domain.

5.3 Sequence search using BLAST

To allow sequence-based searching of gene models, GeneNoteBook implements a wrapper around BLAST [281]. Users can submit fasta formatted DNA or protein sequences to BLAST against GeneNoteBooks genome annotations. BLAST jobs are handled by GeneNoteBooks job queue such that the server will not be flooded with blast jobs and results can be stored. Every BLAST job has a unique ID with a corresponding URL so that BLAST jobs can be retrieved at a later moment. Results are presented in two ways: (1) as an SVG visualization showing the blast hits relative to the query sequence and (2) as a list with hits sorted by E-value including more detailed information about the hits. Additionally, users can send the results to the gene table interface to make additional queries on the BLAST hits, or to download the corresponding data.

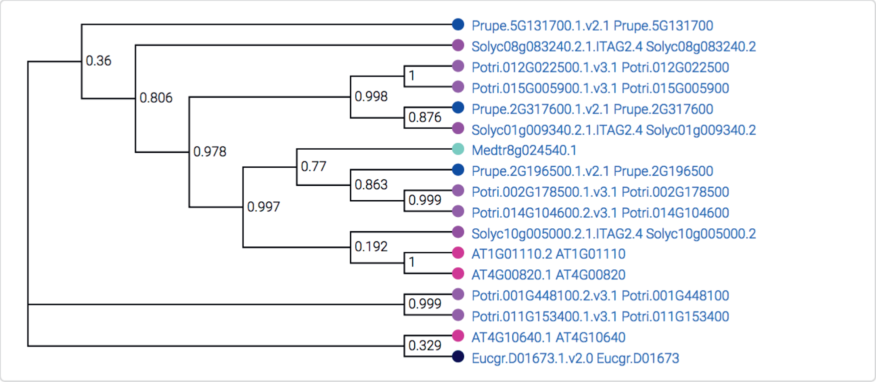

0.36
0.806
0.978
0.998
1
0.876
0.77
0.863
0.999
0.997
0.192
1
0.999
0.329
Prupe.5G131700.1.v2.1 Prupe.5G131700
Solyc08g083240.2.1.ITAG2.4 Solyc08g083240.2
Potri.012G022500.1.v3.1 Potri.012G022500
Potri.015G005900.1.v3.1 Potri.015G005900
Prupe.2G317600.1.v2.1 Prupe.2G317600
Solyc01g009340.2.1.ITAG2.4 Solyc01g009340.2
Medtr8g024540.1
Prupe.2G196500.1.v2.1 Prupe.2G196500
Potri.002G178500.1.v3.1 Potri.002G178500
Potri.014G104600.2.v3.1 Potri.014G104600
Solyc10g005000.2.1.ITAG2.4 Solyc10g005000.2
AT1G01110.2 AT1G01110
AT4G00820.1 AT4G00820
Potri.001G448100.2.v3.1 Potri.001G448100
Potri.011G153400.1.v3.1 Potri.011G153400
AT4G10640.1 AT4G10640
Eucgr.D01673.1.v2.0 Eucgr.D01673

Figure 5.6: Phylogenetic tree visualization for an ortholog group that contains genes from several plant species. GeneNoteBook presents genes that can be linked by their gene ID as hyperlinks to the single gene page of that gene. Additionally, internal nodes display confidence values that are parsed from the Newick formatted tree file used as input. Tip nodes are colored according to the corresponding organism: *Prunus persica* (blue), *Solanum lycopersicum* (ligh purple), *Populus trichocarpa* (dark purple), *Medicago truncatula* (teal), *Arabidopsis thaliana* (pink), *Eucalyptus grandis* (black). This ortholog group was identified by OrthoFinder [285], the phylogenetic tree was constructed using the neighbor joining algorithm [290] on a MAFFT [344] multiple sequence alignment.

5.4 Use cases

These use cases are based on workflows from our previous work on *Parasponia andersonii*, a tropical tree for which we newly sequenced and annotated its genome (see chapter 3). We have populated a GeneNoteBook server with the genomes and protein-coding gene annotations of *Medicago truncatula*, *Arabidopsis thaliana* and *Parasponia andersonii*. Additionally, we have added ortholog group information that includes these species, as well as protein domain predictions quantified gene expression levels for *P. andersonii*.

5.4.1 Visualizing gene expression of BLAST results

Whenever a new gene is discovered to be involved in rhizobium symbiosis in a species that we have not included in our ortholog groups (for example *Lotus japonicus*), we use the GeneNoteBook BLAST functionality to quickly determine the *P. andersonii* homologs and their expression levels in various tissues.

Step 1. Obtain the original nucleotide/protein sequence for the gene

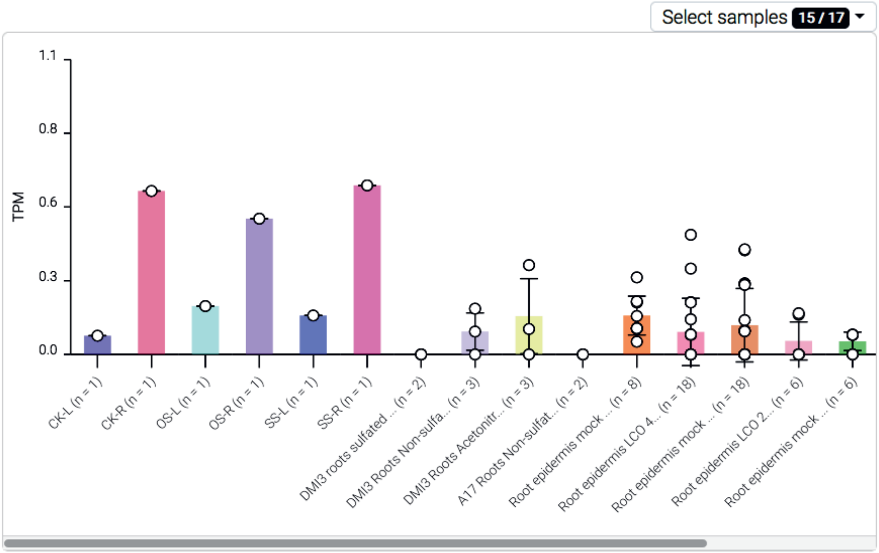


Figure 5.7: GeneNoteBook SVG barplot of expression levels. Quantified gene expression is shown for the *Medicago truncatula* gene Medtr3g082160 in several publicly available transcriptome datasets (BioProjects PRJNA283476 and PRJNA255840). Experiments with multiple replicates are grouped together into replica groups, bars are automatically colored according to replica group. Labels per bar indicate the name of the replica group and automatically add the number of data points. Individual data points are plotted over a bar that represents the mean expression level, with error bars indicating the standard error. Gene expression was quantified using kallisto [329]. TPM = transcripts per million.

of interest

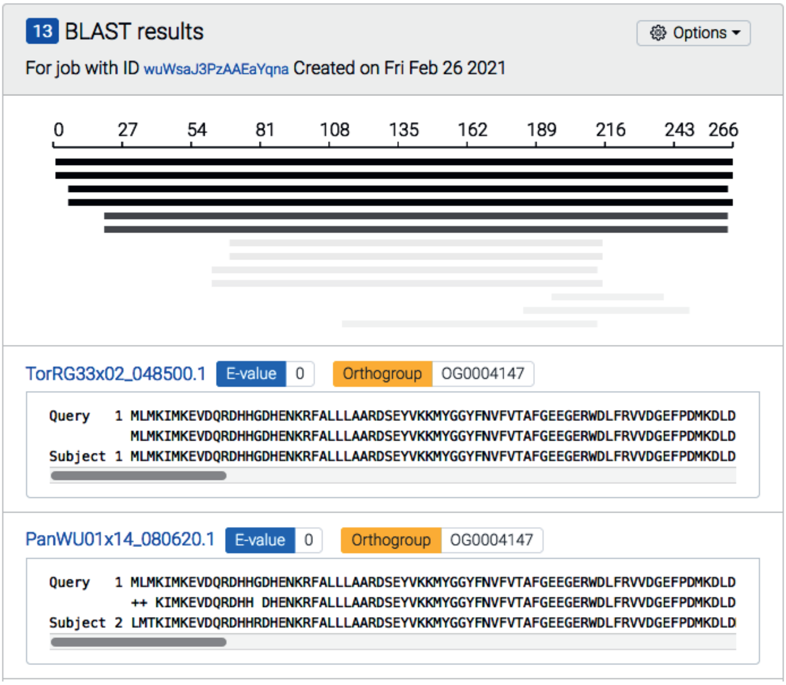
Step 2. BLAST the gene sequence against the *P. andersonii* genome annotation

Step 3. Select the Gene Expression visualization from the options menu in the BLAST results section

Step 4. Optionally, send the BLAST results to the Gene Table section for additional queries

5.4.2 Downloading protein sequences of all genes in an ortholog group

We predominantly use ortholog group phylogenetic trees constructed with the neighbor joining algorithm in our GeneNoteBook. Whenever we doubt



13 BLAST results

For job with ID wuWsaJ3PzAAEaYqna Created on Fri Feb 26 2021

TorRG33x02_048500.1 E-value 0 Orthogroup OG0004147

```
Query   1 MLMKIMKEVDQRDHHGDHENKRFALLLAARDSEYVKKMYGGYFNVFVTAFGEEGERWDLFRVVDGEFPDMKDLD
          MLMKIMKEVDQRDHHGDHENKRFALLLAARDSEYVKKMYGGYFNVFVTAFGEEGERWDLFRVVDGEFPDMKDLD
Subject 1 MLMKIMKEVDQRDHHGDHENKRFALLLAARDSEYVKKMYGGYFNVFVTAFGEEGERWDLFRVVDGEFPDMKDLD
```

PanWU01x14_080620.1 E-value 0 Orthogroup OG0004147

```
Query   1 MLMKIMKEVDQRDHHGDHENKRFALLLAARDSEYVKKMYGGYFNVFVTAFGEEGERWDLFRVVDGEFPDMKDLD
          ++ KIMKEVDQRDHH DHENKRFALLLAARDSEYVKKMYGGYFNVFVTAFGEEGERWDLFRVVDGEFPDMKDLD
Subject 2 LMTKIMKEVDQRDHHRDHENKRFALLLAARDSEYVKKMYGGYFNVFVTAFGEEGERWDLFRVVDGEFPDMKDLD
```

Figure 5.8: BLAST result visualization. The figure depicts the location of the BLAST hits relative to the query, with a darker color indicating a higher bit-score. Below the visualization all hits are listed with their corresponding sequence alignments and a hyperlink to the corresponding gene. The options menu in the top right corner allows users to select various visualizations for the BLAST hits, including gene expression levels and protein domain predictions.

the accuracy of a phylogenetic tree, we download the protein sequences of all genes in the phylogenetic tree and run more sophisticated tree reconstruction methods based on maximum likelihood or Bayesian inference.

Step 1. For a gene of interest, collect the orthology group ID

Step 2. Select all genes that belong to the ortholog group in the Gene Table by enabling the ortholog group column from the Select Column menu and pasting the ortholog ID in the Filter → Equals box

Step 3. Check the Select All checkbox in the top right of the Gene Table

Step 4. Click Download

Step 5. Select Sequences in the download menu

Step 6. Download the protein sequences of these genes

5.4.3 Annotating orthologs of known genes

We started our project on the newly sequenced and annotated genome of *P. andersonii* by taking genes of *M. trucatula* that are known to be involved in rhizobium symbiosis, and finding their *P. andersonii* orthologs. We subsequently named the *P. andersonii* genes through the GeneNoteBook editing functionality. Naming a gene is a quick way of ensuring easy retrieval of the gene at later stages. In this fashion, we named 2000 genes with a group of six people over a period of a few weeks.

Step 1. Search for an *M. truncatula* gene that is involved in symbiosis

Step 2. Interpret the ortholog group phylogenetic tree to identify the *P. andersonii* ortholog

Step 3. Browse to the *P. andersonii* ortholog and add a name attribute

5.4.4 Browsing differentially expressed genes

In a subsequent experiment, we determined gene expression levels in uninoculated roots and in root nodules that contain rhizobium bacteria (see chapter 3). We determined significant differential expression using external tools, and used GeneNoteBook to identify what these differentially expressed genes are. Specifically, we looked for known symbiosis genes by querying on the names from examples 5.4.1 and 5.4.3.

Step 1. Determine differential expression, and obtain a list of gene IDs for genes that are differentially expressed between uninoculated roots and root nodules

Step 2. Add a GeneNoteBook query for these gene IDs by pasting a list of new-line separated gene IDs in the query box

Step 3. Add a GeneNoteBook query for genes that have the Name attribute by first enabling the Name attribute in the Select Columns menu and subsequently selecting Filter → Present in the Name attribute column header.

5.5 Implementation details

GeneNoteBook is a meteor.js web app consisting of a node.js server and a JavaScript browser client. Meteor.js was chosen for three main reasons. First, meteor.js apps are designed to work with the document-oriented NoSQL database MongoDB. The NoSQL document model of MongoDB provides a flexible data model. This is convenient, since the types of data used in GeneNoteBook are often not easily represented in tabular form (i.e. a gene model is typically represented as a directed acyclic graph of exons). Second, meteor.js comes with several default functionalities such as a user account system and a job-queue system. Third, since one of the key features of GeneNoteBook is the ability to edit gene attributes, GeneNoteBook uses the meteor.js publication-subscription system over a WebSocket connection to be able to actively push changed data to connected clients. This ensures that users always receive up-to-date data and allows GeneNoteBook to function as a real-time dynamic web application. For dynamic rendering, both for the HTML user interface and the SVG visualizations, GeneNoteBook uses React.js.

5.5.1 Data types and file formats

GeneNoteBook integrates queries and visualizations for various data types (Table 5.1). The minimal amount of required data needed to run GeneNoteBook is a reference genome sequence in FASTA format, and a protein coding gene annotation in GFF3 format. Since GeneNoteBook uses the GFF3 ID attribute to uniquely identify genes, every gene interval in the GFF3 file must have a unique ID. Commonly used gene prediction algorithms generally fulfill this requirement, and GeneNoteBook is designed to work on result files from Maker [277], EvidenceModeler [278] and BRAKER [275].

Table 5.1: GeneNoteBook data types and file formats

Data type	**File format**	**Required?**
Reference genome sequence	.fasta	yes
Protein coding gene annotation	.gff3	yes
Interproscan protein domain prediction	.gff3 (interproscan)	no
Ortholog group phylogenetic trees	.newick (Orthofinder)[1]	no
Transriptome quantification	.tsv (Kallisto)[2]	no

[1] The Kallisto tab delimited file format has the following five fields: transcript ID, transcript length, effective transcript length, est counts, transcripts per million. The first line is a header and is skipped during parsing. Only the transcript ID, est counts, and transcripts per million fields are parsed.

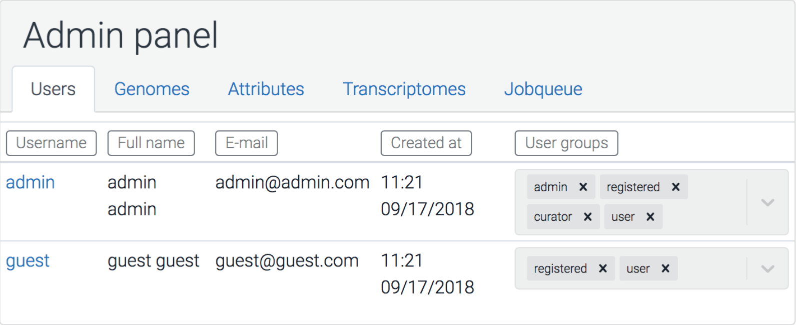
[2] Every ortholog group phylogenetic tree should be in a separate file. The file name is treated as the orthogroup identifier.

5.5.2 User accounts

To handle data permissions and allow users to save BLAST results GeneNoteBook uses a minimalistic user account system. By default, four access levels are used to distinguish between users: 1) registered, 2) user, 3) curator and 4) administrator. If someone registers an account they become 'registered', but have no access to any data yet. Once an administrator changes the account to 'user', they can start browsing data. Curators are allowed to make changes to gene attributes. Administrators have full access to all data, can assign permissions to other users and have access to the administrator menu of the GeneNoteBook instance. The administrator menu can be used to monitor and configure user profiles, permissions, BLAST databases, gene attributes, transcriptome datasets and the job queue (figure 5.9).

5.5.3 Job queue

To execute long-running processes like BLAST and preparing data for downloading, GeneNoteBook uses a job queue system. The job queue can be configured to process up to a specified maximum number of jobs in parallel. This ensures that the machine hosting GeneNoteBook will not be overloaded, but also allows upscaling for GeneNoteBook instances that are heavily used. The job queue is implemented in the database such that no external configuration is necessary. Additionally, by storing job status and results in the database, jobs can be tracked in real-time using



Username	Full name	E-mail	Created at	User groups
admin	admin admin	admin@admin.com	11:21 09/17/2018	admin × registered × curator × user ×
guest	guest guest	guest@guest.com	11:21 09/17/2018	registered × user ×

Figure 5.9: The administrator section of GeneNoteBook. This section gives an overview of registered accounts and user permissions. Additionally, the accessibility of loaded genomes, transcriptomes and annotation tracks can be configured, and current and old jobs in the job queue can be monitored and configured.The administrator section of GeneNoteBook. This section gives an overview of registered accounts and user permissions. Additionally, the accessibility of loaded genomes, transcriptomes and annotation tracks can be configured, and current and old jobs in the job queue can be monitored and configured.

GeneNoteBooks reactive publication/subscription system. To prevent old jobs from accumulating, by default a cleaning job runs every 20 minutes to remove jobs that are older than one month.

5.6 Installation options

GeneNoteBook depends on node.js, MongoDB and BLAST. To allow several varieties of dependency management and configuration, GeneNoteBook can be installed and run in three ways: (1) download a release tarball from github, (2) install through bioconda [345] or (3) run as a Docker container. When downloading the release tarball from github, all dependencies must be installed manually and the MongoDB daemon must be started manually before GeneNoteBook can start. When using the bioconda distribution, all dependencies are installed by bioconda, but again the MongoDB daemon must be started manually. When using the Docker distribution, all dependencies are installed in the GeneNoteBook Docker container and a separate Docker container for the MongoDB daemon is started automatically. More information on how to configure a GeneNoteBook deployment can be found in the online documentation at `http://genenotebook.github.io`.

5.7 Conclusion

GeneNoteBook is specifically geared towards comparative genomics, since it is designed to store multiple genomes. We have successfully used GeneNoteBook for the comparison of several plants from the genera *Parasponia* and *Trema* to study the evolution of rhizobium symbiosis [213]. To demonstrate the potential of GeneNoteBook, a public instance hosting various plant genomes is available through the online documentation. These projects have demonstrated that GeneNoteBook is useful for both experimental biologists and bioinformatics researchers. This integrative approach facilitates studies of newly sequenced organisms compared to related organisms or well-annotated model organisms. Whereas our examples include plants, GeneNoteBook permits genomic data from any organism, even over large evolutionary distances. Additionally, GeneNoteBook offers several options for a smooth installation and configuration, such as a Bioconda [345] distribution and a Docker image. As such, GeneNoteBook has the potential to be used in a wide range of genome projects.

Acknowledgements

We thank Wouter Kohlen, Eef Jonkheer and Janani Durairaj for feedback on the user interface and for testing installation procedures.

Funding

This work has been supported by the Netherlands Organization for Scientific Research [NWO-VICI 865.13.001 to R.G.] and the European Research Council [ERC-2011-AdG-294790 to T.B.]

Chapter 6

General discussion

This thesis consists largely of bioinformatic approaches to study the molecular evolution of plants engaging in a symbiotic interaction with nitrogen-fixing bacteria. The main findings of this thesis fall into two broad categories: (1) The main biological findings consist of a renewed hypothesis on how nodulation evolved in plants and several testable predictions on which genes play a role in the symbiosis, including previously unstudied candidates, and (2) the technological contribution in the field of bioinformatics comprises a computational framework for the comparative analysis of transcriptional networks and intuitive data visualization approaches. This discussion chapter briefly recapitulates these main findings to put them in a broader context and provides an outlook on future research on nitrogen-fixing plant-microbe symbioses.

6.1 A new perspective on the evolution of nodulation

Starting from the one century-old question why some plants can engage in nitrogen-fixing symbiosis and other can not [6], this thesis set out to study the molecular evolution of rhizobium symbiosis in plants at a genome-wide scale. The scattered phylogenetic distribution of plant species engaging in nitrogen-fixing root-nodule symbiosis was previously considered to be the result of multiple independent evolutionary events [21] (figure 1.1a). This multiple-gain hypothesis has been described as most parsimonious, as it requires the least amount of evolutionary gains and losses of nodulation [14]. Additionally, the multiple-gain hypothesis separates the evolutionary origin(s) of rhizobium symbiosis from the origin(s) of *Frankia* symbiosis. Whereas common features between Frankia and rhizobium symbioses had been known (see chapter 2), it was generally thought these commonalities were the result of an unobserved predisposition that primed some species to be susceptible for acquiring the ability to nodulate (figure 1.1a) [15]. However, none of these previous studies provide insight into how the molecular mechanisms underlying nitrogen-fixing symbiosis evolved to give rise to the nodulation trait. To establish a molecular perspective on the multiple occurrences of rhizobium symbiosis, newly sequenced genomes of symbiotic *Parasponia* and non-nodulating *Trema* were analyzed together with publicly available datasets for known nodulating and non-nodulating lineages.

Since rhizobium symbiosis co-opted part of the molecular signalling pathway used in arbuscular mycorrhizal (AM) symbiosis [44, 73],the phylogenomic approach taken in this thesis took inspiration from previous studies on the evolution of AM symbiosis. Specifically, an approach similar to Delaux et al. [51] was used to identify protein-coding genes that are present in symbiotic species and absent in non-symbiotic species. Unlike AM symbiosis, which is considered to have a single origin at the emergence of land plants 450 million years ago [253], rhizobium symbiosis was generally thought to have evolved independently at least twice: at least once in legumes (Fabaceae) and once in *Parasponia* (Cannabaceae) [21]. As a consequence, where phylogenomic analysis of AM symbiosis focussed on gene loss from the start, we hypothesized that gene gain led to the evolution of rhizobium symbiosis.

Chapter 3 establishes that gene loss plays a major role in the absence / presence of rhizobium symbiosis in the Rosales, a finding that has subsequently been corroborated for both *Frankia* and rhizobium symbiosis in the nitrogen-fixing clade using broader taxonomic sampling [295]. Specifically, it seems that three known symbiosis genes are predominantly lost in non-nodulating plant species in the nitrogen-fixing clade: *NOD FACTOR PERCEPTION 2* (*NFP2*), *NODULE INCEPTION* (*NIN*), and *RHIZOBIUM-DIRECTED POLAR GROWTH* (*RPG*). This novel finding has redefined how we view the evolution of nitrogen-fixing symbiosis. As a consequence it is now more likely that there was a single gain of nodulation with multiple independent losses (figure 1.1b).

The renewed appreciation of the single-gain hypothesis has important implications for our understanding of the molecular mechanisms of nodulation. Specifically, the single-gain hypothesis raises several new questions.

6.1.1 Why did so many lineages lose nodulation?

In all lineages that lost nitrogen-fixing symbiosis, the benefits of the symbiosis likely no longer outweighed its drawbacks. The added benefit of microbe-provided nitrogen is limited when sufficient environmental nitrogen is present. This idea is reinforced by the observation that legumes, actinorhizal plants and *Parasponia* can prevent symbiotic interactions under high environmental nitrogen concentrations [346]. Whereas this control of symbiotic interactions by the plant is an active and reversible process, an extended release of dependence on symbiont-provided nitrogen could have led to permanent abolishment of symbiotic interactions. To motivate why microbe-provided nitrogen became redundant, it has been suggested that a decrease in atmospheric CO_2 concentrations could be a driving force behind a relaxed selective pressure on nitrogen-fixing symbiosis [343]. In this perspective, the rate-limiting step in plant growth changed from being nitrogen-based to being carbon-based. In other mutualisms, similar breakdowns in mutualistic dependence have previously been reported [347]. Taken together, following the 'use it or lose it' principle, extant lineages that lost the ability to form a nitrogen-fixing symbiosis simply did not need the microbe-provided nitrogen.

6.1.2 What did the ancestral nitrogen-fixing symbiosis look like?

An important implication of a single gain of nitrogen-fixing symbiosis at the base of the nitrogen-fixing clade is that this hypothesized symbiotic ancestor has given rise to extant taxa that exclusively form mutualistic interactions with either *Frankia* or rhizobium bacteria. As a consequence, it seems plausible that the symbiotic ancestor at the base of the nitrogen fixation clade formed a mutualistic interaction with either *Frankia* or rhizobium bacteria, and that subsequently symbiont switches have occurred. The concept of symbiont switching is not confined to nodulation, and it has been described for the arbuscular mycorrhizal and ectomycorrhizal symbioses as well [348]. Since *Frankia* bacteria have intrinsic mechanisms to protect their nitrogenase enzyme from high oxygen concentrations [257], they potentially require simpler symbiotic mechanisms from the plant. It has therefore been hypothesized that the ancestral symbiotic plant formed a mutualistic interaction with *Frankia* bacteria [343].

The notion that ancestral nodulation used *Frankia* is supported by the recent finding that the development of actinorhizal-type nodules[1] and legume-type nodules is more similar than previously thought [349]. Additionally, mutation of the *Medicago truncatula* transcriptional regulator *NODULE ROOT1* (*MtNOOT1*) led to a switch from a legume-type nodule to an actinorhizal-type nodule. Taken together, this suggests a shared evolutionary origin of both types of nodules, with the actinorhizal nodule type being ancestral [349]. Furthermore, these reported findings dispute an earlier idea that the ancestral nitrogen-fixing symbiosis only consisted of intracellularly housed microbial symbionts, and that the ability to form nodules evolved multiple times independently [350]. In this perspective, intracellular infection acts as the predisposition to acquire the nodulation trait. This notion of intracellular infection as predisposition currently cannot be verified, but seems unlikely since intracellular infection also occurs outside of the nitrogen-fixing clade in species from the genus *Gunnera* [351]. Ultimately, conclusive evidence on whether the ancestral nitrogen-fixing symbiosis formed nodules will have to come from the fossil record

[1] This includes nodules formed by rhizobium bacteria on the roots of *Parasponia* species.

[343, 350].

Given that there are both symbiotic *Frankia* and rhizobium strains that do and do not produce lipochitooligosaccharides (LCOs), an important question is how LCOs were involved in the ancestral symbiosis. The LCO-receptor *NFP2* was duplicated at the base of the nitrogen-fixing clade (see chapter 3). This makes it likely that LCOs were involved in the ancestral symbiosis, indicating that current symbioses that do not require LCO signalling would have lost this trait. Whether LCO perception has been the first evolutionary step towards nitrogen-fixing mutualism is unclear, but seems unlikely given the involvement of LCO signalling in the ancestral arbuscular mycorrhizal symbiosis.

6.1.3 What is the prospect for engineering?

The observation that nodulation likely originated once and was lost many times has several important implications for engineering nitrogen-fixing symbiosis in crops. At minimum the consistently lost genes will have to be reconstituted in non-nodulating species of the nitrogen-fixing clade. Additionally, the possibility of species-specific losses is high and will have to be taken into account. A discouraging parallel can be found in the loss of limbs in snakes [352]. There, careful comparative analysis identified the loss of a transcriptional enhancer crucial for limb development. Upon reconstituting a functional enhancer sequence, expression of the enhancer target was returned to normal, but did not result in snakes with limbs. Ultimately, the authors concluded that other processes had been lost as the result of “functional erosion”. Whether similar functional erosion has occurred in species that lost the ability to form nodules is unknown. However, to minimize the effects of functional erosion, engineering efforts should focus on lineages that have lost nodulation only recently, such as *Trema*.

6.2 Loss of genes essential for nodulation

The observation in chapter 3 that in the Rosales loss of only three symbiosis genes (*NFP2*, *NIN*, and *RPG*) strongly correlates with the absence of nodulation raises the question why specifically these three genes are lost. The aforementioned argument on why nitrogen-fixation as a whole was lost can be extended to the loss of individual genes. Again, following the principle of “use it or lose it”, only genes that are dispensable are lost in evolution [353]. This implies that consistently lost symbiosis genes are likely exclusively used in nitrogen-fixing symbiosis and were therefore not involved in other essential processes. To further explore this prediction, the following section highlights several recent findings on *NFP2*, *NIN*, and *RPG*, and whether these genes indeed function exclusively in nitrogen-fixing symbiosis.

6.2.1 NFP2

NFP2 is a lysin motif-containing (LysM) receptor involved in detecting rhizobial LCOs. Recent work showed that *NFP2* arose from a duplication event at the base of the nitrogen-fixing clade, giving rise to *NFP1* and *NFP2* orthogroups [354]. The *NFP2* orthogroup contains receptors that are essential for rhizobium LCO recognition in legumes (e.g. *MtNFP* in *Medicago truncatula* and *LjNFR5* in *Lotus japonicus*) and the non-legume *Parasponia* (*PanNFP2*). However, it is unlikely that *NFP2*-type genes neofunctionalized to become exclusive for nodulation, as the non-nodulating legume *Cercis canadensis* has a functional *NFP2* gene, indicating there it is used in processes other than detecting bacterial LCOs. Furthermore, all actinorhizal species from the Fagales have lost *NFP2*. Finally, the *Pannfp1* mutant in *Parasponia andersonii* has a reduced number of nodules compared to wildtype, indicating *P. andersonii* uses both *PanNFP1* and *PanNFP2* in detecting rhizobium LCOs.

In a broader perspective, other members of the LysM receptor gene family are implicated in recognition of arbuscular mycorrhizae-derived LCO molecules [167, 355, 356]. Given that AM symbiosis predates the origin of nitrogen-fixing symbiosis, it is likely that the ancestral *NFP* gene had the capacity to detect LCO molecules. This idea is supported by the observation that *NFP* proteins of two species outside of the nitrogen-fixing clade (*Petunia hybrida* and *Solanum lycopersicum*) can (partially) transcomplement *Mtnfp* and *Ljnfr5* mutant phenotypes in *M. truncatula* and *L. japonicus*. In line with this it cannot be excluded that the ancestral *NFP* receptor protein already enabled LCO-based nitrogen-fixing symbiosis, after which the duplication into *NFP1* and *NFP2* allowed for partial neo- or subfunctionalization.

One key aspect of an ancestral LCO-based nitrogen-fixing symbiosis is that the ancestral symbiont should have produced LCOs. This chapter previously discussed that a single gain of nitrogen-fixing symbiosis implies a single ancestral symbiont, and subsequent symbiont switches. The recent findings on the duplication of *NFP* shed some light on whether the ancestral symbiont was likely rhizobium or *Frankia*. In both rhizobium and *Frankia*, lineages that do and do not produce LCOs exist. Combined with the observation that only actinorhizal lineages interacting with (potential)

LCO-producing *Frankia* have a functional *NFP2* gene [354], the ancestral symbiosis likely predates the *NFP1-NFP2* divergence and involved an LCO producing *Frankia*. To further substantiate this hypothesis, functional studies on *NFP1* and *NFP2* in actinorhizal lineages are indispensable.

6.2.2 NIN

NIN is a transcription factor that is crucial for nodulation in legumes. Recent work in the *Parasponia*-rhizobium and *Casuarina-Frankia* symbioses revealed similar roles [159, 357]. It is therefore likely that *NIN* functions as a master regulator in nodulation throughout the nitrogen-fixing clade.

As *NIN* is indispensable for nodulation, but not used in arbuscular mycorrhizal symbiosis [358], it most likely adopted a new function at the base of the nitrogen-fixing clade. Whether this was complete neofunctionalization or subfunctionalization is currently not known. However, subfunctionalization is plausible given the presence of *NIN* in the non-symbiotic *Ziziphus jujuba* (Rosales) and *Juglans regia* (Fagales), two species that are thought to have lost nodulation [213, 295]. The distinction between sub- and neofunctionalization is important, as it implies a non-nodulating role for *NIN* in the nitrogen-fixing clade that is currently unknown. Identifying the possible non-nodulating role of *NIN* can help shed light on the transition from non-nodulating to nodulating function of the protein.

Since *NIN* is a transcription factor, transcriptional rewiring likely plays a role in the evolution of *NIN*. As transcription factors are regulators that themselves can be regulated, transcriptional rewiring refers to the (possibly indirect) effects of changes in regulatory interactions during evolution. In the case of *NIN*, there are two ways transcriptional rewiring can have occurred: (1) *NIN*-targets are different for nodulating and non-nodulating plants, or (2) *NIN* itself is targeted differently in nodulating plants. In both scenarios, the change would have occurred at the root of the nitrogen-fixing clade. Given the age of the nitrogen-fixing clade, it is entirely possible that currently both scenarios can be observed as the result of multiple evolutionary events. However, to understand the molecular evolution of the function of *NIN* in nodulation, the order of events has to be determined. This can be done with the approach that was used in chapter 3 to identify gene loss events by instead looking for shared transcriptional interactions across several species.

In an attempt to characterize evolutionary rewiring in the transcriptional network surrounding *NIN*, chapter 4 sought to identify to what extent conserved transcriptional interactions exist in a range of nodulating and non-nodulating species. Unfortunately, current publicly available datasets and algorithms proved insufficient to reconstruct transcriptional networks involved in symbiosis from bulk RNA sequencing data. Nevertheless, the computational framework to compare transcriptional interactions between species based on orthogroups will be of use once it is possible to accurately measure or predict these interactions in multiple species.

Currently, the crude readout of transcriptional activity that is provided by bulk RNA sequencing can still be useful for identifying general gene-expression patterns. For example, bulk RNA sequencing of *P.andersonii* and *P. andersonii* × *Trema orientalis* F_1 hybrid nodules and roots allowed for the identification of two separate transcriptional programs for nodule organogenesis and intracellular infection (Chapter 3). However, because bulk RNA sequencing data is pooled across cell types and tissues, using this data alone is not sufficient to make accurate statements on mechanistic interactions. One solution to this problem is isolating different tissues to a greater resolution, for example using laser microdissection as in Roux et al. [226]. Alternatively, to achieve a cell-specific resolution, single-cell RNA sequencing techniques could be employed [359].

Ultimately, to study how transcriptional rewiring is involved in the evolution of rhizobium symbiosis, it is essential that transcription factor targets can be identified accurately (see chapter 4). Future studies on transcriptional rewiring will require both novel data and accompanying algorithms. Quantifying gene expression in time series experiments is a straightforward way to incorporate directionality into predicting transcriptional networks that can be applied in most systems. A more direct approach is to quantify gene expression in a knockout experiment, but this is not feasible in most species. Alternatively, determining expression quantitative trait loci (eQTL) is a powerful way of linking genetic variation to gene expression for model systems that have a mapping population [360, 361]. In species for which no mapping population can be established, genome-wide association studies (GWAS) on gene expression can provide a similar link between genetic variation and gene expression [362]. Finally, several medium- to high-throughput methods exist for directly inferring transcriptional networks: chromatin immunoprecipitation sequencing (ChIP-seq) [363], DNA affinity purification sequencing (DAP-seq) [326], and DNase I hypersensitive sites sequencing (DNAse-seq) [325]. However, these methods always require a significant amount of lab work and often can only be performed in systems amenable to genetic modification. Taken together, whereas there are several opportunities to increase the accuracy of predicting transcriptional networks, doing so in a range of symbiotic and non-symbiotic plant species remains a challenge.

In the meantime, low-throughput studies have added a few more pieces to the puzzle that is the evolution of symbiotic function of *NIN*. Whereas several potential targets and regulators of *NIN* have been described [32, 33, 40, 43, 218, 316, 317, 357] in a symbiotic context, it is unclear if orthologs of the same proteins outside of the nitrogen-fixing clade interact in a similar fashion. A recent addition to our understanding of symbiotic *NIN* targets is the observation that *NIN* targets *LATERAL ORGAN BOUNDARIES-DOMAIN 16* (*LBD16*) in *M. truncatula* and *L. japonicus*, a protein involved in lateral root formation [364, 365]. Whether the interaction between *NIN* and *LBD16* is fully specific for symbiosis is not clear [43]. However, in *Arabidopsis thaliana*, the *NIN*-paralog *NIN-LIKE-PROTEIN7* (*NLP7*), interacts with an *LBD16* ortholog [366]. Additionally, it was recently found that *NIN* physically interacts with paralogous members of the *NLP* gene family [367] in *M. truncatula*. Combined with the observations that multiple members of the *NLP* gene family can bind to the same promotor elements in both *Arabidopsis thaliana* [368] and *L. japonicus* [369], transcriptional rewiring in the *NLP* gene family seems to be linked with the evolution of nodulation [43]. This suggests that the symbiotic transcriptional network including *NIN* might include regulators and targets of other *NLP* genes, and therefore be much larger than currently is known. An approach similar to the one described in chapter 4 will be indispensable for studying the evolution of *NIN*, ideally based on the high-throughput techniques previously described.

6.2.3 RPG

RPG is commonly thought to be involved in the infection process that follows after bacterial recognition. The *rpg* phenotype has been described as being characterized by delayed and abnormal root hair curling and infection threads in *M. truncatula* [229]. Recent work has shown that *NIN* is required for *RPG* expression, and that *NIN* seems to bind to the *RPG* promoter *in vivo* [43]. In addition, it was found that several other infection-related proteins form a complex that is crucial for polar growth of infection threads in *Medicago* [43], the process after which *RPG* was named. Some of these polar growth related proteins also display a *NIN*-dependent expression pattern. As such, a pattern emerges that in legumes a *NIN*-regulated protein complex involves *RPG* and is tightly linked to symbiotic bacterial infection. However, it is now also clear that *RPG* is not required for functional nitrogen-fixing symbiosis in all cases: both *Arachis hypogaea* (peanut) and *Aeschynomene evenia* do not have an *RPG* ortholog, and instead form symbiotic interactions without the formation of canonical infection threads [295, 370, 371]. Furthermore, is is now clear that an *RPG* ortholog is present in the non-nodulating black raspberry (*Rubus occidentalis*) [295]. Taken together, it is unlikely that evolution of *RPG* has been causal for the origin of nodulation. Similar to *NFP2* and *NIN*, differentiating the symbiotic from non-symbiotic function of *RPG* will require studies on non-nodulating species within and outside of the nitrogen-fixing clade.

6.2.4 What about other genes?

Whereas the previous section highlighted novel findings and open questions on the evolution of three known nodulation genes, chapter 3 identified four more genes that are consistently lost in *Trema* species (figure 3.6). The nodule-specific expression pattern of these four genes in *P. andersonii* suggests they are adapted to function in nodulation in *Parasponia*. As such, the pseudogenized orthologs in *Trema* likely represent cases of functional erosion, and therefore could be prime targets for re-engineering nodulation in this lineage.

Relatedly, an important consideration in interpreting the findings of this thesis is the taxonomic diversity of the studied species. Whereas in this thesis care was taken to include a diverse set of species from the nitrogen-fixing clade with relevant outgroups, the notion that a handful of species can represent the taxonomic diversity of all four orders in the nitrogen-fixing clade is a strong simplification. This taxonomic perspective is especially important when asserting that presence/absence patterns must be consistent across all species: the increased taxonomic sampling used in Griesmann et al. [295] and Rutten et al. [354] has revealed several deviations from the presence/absence patterns described in chapter 3. Consequently, there is now no single gene known that is consistenly present in nodulating lineages and absent in non-nodulating lineages. As a result, studying the molecular evolution of nitrogen-fixing symbiosis now deals with two new questions: (1) what lineage-specific losses and adaptations have occurred within the nitrogen-fixing clade, and (2) what makes lineages within the nitrogen-fixing clade different from other lineages? Answering both questions will require increased taxonomic sampling. For example, the genus *Dryas* has been described to include a nodulating / non-nodulating pair similar to *Parasponia* / *Trema* [372]. Including these species in future analyses will identify relevant lineage-specific losses and adaptations.

Fortunately, it is now feasible to assemble, annotate, and analyze the genomes of previously unstudied plant species within the time frame of a single PhD project. Nevertheless, there are some clear limitations in several of the molecular techniques and associated algorithms used in this thesis. For example, the use of short-read sequencing data for eukaryotic genome assembly has been largely superseded by third generation long-read sequencing technologies [373]. In chapter 3 the assembly of *Parasponia* and *Trema* genomes using short-read data took weeks on a high-performance compute facility and resulted in usable but fragmented assemblies. In contrast, it has recently become possible to assemble the genome of *A. thaliana* to a higher accuracy in one hour on a single computer [374] due to the use of long-read sequencing data. Given that it is now clear that some of the conclusions in chapter 3 might be influenced by limited taxonomic sampling, future studies including more species will benefit from faster, cheaper, and more accurate genome assemblies.

Given the far-reaching conclusions that result from the identification of symbiotic gene loss, a critical aspect of comparative studies focussing on newly sequenced genomes is the accuracy of predicted protein-coding gene models. The effect of incorrectly annotated genes on potential biological conclusions can be summarized with two examples: (1) automated curation of protein-coding genes in *Parasponia* and *Trema* revealed that $\sim$15% of orthologous genes were incorrectly annotated to be different, and (2) the *nin* pseudogene in *Prunus persica* was incorrectly annotated as functional protein-coding gene. The simultaneous annotation of two closely related genomes in chapter 3 was instrumental in minimizing genome-specific artefacts. Recent algorithmic developments acknowledge this need and tackle the challenge of annotating multiple related genomes simultaneously [375], or focus on postprocessing of multiple annotations [376]. In future efforts, novel sequenced genomes are only as useful as the quality of their annotation.

6.3 Conclusion

This thesis has renewed interest in the single-gain hypothesis of rhizobium symbiosis (figure 1.1). As a consequence, both rhizobium and *Frankia* symbiosis are now thought to share a single ancestral origin. Future efforts will focus on testing various mechanistic predictions that are a result of this altered perspective, including more detailed work on transcriptional rewiring. High-throughput molecular approaches are indispensable for these future efforts. Consequently there will have to be coordinated efforts to develop the required computational tools. Currently, we are one step closer to a unified mechanistic perspective on the evolutionary origin of nitrogen-fixing plant-microbe symbiosis.

References

[1] Boucher DH. *The Biology of Mutualism: Ecology and Evolution*. Oxford University Press, 1985.

[2] Zimmer S et al. “Effects of soybean variety and Bradyrhizobium strains on yield, protein content and biological nitrogen fixation under cool growing conditions in Germany”. *European Journal of Agronomy* 72 (2016), pp. 38–46.

[3] Warembourg FR et al. “Economy of Symbiotically Fixed Nitrogen in Red Clover (Trifolium pratenseL.)” *Annals of Botany* 80.4 (1997), pp. 515–523.

[4] FAO. *World fertilizer trends and outlook to 2022*. Tech. rep. FOOD and AGRICULTURE ORGANIZATION OF THE UNITED NATIONS, 2019.

[5] Hollaender A et al., eds. *Genetic Engineering for Nitrogen Fixation*. Springer, Boston, MA, 1977.

[6] Burrill TJ and Hansen R. “Is symbiosis possible between legume bacteria and non-legume plants?” *Agr. Exp. Sta. Bull* 202 (1917), pp. 115–181.

[7] Khush GS and Bennett J. *Nodulation and Nitrogen Fixation in Rice: Potential and Prospects*. Int. Rice Res. Inst., 1992.

[8] Rogers C and Oldroyd GED. “Synthetic biology approaches to engineering the nitrogen symbiosis in cereals”. *Journal of Experimental Botany* 65.8 (2014), pp. 1939–1946.

[9] Soltis DE et al. “Chloroplast gene sequence data suggest a single origin of the predisposition for symbiotic nitrogen fixation in angiosperms”. *Proceedings of the National Academy of Sciences of the United States of America* 92.7 (1995), pp. 2647–2651.

[10] Sun M et al. “Phylogeny of the Rosidae: A dense taxon sampling analysis”. *Journal of Systematics and Evolution* 54.4 (2016), pp. 363–391.

[11] Li HL et al. “Large-scale phylogenetic analyses reveal multiple gains of actinorhizal nitrogen-fixing symbioses in angiosperms associated with climate change”. *Scientific Reports* 5 (2015), p. 14023.

[12] Swensen SM. “The Evolution of Actinorhizal Symbioses: Evidence for Multiple Origins of the Symbiotic Association”. *American Journal of Botany* 83.11 (1996), pp. 1503–1512.

[13] Doyle JJ. “Phylogenetic perspectives on nodulation: evolving views of plants and symbiotic bacteria”. *Trends in Plant Science* 3.12 (1998), pp. 473–478.

[14] Doyle JJ. “Phylogenetic perspectives on the origins of nodulation”. *Molecular plant-microbe interactions: MPMI* 24.11 (2011), pp. 1289–1295.

[15] Werner GDA et al. “A single evolutionary innovation drives the deep evolution of symbiotic N2-fixation in angiosperms”. *Nature Communications* 5.1 (2014), p. 4087.

[16] Vessey JK et al. “Root-based N 2 -fixing symbioses: Legumes, actinorhizal plants, Parasponia sp. and cycads”. *Plant and Soil* 266.1 (2005), pp. 205–230.

[17] Sato S et al. “Genome structure of the legume, Lotus japonicus”. *DNA research: an international journal for rapid publication of reports on genes and genomes* 15.4 (2008), pp. 227–239.

[18] Young ND et al. “The Medicago genome provides insight into the evolution of rhizobial symbioses”. *Nature* 480.7378 (2011), pp. 520–524.

[19] Schmutz J et al. “Genome sequence of the palaeopolyploid soybean”. *Nature* 463.7278 (2010), pp. 178–183.

[20] Behm JE et al. “Parasponia: a novel system for studying mutualism stability”. *Trends in Plant Science* 19.12 (2014), pp. 757–763.

[21] Geurts R et al. “Exploiting an ancient signalling machinery to enjoy a nitrogen fixing symbiosis”. *Current Opinion in Plant Biology* 15.4 (2012), pp. 438–443.

[22] Honma MA et al. “Rhizobium meliloti nodD genes mediate host-specific activation of nodABC”. *Journal of bacteriology* 172.2 (1990), pp. 901–911.

[23] Arrighi JF et al. “The Medicago truncatula Lysine Motif-Receptor-Like Kinase Gene Family Includes NFP and New Nodule-Expressed Genes”. *Plant Physiology* 142.1 (2006), pp. 265–279.

[24] Limpens E et al. “LysM domain receptor kinases regulating rhizobial Nod factor-induced infection”. *Science* 302.5645 (2003), pp. 630–633.

[25] Madsen EB et al. “A receptor kinase gene of the LysM type is involved in legume perception of rhizobial signals”. *Nature* 425.6958 (2003), pp. 637–640.

[26] Radutoiu S et al. “Plant recognition of symbiotic bacteria requires two LysM receptor-like kinases”. *Nature* 425.6958 (2003), pp. 585–592.

[27] Broghammer A et al. “Legume receptors perceive the rhizobial lipochitin oligosaccharide signal molecules by direct binding”. *Proceedings of the National Academy of Sciences of the United States of America* 109.34 (2012), pp. 13859–13864.

[28] Rose CM et al. “Rapid phosphoproteomic and transcriptomic changes in the rhizobia-legume symbiosis”. *Molecular & cellular proteomics: MCP* 11.9 (2012), pp. 724–744.

[29] Ehrhardt DW et al. “Calcium spiking in plant root hairs responding to Rhizobium nodulation signals”. *Cell* 85.5 (1996), pp. 673–681.

[30] Lévy J et al. “A putative Ca2+ and calmodulin-dependent protein kinase required for bacterial and fungal symbioses”. *Science* 303.5662 (2004), pp. 1361–1364.

[31] Yano K et al. “CYCLOPS, a mediator of symbiotic intracellular accommodation”. *Proceedings of the National Academy of Sciences of the United States of America* 105.51 (2008), pp. 20540–20545.

[32] Singh S et al. “CYCLOPS, a DNA-binding transcriptional activator, orchestrates symbiotic root nodule development”. *Cell host & microbe* 15.2 (2014), pp. 139–152.

[33] Soyano T et al. “Nodule inception directly targets NF-Y subunit genes to regulate essential processes of root nodule development in Lotus japonicus”. *PLoS genetics* 9.3 (2013), e1003352.

[34] Oldroyd GED et al. “The rules of engagement in the legume-rhizobial symbiosis”. *Annual review of genetics* 45 (2011), pp. 119–144.

[35] Garrocho-Villegas V et al. “Plant hemoglobins: what we know six decades after their discovery”. *Gene* 398.1-2 (2007), pp. 78–85.

[36] Bergersen FJ et al. “Effects of O2 Concentrations and Various Haemoglobins on Respiration and Nitrogenase Activity of Bacteroids from Stem and Root Nodules of Sesbania rostrata and of the Same Bacteria from Continuous Cultures”. *Microbiology* 132.12 (1986), pp. 3325–3336.

[37] Udvardi M and Poole PS. “Transport and metabolism in legume-rhizobia symbioses”. *Annual review of plant biology* 64 (2013), pp. 781–805.

[38] Reid DE et al. “Molecular mechanisms controlling legume autoregulation of nodulation”. *Annals of Botany* 108.5 (2011), pp. 789–795.

[39] Streeter J and Wong PP. “Inhibition of legume nodule formation and N2 fixation by nitrate”. *Critical Reviews in Plant Sciences* 7.1 (1988), pp. 1–23.

[40] Soyano T and Hayashi M. “Transcriptional networks leading to symbiotic nodule organogenesis”. *Current Opinion in Plant Biology* 20 (2014), pp. 146–154.

[41] Freeman M. “Feedback control of intercellular signalling in development”. *Nature* 408.6810 (2000), pp. 313–319.

[42] Untergasser A et al. “One-step Agrobacterium mediated transformation of eight genes essential for rhizobium symbiotic signaling using the novel binary vector system pHUGE”. *PLOS One* 7.10 (2012), e47885.

[43] Liu J et al. “A Remote cis-Regulatory Region Is Required for NIN Expression in the Pericycle to Initiate Nodule Primordium Formation in Medicago truncatula”. *The Plant Cell* (2019).

[44] Oldroyd GED. “Speak, friend, and enter: signalling systems that promote beneficial symbiotic associations in plants”. *Nature reviews. Microbiology* 11 (2013), p. 252.

[45] Parniske M. “Arbuscular mycorrhiza: the mother of plant root endosymbioses”. *Nature Reviews. Microbiology* 6.10 (2008), pp. 763–775.

[46] Gough C and Cullimore J. “Lipo-chitooligosaccharide signaling in endosymbiotic plant-microbe interactions”. *Molecular Plant-Microbe Interactions* 24.8 (2011), pp. 867–878.

[47] Op den Camp R et al. “LysM-type mycorrhizal receptor recruited for rhizobium symbiosis in nonlegume Parasponia”. *Science* 331.6019 (2011), pp. 909–912.

[48] Gherbi H et al. “SymRK defines a common genetic basis for plant root endosymbioses with arbuscular mycorrhiza fungi, rhizobia, and Frankiabacteria”. *Proceedings of the National Academy of Sciences of the United States of America* 105.12 (2008), pp. 4928–4932.

[49] Harrison MJ. “The Arbuscular Mycorrhizal Symbiosis”. *Plant-Microbe Interactions*. Ed. by Stacey G and Keen NT. Boston, MA: Springer US, 1997, pp. 1–34.

[50] Delaux PM et al. “Evolution of the plant–microbe symbiotic ‘toolkit’”. *Trends in Plant Science* 18.6 (2013), pp. 298–304.

[51] Delaux PM et al. “Comparative phylogenomics uncovers the impact of symbiotic associations on host genome evolution”. *PLoS genetics* 10.7 (2014), e1004487.

[52] Bravo A et al. “Genes conserved for arbuscular mycorrhizal symbiosis identified through phylogenomics”. *Nature plants* 2 (2016), p. 15208.

[53] Shendure J et al. “DNA sequencing at 40: past, present and future”. *Nature* 550.7676 (2017), pp. 345–353.

[54] Franklin RE et al. “The structure of sodium thymonucleate fibres. I. The influence of water content”. *Acta crystallographica* 6.8-9 (1953), pp. 673–677.

[55] Watson JD and Crick FH. “Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid”. *Nature* 171.4356 (1953), pp. 737–738.

[56] Sanger F and Tuppy H. “The amino-acid sequence in the phenylalanyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates”. *Biochemical Journal* 49.4 (1951), pp. 481–490.

[57] Sanger F and Tuppy H. “The amino-acid sequence in the phenylalanyl chain of insulin. I. The identification of lower peptides from partial hydrolysates”. *Biochemical Journal* 49.4 (1951), pp. 463–481.

[58] Sanger F et al. “DNA sequencing with chain-terminating inhibitors”. *Proceedings of the National Academy of Sciences of the United States of America* 74.12 (1977), pp. 5463–5467.

[59] Staden R. “A strategy of DNA sequencing employing computer programs”. *Nucleic Acids Research* 6.7 (1979), pp. 2601–2610.

[60] Felsenstein J. “Evolutionary trees from DNA sequences: a maximum likelihood approach”. *Journal of molecular evolution* 17.6 (1981), pp. 368–376.

[61] International Human Genome Sequencing Consortium. “Finishing the euchromatic sequence of the human genome”. *Nature* 431.7011 (2004), pp. 931–945.

[62] Goff SA et al. “Chapter Three - The Evolution of Plant Gene and Genome Sequencing”. *Advances in Botanical Research*. Ed. by Paterson AH. Vol. 69. Academic Press, 2014, pp. 47–90.

[63] Arabidopsis Genome Initiative. “Analysis of the genome sequence of the flowering plant Arabidopsis thaliana”. *Nature* 408.6814 (2000), pp. 796–815.

[64] Hesper B and Hogeweg P. “Bioinformatica: een werkconcept”. *Kameleon* 1.6 (1970), pp. 28–29.

[65] Hogeweg P. “The roots of bioinformatics in theoretical biology”. *PLoS computational biology* 7.3 (2011), e1002021.

[66] Sboner A et al. “The real cost of sequencing: higher than you think!” *Genome Biology* 12.8 (2011), p. 125.

[67] Gauthier J et al. “A brief history of bioinformatics”. *Briefings in Bioinformatics* 20.6 (2019), pp. 1981–1996.

[68] De Vega JJ et al. “Red clover (Trifolium pratense L.) draft genome provides a platform for trait improvement”. *Scientific Reports* 5 (2015), p. 17394.

[69] Varshney RK et al. “Draft genome sequence of chickpea (Cicer arietinum) provides a resource for trait improvement”. *Nature Biotechnology* 31.3 (2013), pp. 240–246.

[70] Velasco R et al. “The genome of the domesticated apple (Malus× domestica Borkh.)” *Nature Genetics* 42.10 (2010), pp. 833–839.

[71] Verde I et al. “The high-quality draft genome of peach (Prunus persica) identifies unique patterns of genetic diversity, domestication and genome evolution”. *Nature Genetics* 45.5 (2013), pp. 487–494.

[72] Huang S et al. “The genome of the cucumber, Cucumis sativus L”. *Nature Genetics* 41.12 (2009), pp. 1275–1281.

[73] Holmer R et al. “Commonalities in Symbiotic Plant-Microbe Signalling”. *Advances in Botanical Research*. Vol. 82. Elsevier, 2017, pp. 187–221.

[74] Jones DL et al. “Carbon flow in the rhizosphere: carbon trading at the soil–root interface”. *Plant and soil* 321.1-2 (2009), pp. 5–33.

[75] Coleman-Derr D and Tringe SG. “Building the crops of tomorrow: advantages of symbiont-based approaches to improving abiotic stress tolerance”. *Frontiers in microbiology* 5 (2014), p. 283.

[76] Mendes R et al. “The rhizosphere microbiome: significance of plant beneficial, plant pathogenic, and human pathogenic microorganisms”. *FEMS microbiology reviews* 37.5 (2013), pp. 634–663.

[77] Redecker D et al. “An evidence-based consensus for the classification of arbuscular mycorrhizal fungi (Glomeromycota)”. *Mycorrhiza* 23.7 (2013), pp. 515–531.

[78] Remy W et al. “Four hundred-million-year-old vesicular arbuscular mycorrhizae”. *Proceedings of the National Academy of Sciences of the United States of America* 91.25 (1994), pp. 11841–11843.

[79] Simon L et al. “Origin and diversification of endomycorrhizal fungi and coincidence with vascular land plants”. *Nature* 363.6424 (1993), pp. 67–69.

[80] Wang B and Qiu YL. “Phylogenetic distribution and evolution of mycorrhizas in land plants”. *Mycorrhiza* 16.5 (2006), pp. 299–363.

[81] Balestrini R and Bonfante P. “Cell wall remodeling in mycorrhizal symbiosis: a way towards biotrophism”. *Frontiers in plant science* 5 (2014), p. 237.

[82] Martin F et al. “Unearthing the roots of ectomycorrhizal symbioses”. *Nature reviews. Microbiology* 14.12 (2016), pp. 760–773.

[83] Smith SE and Read DJ. *Mycorrhizal Symbiosis*. Academic Press, 2010.

[84] Tedersoo L and Smith ME. “Lineages of ectomycorrhizal fungi revisited: Foraging strategies and novel lineages revealed by sequences from belowground”. *Fungal biology reviews* 27.3 (2013), pp. 83–99.

[85] Plett JM and Martin F. “Poplar root exudates contain compounds that induce the expression of MiSSP7 in Laccaria bicolor”. *Plant Signaling & Behavior* 7.1 (2012), pp. 12–15.

[86] Kohler A et al. “Convergent losses of decay mechanisms and rapid turnover of symbiosis genes in mycorrhizal mutualists”. *Nature Genetics* 47.4 (2015), pp. 410–415.

[87] Navarro-Ródenas A et al. “Laccaria bicolor aquaporin LbAQP1 is required for Hartig net development in trembling aspen (Populus tremuloides)”. *Plant, cell & environment* 38.11 (2015), pp. 2475–2486.

[88] Vayssières A et al. “Development of the Poplar-Laccaria bicolor Ectomycorrhiza Modifies Root Auxin Metabolism, Signaling, and Response”. *Plant Physiology* 169.1 (2015), pp. 890–902.

[89] Gea L et al. “Structural aspects of ectomycorrhiza of Pinus pinaster (Ait.) Sol. formed by an IAA-overproducer mutant of Hebeloma cylindrosporum Romagnési”. *The New Phytologist* 128.4 (1994), pp. 659–670.

[90] Gtari M et al. “Diversity of Frankia Strains, Actinobacterial Symbionts of Actinorhizal Plants”. *Symbiotic Endophytes*. Ed. by Aroca R. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 123–148.

[91] Pawlowski K and Demchenko KN. “The diversity of actinorhizal symbiosis”. *Protoplasma* 249.4 (2012), pp. 967–979.

[92] Remigi P et al. “Symbiosis within Symbiosis: Evolving Nitrogen-Fixing Legume Symbionts”. *Trends in microbiology* 24.1 (2016), pp. 63–75.

[93] Geurts R et al. “What does it take to evolve a nitrogen-fixing endosymbiosis?” *Trends in Plant Science* (2016).

[94] Beauchemin NJ et al. “Casuarina root exudates alter the physiology, surface properties, and plant infectivity of Frankia sp. strain CcI3”. *Applied and environmental microbiology* 78.2 (2012), pp. 575–580.

[95] Ferrer JL et al. “Structure and function of enzymes involved in the biosynthesis of phenylpropanoids”. *Plant Physiology and biochemistry: PPB / Societe francaise de physiologie vegetale* 46.3 (2008), pp. 356–370.

[96] Falcone Ferreyra ML et al. “Flavonoids: biosynthesis, biological functions, and biotechnological applications”. *Frontiers in plant science* 3 (2012), p. 222.

[97] Garg N and Singla P. “Stimulation of nitrogen fixation and trehalose biosynthesis by naringenin (Nar) and arbuscular mycorrhiza (AM) in chickpea under salinity stress”. *Plant Growth Regulation* 80.1 (2016), pp. 5–22.

[98] Weston LA and Mathesius U. “Flavonoids: their structure, biosynthesis and role in the rhizosphere, including allelopathy”. *Journal of chemical ecology* 39.2 (2013), pp. 283–297.

[99] Kikuchi K et al. “Flavonoids induce germination of basidiospores of the ectomycorrhizal fungus Suillus bovinus”. *Mycorrhiza* 17.7 (2007), pp. 563–570.

[100] Abdel-Lateif K et al. “Silencing of the chalcone synthase gene in C asuarina glauca highlights the important role of flavonoids during nodulation”. *The New Phytologist* 199.4 (2013), pp. 1012–1021.

[101] Bécard G et al. “Extensive In Vitro Hyphal Growth of Vesicular-Arbuscular Mycorrhizal Fungi in the Presence of CO2 and Flavonols”. *Applied and environmental microbiology* 58.3 (1992), pp. 821–825.

[102] Hartwig UA et al. “Flavonoids Released Naturally from Alfalfa Seeds Enhance Growth Rate of Rhizobium meliloti”. *Plant Physiology* 95.3 (1991), pp. 797–803.

[103] Sayed WF and Wheeler CT. “Effect of the flavonoid quercetin on culture and isolation ofFrankia fromCasuarina root nodules”. *Folia microbiologica* 44.1 (1999), p. 59.

[104] Mulligan JT and Long SR. “A family of activator genes regulates expression of Rhizobium meliloti nodulation genes”. *Genetics* 122.1 (1989), pp. 7–18.

[105] Chen XC et al. “Modulating DNA bending affects NodD-mediated transcriptional control in Rhizobium leguminosarum”. *Nucleic Acids Research* 33.8 (2005), pp. 2540–2548.

[106] Limpens E et al. “Lipochitooligosaccharides modulate plant host immunity to enable endosymbioses”. *Annual review of phytopathology* 53 (2015), pp. 311–334.

[107] Auguy F et al. “Activation of the isoflavonoid pathway in actinorhizal symbioses”. *Functional plant biology: FPB* 38.9 (2011), pp. 690–696.

[108] Lin K et al. “Single nucleus genome sequencing reveals high similarity among nuclei of an endomycorrhizal fungus”. *PLoS genetics* 10.1 (2014), e1004078.

[109] Maillet F et al. “Fungal lipochitooligosaccharide symbiotic signals in arbuscular mycorrhiza”. *Nature* 469.7328 (2011), pp. 58–63.

[110] Genre A et al. “Short-chain chitin oligomers from arbuscular mycorrhizal fungi trigger nuclear Ca2+ spiking in Medicago truncatula roots and their production is enhanced by strigolactone”. *New* (2013).

[111] Tisserant E et al. “Genome of an arbuscular mycorrhizal fungus provides insight into the oldest plant symbiosis”. *Proceedings of the National Academy of Sciences of the United States of America* 110.50 (2013), pp. 20117–20122.

[112] Persson T et al. “Candidatus Frankia Datiscae Dg1, the Actinobacterial Microsymbiont of Datisca glomerata, Expresses the Canonical nod Genes nodABC in Symbiosis with Its Host Plant”. *PLOS One* 10.5 (2015), e0127630.

[113] Reddy PM et al. “Flavonoids as signaling molecules and regulators of root nodule development”. *Dynamic soil, Dynamic plant* 1.2 (2007), pp. 83–94.

[114] Kosslak RM et al. “Induction of Bradyrhizobium japonicum common nod genes by isoflavones isolated from Glycine max”. *Proceedings of the National Academy of Sciences of the United States of America* 84.21 (1987), pp. 7428–7432.

[115] Zuanazzi JAS et al. “Production of Sinorhizobium meliloti nod Gene Activator and Repressor Flavonoids from Medicago sativa Roots”. *Molecular plant-microbe interactions: MPMI* 11.8 (1998), pp. 784–794.

[116] Peters NK and Long SR. “Alfalfa Root Exudates and Compounds which Promote or Inhibit Induction of Rhizobium meliloti Nodulation Genes”. *Plant Physiology* 88.2 (1988), pp. 396–400.

[117] Wasson AP et al. “Silencing the flavonoid pathway in Medicago truncatula inhibits root nodule formation and prevents auxin transport regulation by rhizobia”. *The Plant Cell* 18.7 (2006), pp. 1617–1629.

[118] Becard G et al. “Flavonoids are not necessary plant signal compounds in arbuscular mycorrhizal symbioses”. *MPMI-Molecular Plant Microbe Interactions* 8.2 (1995), pp. 252–258.

[119] Ellouze W et al. “Phytochemicals and spore germination: At the root of AMF host preference?” *Applied soil ecology: a section of Agriculture, Ecosystems & Environment* 60 (2012), pp. 98–104.

[120] Brown DE et al. “Flavonoids act as negative regulators of auxin transport in vivo in arabidopsis”. *Plant Physiology* 126.2 (2001), pp. 524–535.

[121] Mathesius U et al. “Auxin transport inhibition precedes root nodule formation in white clover roots and is regulated by flavonoids and derivatives of chitin oligosaccharides”. *The Plant Journal* 14.1 (1998), pp. 23–34.

[122] Deinum EE et al. “Modeling a cortical auxin maximum for nodulation: different signatures of potential strategies”. *Frontiers in plant science* 3 (2012), p. 96.

[123] Hirsch AM et al. “Early nodulin genes are induced in alfalfa root outgrowths elicited by auxin transport inhibitors”. *Proceedings of the National Academy of Sciences of the United States of America* 86.4 (1989), pp. 1244–1248.

[124] Ng JLP et al. “The Control of Auxin Transport in Parasitic and Symbiotic Root–Microbe Interactions”. *Plants* 4.3 (2015), pp. 606–643.

[125] Pandey A et al. “Corrigendum: Emerging Roles of Strigolactones in Plant Responses to Stress and Development”. *Frontiers in plant science* 7 (2016), p. 860.

[126] Brewer PB et al. “Diverse roles of strigolactones in plant development”. *Molecular plant* 6.1 (2013), pp. 18–28.

[127] Alder A et al. “The path from β-carotene to carlactone, a strigolactone-like plant hormone”. *Science* 335.6074 (2012), pp. 1348–1351.

[128] Zhang Y et al. “Rice cytochrome P450 MAX1 homologs catalyze distinct steps in strigolactone biosynthesis”. *Nature chemical biology* 10.12 (2014), pp. 1028–1033.

[129] Al-Babili S and Bouwmeester HJ. “Strigolactones, a novel carotenoid-derived plant hormone”. *Annual review of plant biology* 66 (2015), pp. 161–186.

[130] Hamiaux C et al. “DAD2 is an α/β hydrolase likely to be involved in the perception of the plant branching hormone, strigolactone”. *Current Biology* 22.21 (2012), pp. 2032–2036.

[131] Zhao LH et al. “Destabilization of strigolactone receptor DWARF14 by binding of ligand and E3-ligase signaling effector DWARF3”. *Cell research* 25.11 (2015), pp. 1219–1236.

[132] Akiyama K et al. “Structural requirements of strigolactones for hyphal branching in AM fungi”. *Plant & Cell Physiology* 51.7 (2010), pp. 1104–1117.

[133] Besserer A et al. “Strigolactones stimulate arbuscular mycorrhizal fungi by activating mitochondria”. *PLoS biology* 4.7 (2006), e226.

[134] Besserer A et al. “GR24, a synthetic analog of strigolactones, stimulates the mitosis and growth of the arbuscular mycorrhizal fungus Gigaspora rosea by boosting its energy metabolism”. *Plant Physiology* 148.1 (2008), pp. 402–413.

[135] Tsai SM and Phillips DA. “Flavonoids released naturally from alfalfa promote development of symbiotic glomus spores in vitro”. *Applied and environmental microbiology* 57.5 (1991), pp. 1485–1488.

[136] Tsuzuki S et al. “Strigolactone-Induced Putative Secreted Protein 1 Is Required for the Establishment of Symbiosis by the Arbuscular Mycorrhizal Fungus Rhizophagus irregularis”. *Molecular plant-microbe interactions: MPMI* 29.4 (2016), pp. 277–286.

[137] Steinkellner S et al. “Flavonoids and strigolactones in root exudates as signals in symbiotic and pathogenic plant-fungus interactions”. *Molecules* 12.7 (2007), pp. 1290–1306.

[138] Dor E et al. “The synthetic strigolactone GR24 influences the growth pattern of phytopathogenic fungi”. *Planta* 234.2 (2011), pp. 419–427.

[139] Peláez-Vico MA et al. “Strigolactones in the Rhizobium-legume symbiosis: Stimulatory effect on bacterial surface motility and down-regulation of their levels in nodulated plants”. *Plant science: an international journal of experimental plant biology* 245 (2016), pp. 119–127.

[140] De Cuyper C et al. “From lateral root density to nodule number, the strigolactone analogue GR24 shapes the root architecture of Medicago truncatula”. *Journal of Experimental Botany* 66.13 (2015), p. 4091.

[141] Zeijl A van et al. “The strigolactone biosynthesis gene DWARF27 is co-opted in rhizobium symbiosis”. *BMC plant biology* 15 (2015), p. 260.

[142] Gomez-Roldan V et al. “Strigolactone inhibition of shoot branching”. *Nature* 455.7210 (2008), pp. 189–194.

[143] Kohlen W et al. “The tomato CAROTENOID CLEAVAGE DIOXYGENASE 8 (SlCCD8) regulates rhizosphere signaling, plant architecture and affects reproductive development through strigolactone biosynthesis”. *The New Phytologist* 196.2 (2012), pp. 535–547.

[144] Kretzschmar T et al. “A petunia ABC protein controls strigolactone-dependent symbiotic signalling and branching”. *Nature* 483.7389 (2012), pp. 341–344.

[145] Liu J et al. “Carotenoid cleavage dioxygenase 7 modulates plant growth, reproduction, senescence, and determinate nodulation in the model legume Lotus japonicus”. *Journal of Experimental Botany* 64.7 (2013), pp. 1967–1981.

[146] Vogel JT et al. “SlCCD7 controls strigolactone biosynthesis, shoot branching and mycorrhiza-induced apocarotenoid formation in tomato”. *The Plant Journal* 61.2 (2010), pp. 300–311.

[147] Foo E et al. “Strigolactones and the regulation of pea symbioses in response to nitrate and phosphate deficiency”. *Molecular plant* 6.1 (2013), pp. 76–87.

[148] Liu W et al. “Strigolactone biosynthesis in Medicago truncatula and rice requires the symbiotic GRAS-type transcription factors NSP1 and NSP2”. *The Plant Cell* 23.10 (2011), pp. 3853–3865.

[149] Catoira R et al. “Four genes of Medicago truncatula controlling components of a nod factor transduction pathway”. *The Plant Cell* 12.9 (2000), pp. 1647–1666.

[150] Oldroyd GED and Long SR. “Identification and Characterization of Nodulation-Signaling Pathway 2, a Gene of Medicago truncatula Involved in Nod Factor Signaling”. *Plant Physiology* 131.3 (2003), pp. 1027–1032.

[151] Heckmann AB et al. “Lotus japonicus nodulation requires two GRAS domain regulators, one of which is functionally conserved in a non-legume”. *Plant Physiology* 142.4 (2006), pp. 1739–1750.

[152] Yoshida S et al. “The D3 F-box protein is a key component in host strigolactone responses essential for arbuscular mycorrhizal symbiosis”. *The New Phytologist* 196.4 (2012), pp. 1208–1216.

[153] Gutjahr C et al. “Rice perception of symbiotic arbuscular mycorrhizal fungi requires the karrikin receptor complex”. *Science* 350.6267 (2015), pp. 1521–1524.

[154] Scaffidi A et al. “Strigolactone Hormones and Their Stereoisomers Signal through Two Related Receptor Proteins to Induce Different Physiological Responses in Arabidopsis”. *Plant Physiology* 165.3 (2014), pp. 1221–1232.

[155] Garcia K et al. “Molecular signals required for the establishment and maintenance of ectomycorrhizal symbioses”. *The New Phytologist* 208.1 (2015), pp. 79–87.

[156] Capoen W et al. “Nuclear membranes control symbiotic calcium signaling of legumes”. *Proceedings of the National Academy of Sciences of the United States of America* 108.34 (2011), pp. 14348–14353.

[157] Charpentier M et al. “Nuclear-localized cyclic nucleotide–gated channels mediate symbiotic calcium oscillations”. *Science* (2016).

[158] Imaizumi-Anraku H et al. “Plastid proteins crucial for symbiotic fungal and bacterial entry into plant roots”. *Nature* 433.7025 (2005), pp. 527–531.

[159] Clavijo F et al. “The Casuarina NIN gene is transcriptionally activated throughout Frankia root infection as well as in response to bacterial diffusible signals”. *The New Phytologist* 208.3 (2015), pp. 887–903.

[160] Marsh JF et al. “Medicago truncatula NIN is essential for rhizobial-independent nodule organogenesis induced by autoactive calcium/calmodulin-dependent protein kinase”. *Plant Physiology* 144.1 (2007), pp. 324–335.

[161] Schauser L et al. “A plant regulator controlling development of symbiotic root nodules”. *Nature* 402.6758 (1999), pp. 191–195.

[162] Gobbato E et al. “A GRAS-type transcription factor with a specific function in mycorrhizal signaling”. *Current Biology* 22.23 (2012), pp. 2236–2241.

[163] Buist G et al. “LysM, a widely distributed protein motif for binding to (peptido)glycans”. *Molecular microbiology* 68.4 (2008), pp. 838–847.

[164] Antolín-Llovera M et al. “Cleavage of the SYMBIOSIS RECEPTOR-LIKE KINASE ectodomain promotes complex formation with Nod factor receptor 5”. *Current Biology* 24.4 (2014), pp. 422–427.

[165] Zhang X et al. “The receptor kinase CERK1 has dual functions in symbiosis and immunity signalling”. *The Plant Journal* 81.2 (2015), pp. 258–267.

[166] Czaja LF et al. “Transcriptional responses toward diffusible signals from symbiotic microbes reveal MtNFP-and MtDMI3-dependent reprogramming of host gene expression by arbuscular mycorrhizal fungal lipochitooligosaccharides”. *Plant Physiology* 159.4 (2012), pp. 1671–1685.

[167] De Mita S et al. “Evolution of a symbiotic receptor through gene duplications in the legume–rhizobium mutualism”. *The New Phytologist* 201.3 (2014), pp. 961–972.

[168] Rasmussen SR et al. “Intraradical colonization by arbuscular mycorrhizal fungi triggers induction of a lipochitooligosaccharide receptor”. *Scientific Reports* 6 (2016), p. 29733.

[169] Buendia L et al. “The LysM receptor-like kinase SlLYK10 regulates the arbuscular mycorrhizal symbiosis in tomato”. *The New Phytologist* 210.1 (2016), pp. 184–195.

[170] Miyata K et al. “The bifunctional plant receptor, OsCERK1, regulates both chitin-triggered immunity and arbuscular mycorrhizal symbiosis in rice”. *Plant & Cell Physiology* 55.11 (2014), pp. 1864–1872.

[171] Nakagawa T et al. “From defense to symbiosis: limited alterations in the kinase domain of LysM receptor-like kinases are crucial for evolution of legume–Rhizobium symbiosis”. *The Plant Journal* 65.2 (2011), pp. 169–180.

[172] Ben C et al. “Natural diversity in the model legume Medicago truncatula allows identifying distinct genetic mechanisms conferring partial resistance to Verticillium wilt”. *Journal of Experimental Botany* 64.1 (2013), pp. 317–332.

[173] Rey T et al. “NFP, a LysM protein controlling Nod factor perception, also intervenes in Medicago truncatula resistance to pathogens”. *The New Phytologist* 198.3 (2013), pp. 875–886.

[174] Rey T et al. “Medicago truncatula symbiosis mutants affected in the interaction with a biotrophic root pathogen”. *The New Phytologist* 206.2 (2015), pp. 497–500.

[175] Pietraszewska-Bogiel A et al. “Interaction of Medicago truncatula lysin motif receptor-like kinases, NFP and LYK3, produced in Nicotiana benthamiana induces defence-like responses”. *PLOS One* 8.6 (2013), e65055.

[176] Moling S et al. “Nod factor receptors form heteromeric complexes and are essential for intracellular infection in medicago nodules”. *The Plant Cell* 26.10 (2014), pp. 4188–4199.

[177] Haney CH and Long SR. “Plant flotillins are required for infection by nitrogen-fixing bacteria”. *Proceedings of the National Academy of Sciences of the United States of America* 107.1 (2010), pp. 478–483.

[178] Lefebvre B et al. “A remorin protein interacts with symbiotic receptors and regulates bacterial infection”. *Proceedings of the National Academy of Sciences of the United States of America* 107.5 (2010), pp. 2343–2348.

[179] Kaku H et al. “Plant cells recognize chitin fragments for defense signaling through a plasma membrane receptor”. *Proceedings of the National Academy of Sciences of the United States of America* 103.29 (2006), pp. 11086–11091.

[180] Hayafune M et al. “Chitin-induced activation of immune signaling by the rice receptor CEBiP relies on a unique sandwich-type dimerization”. *Proceedings of the National Academy of Sciences of the United States of America* 111.3 (2014), E404–13.

[181] Shimizu T et al. “Two LysM receptor molecules, CEBiP and OsCERK1, cooperatively regulate chitin elicitor signaling in rice”. *The Plant Journal* 64.2 (2010), pp. 204–214.

[182] Brewin NJ. “Plant Cell Wall Remodelling in the Rhizobium–Legume Symbiosis”. *Critical Reviews in Plant Sciences* 23.4 (2004), pp. 293–316.

[183] Tisa LS et al. “What stories can the Frankia genomes start to tell us?” *Journal of biosciences* 38.4 (2013), pp. 719–726.

[184] Fabre S et al. “Nod Factor-Independent Nodulation in Aeschynomene evenia Required the Common Plant-Microbe Symbiotic Toolkit”. *Plant Physiology* 169.4 (2015), pp. 2654–2664.

[185] Giraud E et al. “Legumes symbioses: absence of Nod genes in photosynthetic bradyrhizobia”. *Science* 316.5829 (2007), pp. 1307–1312.

[186] Chabaud M et al. “Chitinase-resistant hydrophilic symbiotic factors secreted by Frankia activate both Ca2+ spiking and NIN gene expression in the actinorhizal plant Casuarina glauca”. *The New Phytologist* 209.1 (2015), pp. 86–93.

[187] Franche C et al. “New insights in the molecular events underlying actinorhizal nodulation in the tropical tree Casuarina glauca”. *BMC proceedings* 5.7 (2011), O33.

[188] Granqvist E et al. “Bacterial-induced calcium oscillations are common to nitrogen-fixing associations of nodulating legumes and non-legumes”. *The New Phytologist* 207.3 (2015), pp. 551–558.

[189] Okazaki S et al. “Rhizobium–legume symbiosis in the absence of Nod factors: two possible scenarios with or without the T3SS”. *The ISME journal* 10.1 (2016), pp. 64–74.

[190] Okazaki S et al. “Hijacking of leguminous nodulation signaling by the rhizobial type III secretion system”. *Proceedings of the National Academy of Sciences of the United States of America* 110.42 (2013), pp. 17131–17136.

[191] Pieterse CMJ et al. “Hormonal modulation of plant immunity”. *Annual review of cell and developmental biology* 28 (2012), pp. 489–521.

[192] Martinez-Abarca F et al. “Involvement of salicylic acid in the establishment of the Rhizobium meliloti-alfalfa symbiosis”. *Molecular plant-microbe interactions: MPMI* 11.2 (1998), pp. 153–155.

[193] Blilou I et al. “Resistance of pea roots to endomycorrhizal fungus or Rhizobium correlates with enhanced levels of endogenous salicylic acid”. *Journal of Experimental Botany* 50.340 (1999), pp. 1663–1668.

[194] Liang Y et al. “Nonlegumes respond to rhizobial Nod factors by suppressing the innate immune response”. *Science* 341.6152 (2013), pp. 1384–1387.

[195] Tanaka K et al. “Effect of lipo-chitooligosaccharide on early growth of C4 grass seedlings”. *Journal of Experimental Botany* 66.19 (2015), pp. 5727–5738.

[196] Navarro L et al. “DELLAs control plant immune responses by modulating the balance of jasmonic acid and salicylic acid signaling”. *Current Biology* 18.9 (2008), pp. 650–655.

[197] Hou X et al. “DELLAs modulate jasmonate signaling via competitive binding to JAZs”. *Developmental cell* 19.6 (2010), pp. 884–894.

[198] Boter M et al. “Conserved MYC transcription factors play a key role in jasmonate signaling both in tomato and Arabidopsis”. *Genes & development* 18.13 (2004), pp. 1577–1591.

[199] Wild M et al. “The Arabidopsis DELLA RGA-LIKE3 is a direct target of MYC2 and modulates jasmonate signaling responses”. *The Plant Cell* 24.8 (2012), pp. 3307–3319.

[200] Yang DL et al. “Plant hormone jasmonate prioritizes defense over growth by interfering with gibberellin signaling cascade”. *Proceedings of the National Academy of Sciences of the United States of America* 109.19 (2012), E1192–200.

[201] Floss DS et al. “DELLA proteins regulate arbuscule formation in arbuscular mycorrhizal symbiosis”. *Proceedings of the National Academy of Sciences of the United States of America* 110.51 (2013), E5025–34.

[202] Fonouni-Farde C et al. “DELLA-mediated gibberellin signalling regulates Nod factor signalling and rhizobial infection”. *Nature Communications* 7 (2016), p. 12636.

[203] Pimprikar P et al. “A CCaMK-CYCLOPS-DELLA Complex Activates Transcription of RAM1 to Regulate Arbuscule Branching”. *Current Biology* 26.8 (2016), pp. 987–998.

[204] Yu N et al. “A DELLA protein complex controls the arbuscular mycorrhizal symbiosis in plants”. *Cell research* 24.1 (2014), pp. 130–133.

[205] Jin Y et al. “DELLA proteins are common components of symbiotic rhizobial and mycorrhizal signalling pathways”. *Nature Communications* 7 (2016), p. 12433.

[206] Martin F et al. “The genome of Laccaria bicolor provides insights into mycorrhizal symbiosis”. *Nature* 452.7183 (2008), pp. 88–92.

[207] Kloppholz S et al. “A secreted fungal effector of Glomus intraradices promotes symbiotic biotrophy”. *Current Biology* 21.14 (2011), pp. 1204–1209.

[208] Economou A et al. “The Rhizobium nodulation gene nodO encodes a Ca2 (+)-binding protein that is exported without N-terminal cleavage and is homologous to haemolysin and related proteins”. *The EMBO journal* 9.2 (1990), pp. 349–354.

[209] Sutton JM et al. “The nodulation-signaling protein NodO from Rhizobium leguminosarum biovar viciae forms ion channels in membranes”. *Proceedings of the National Academy of Sciences of the United States of America* 91.21 (1994), pp. 9990–9994.

[210] Plett JM et al. “Effector MiSSP7 of the mutualistic fungus Laccaria bicolor stabilizes the Populus JAZ6 protein and represses jasmonic acid (JA) responsive genes”. *Proceedings of the National Academy of Sciences of the United States of America* 111.22 (2014), pp. 8299–8304.

[211] Sun J et al. “Activation of symbiosis signaling by arbuscular mycorrhizal fungi in legumes and rice”. *The Plant Cell* 27.3 (2015), pp. 823–838.

[212] Cesco S et al. “Release of plant-borne flavonoids into the rhizosphere and their role in plant nutrition”. *Plant and soil* 329.1 (2010), pp. 1–25.

[213] Velzen R van et al. “Comparative genomics of the nonlegume Parasponia reveals insights into evolution of nitrogen-fixing rhizobium symbioses”. *Proceedings of the National Academy of Sciences of the United States of America* 115.20 (2018), E4700–E4709.

[214] Ott T et al. “Symbiotic leghemoglobins are crucial for nitrogen fixation in legume root nodules but not for general plant growth and development”. *Current Biology* 15.6 (2005), pp. 531–535.

[215] Stokstad E. “The nitrogen fix”. *Science* 353.6305 (2016), pp. 1225–1227.

[216] Doyle JJ. “Chasing unicorns: Nodulation origins and the paradox of novelty”. *American Journal of Botany* 103.11 (2016), pp. 1865–1868.

[217] Martin FM et al. “Ancestral alliances: Plant mutualistic symbioses with fungi and bacteria”. *Science* 356.6340 (2017).

[218] Vernié T et al. “The NIN Transcription Factor Coordinates Diverse Nodulation Programs in Different Tissues of the Medicago truncatula Root”. *The Plant Cell* 27.12 (2015), pp. 3410–3424.

[219] Clason EW. *The Vegetation of the Upper-Badak Region of Mount Kelut [East Java]*. 1935.

[220] Trinick MJ. “Symbiosis between Rhizobium and the Non-legume, Trema aspera”. *Nature* 244.5416 (1973), pp. 459–460.

[221] Akkermans ADL et al. “Nitrogen-fixing root nodules in Ulmaceae”. *Nature* 274.5667 (1978), pp. 190–190.

[222] Becking JH. “The Rhizobium symbiosis of the nonlegume Parasponia”. *Biological nitrogen fixation* (1992), pp. 497–559.

[223] Marvel DJ et al. “Rhizobium symbiotic genes required for nodulation of legume and nonlegume hosts”. *Proceedings of the National Academy of Sciences of the United States of America* 84.5 (1987), pp. 1319–1323.

[224] Yang MQ et al. “Molecular phylogenetics and character evolution of Cannabaceae”. *Taxon* 62.3 (2013), pp. 473–485.

[225] Op den Camp RHM et al. “Nonlegume Parasponia andersonii deploys a broad rhizobium host range strategy resulting in largely variable symbiotic effectiveness”. *Molecular plant-microbe interactions: MPMI* 25.7 (2012), pp. 954–963.

[226] Roux B et al. “An integrated analysis of plant and bacterial gene expression in symbiotic root nodules using laser-capture microdissection coupled to RNA sequencing”. *The Plant Journal* 77.6 (2014), pp. 817–837.

[227] Combier JP et al. “MtHAP2-1 is a key transcriptional regulator of symbiotic nodule development regulated by microRNA169 in Medicago truncatula”. *Genes & development* 20.22 (2006), pp. 3084–3088.

[228] Baudin M et al. “A Phylogenetically Conserved Group of Nuclear Factor-Y Transcription Factors Interact to Control Nodulation in Legumes”. *Plant Physiology* 169.4 (2015), pp. 2761–2773.

[229] Arrighi JF et al. “The RPG gene of Medicago truncatula controls Rhizobium-directed polar growth during infection”. *Proceedings of the National Academy of Sciences* 105.28 (2008), pp. 9817–9822.

[230] Kistner C et al. “Seven Lotus japonicus genes required for transcriptional reprogramming of the root during fungal and bacterial symbiosis”. *The Plant Cell* 17.8 (2005), pp. 2217–2229.

[231] Deguchi Y et al. “Transcriptome profiling of Lotus japonicus roots during arbuscular mycorrhiza development and comparison with that of nodulation”. *DNA research: an international journal for rapid publication of reports on genes and genomes* 14.3 (2007), pp. 117–133.

[232] Pumplin N et al. “Medicago truncatula Vapyrin is a novel protein required for arbuscular mycorrhizal symbiosis”. *The Plant Journal* 61.3 (2010), pp. 482–494.

[233] Horváth B et al. “Medicago truncatula IPD3 is a member of the common symbiotic signaling pathway required for rhizobial and mycorrhizal symbioses”. *Molecular plant-microbe interactions: MPMI* 24.11 (2011), pp. 1345–1358.

[234] Murray JD et al. “Vapyrin, a gene essential for intracellular progression of arbuscular mycorrhizal symbiosis, is also essential for infection by rhizobia in the nodule symbiosis of Medicago truncatula”. *The Plant Journal* 65.2 (2011), pp. 244–252.

[235] Tóth K et al. “Functional domain analysis of the Remorin protein LjSYMREM1 in Lotus japonicus”. *PLOS One* 7.1 (2012), e30817.

[236] Chiasson DM et al. “Soybean SAT1 (Symbiotic Ammonium Transporter 1) encodes a bHLH transcription factor involved in nodule growth and NH4+ transport”. *Proceedings of the National Academy of Sciences of the United States of America* 111.13 (2014), pp. 4814–4819.

[237] Afkhami ME and Stinchcombe JR. “Multiple mutualist effects on genomewide expression in the tripartite association between Medicago truncatula, nitrogen-fixing bacteria and mycorrhizal fungi”. *Molecular ecology* 25.19 (2016), pp. 4946–4962.

[238] Sturms R et al. “Trema and parasponia hemoglobins reveal convergent evolution of oxygen transport in plants”. *Biochemistry* 49.19 (2010), pp. 4085–4093.

[239] Kakar S et al. “Crystal structures of Parasponia and Trema hemoglobins: differential heme coordination is linked to quaternary structure”. *Biochemistry* 50.20 (2011), pp. 4273–4280.

[240] Żmieńko A et al. “Copy number polymorphism in plant genomes”. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik* 127.1 (2014), pp. 1–18.

[241] Shadle G et al. “Down-regulation of hydroxycinnamoyl CoA: shikimate hydroxycinnamoyl transferase in transgenic alfalfa affects lignification, development and forage quality”. *Phytochemistry* 68.11 (2007), pp. 1521–1529.

[242] Gallego-Giraldo L et al. “Lignin modification leads to increased nodule numbers in alfalfa”. *Plant Physiology* 164.3 (2014), pp. 1139–1150.

[243] Kawaharada Y et al. “Receptor-mediated exopolysaccharide perception controls bacterial infection”. *Nature* 523.7560 (2015), pp. 308–312.

[244] Kawaharada Y et al. “Differential regulation of the Epr3 receptor coordinates membrane-restricted rhizobial colonization of root nodule primordia”. *Nature Communications* 8 (2017), p. 14534.

[245] Borisov AY et al. “The Sym35 gene required for root nodule development in pea is an ortholog of Nin from Lotus japonicus”. *Plant Physiology* 131.3 (2003), pp. 1009–1017.

[246] Natsume S et al. “The Draft Genome of Hop (Humulus lupulus), an Essence for Brewing”. *Plant & Cell Physiology* 56.3 (2014), pp. 428–441.

[247] He N et al. “Draft genome sequence of the mulberry tree Morus notabilis”. *Nature Communications* 4 (2013), p. 2445.

[248] Huang J et al. “The Jujube Genome Provides Insights into Genome Evolution and the Domestication of Sweetness/Acidity Taste in Fruit Trees”. *PLoS genetics* 12.12 (2016), e1006433.

[249] Shulaev V et al. “The genome of woodland strawberry (Fragaria vesca)”. *Nature Genetics* 43.2 (2011), pp. 109–116.

[250] Herendeen PS et al. “A preliminary conspectus of the Allon flora from the Late Cretaceous (Late Santonian) of central Georgia, USA”. *Annals of the Missouri Botanical Garden. Missouri Botanical Garden* (1999), pp. 407–471.

[251] Bruneau A et al. “Phylogenetic patterns and diversification in the caesalpinioid legumes”. *Botany* 86.7 (2008), pp. 697–718.

[252] Nguyen TV et al. “An assemblage of Frankia Cluster II strains from California contains the canonical nod genes and also the sulfotransferase gene nodH”. *BMC genomics* 17.1 (2016), p. 796.

[253] Delaux PM et al. “Algal ancestor of land plants was preadapted for symbiosis”. *Proceedings of the National Academy of Sciences of the United States of America* 112.43 (2015), pp. 13390–13395.

[254] Kamel L et al. “Biology and evolution of arbuscular mycorrhizal symbiosis in the light of genomics”. *The New Phytologist* 213.2 (2017), pp. 531–536.

[255] Winship LJ et al. “The acetylene reduction assay inactivates root nodule uptake hydrogenase in some actinorhizal plants”. *Physiologia Plantarum* 70.2 (1987), pp. 361–366.

[256] Silvester WB. “Oxygen regulation and hemoglobin”. *The Biology of Frankia and Actinorhizal Plants* (1990).

[257] Silvester WB et al. “Oxygen Responses, Hemoglobin, And The Structure And Function Of Vesicles”. *Nitrogen-fixing Actinorhizal Symbioses*. Ed. by Pawlowski K and Newton WE. Dordrecht: Springer Netherlands, 2008, pp. 105–146.

[258] Silvester WB and Winship LJ. “Transient responses of nitrogenase to acetylene and oxygen in actinorhizal nodules and cultured frankia”. *Plant Physiology* 92.2 (1990), pp. 480–486.

[259] Cao Q et al. “Efficiency of Agrobacterium rhizogenes–mediated root transformation of Parasponia and Trema is temperature dependent”. *Plant Growth Regulation* 68.3 (2012), pp. 459–465.

[260] Davey MR et al. “Effective Nodulation of Micro-Propagated Shoots of the Non-Legume Parasponia andersonii by Bradyrhizobium”. *Journal of Experimental Botany* 44.5 (1993), pp. 863–867.

[261] Geurts R and Jong H de. “Fluorescent In Situ Hybridization (FISH) on pachytene chromosomes as a tool for genome characterization”. *Methods in Molecular Biology* 1069 (2013), pp. 15–24.

[262] Gnerre S et al. “High-quality draft assemblies of mammalian genomes from massively parallel sequence data”. *Proceedings of the National Academy of Sciences of the United States of America* 108.4 (2011), pp. 1513–1518.

[263] Boetzer M et al. “Scaffolding pre-assembled contigs using SSPACE”. *Bioinformatics* 27.4 (2011), pp. 578–579.

[264] Parra G et al. “CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes”. *Bioinformatics* 23.9 (2007), pp. 1061–1067.

[265] Simão FA et al. “BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs”. *Bioinformatics* 31.19 (2015), pp. 3210–3212.

[266] Bao W et al. “Repbase Update, a database of repetitive elements in eukaryotic genomes”. *Mobile DNA* 6 (2015), p. 11.

[267] Gremme G et al. “GenomeTools: a comprehensive software library for efficient processing of structured genome annotations”. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 10.3 (2013), pp. 645–656.

[268] Han Y and Wessler SR. “MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences”. *Nucleic Acids Research* 38.22 (2010), e199.

[269] Grabherr MG et al. “Full-length transcriptome assembly from RNA-Seq data without a reference genome”. *Nature Biotechnology* 29.7 (2011), pp. 644–652.

[270] Trapnell C et al. “Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation”. *Nature Biotechnology* 28.5 (2010), pp. 511–515.

[271] Haas BJ et al. “De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis”. *Nature Protocols* 8.8 (2013), pp. 1494–1512.

[272] UniProt Consortium. “UniProt: a hub for protein information”. *Nucleic Acids Research* 43.Database issue (2015), pp. D204–12.

[273] Korf I. “Gene finding in novel genomes”. *BMC Bioinformatics* 5 (2004), p. 59.

[274] Stanke M et al. “Using native and syntenically mapped cDNA alignments to improve de novo gene finding”. *Bioinformatics* 24.5 (2008), pp. 637–644.

[275] Hoff KJ et al. “BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS”. *Bioinformatics* 32.5 (2016), pp. 767–769.

[276] Lomsadze A et al. “Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm”. *Nucleic Acids Research* 42.15 (2014), e119.

[277] Campbell MS et al. “MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations”. *Plant Physiology* 164.2 (2014), pp. 513–524.

[278] Haas BJ et al. “Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments”. *Genome Biology* 9.1 (2008), R7.

[279] Kurtz S et al. “Versatile and open software for comparing large genomes”. *Genome Biology* 5.2 (2004), R12.

[280] Otto TD et al. “RATT: Rapid Annotation Transfer Tool”. *Nucleic Acids Research* 39.9 (2011), e57.

[281] Altschul SF et al. “Basic local alignment search tool”. *Journal of molecular biology* 215.3 (1990), pp. 403–410.

[282] Jones P et al. “InterProScan 5: genome-scale protein function classification”. *Bioinformatics* 30.9 (2014), pp. 1236–1240.

[283] Gene Ontology Consortium. “Creating the gene ontology resource: design and implementation”. *Genome Research* 11.8 (2001), pp. 1425–1433.

[284] Kanehisa M and Goto S. “KEGG: kyoto encyclopedia of genes and genomes”. *Nucleic Acids Research* 28.1 (2000), pp. 27–30.

[285] Emms DM and Kelly S. “OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy”. *Genome Biology* 16 (2015), p. 157.

[286] Swarbreck D et al. “The Arabidopsis Information Resource (TAIR): gene structure and function annotation”. *Nucleic Acids Research* 36.Database issue (2008), pp. D1009–14.

[287] Myburg AA et al. “The genome of Eucalyptus grandis”. *Nature* 510.7505 (2014), pp. 356–362.

[288] Tuskan GA et al. “The genome of black cottonwood, Populus trichocarpa (Torr. & Gray)”. *Science* 313.5793 (2006), pp. 1596–1604.

[289] Goodstein DM et al. “Phytozome: a comparative platform for green plant genomics”. *Nucleic Acids Research* 40.Database issue (2012), pp. D1178–86.

[290] Saitou N and Nei M. “The neighbor-joining method: a new method for reconstructing phylogenetic trees”. *Molecular biology and evolution* 4.4 (1987), pp. 406–425.

[291] Valdar WSJ. “Scoring residue conservation”. *Proteins* 48.2 (2002), pp. 227–241.

[292] Kim D et al. “HISAT: a fast spliced aligner with low memory requirements”. *Nature Methods* 12.4 (2015), pp. 357–360.

[293] Liao Y et al. “featureCounts: an efficient general purpose program for assigning sequence reads to genomic features”. *Bioinformatics* 30.7 (2014), pp. 923–930.

[294] Banba M et al. “Divergence of evolutionary ways among common sym genes: CASTOR and CCaMK show functional conservation between two symbiosis systems and constitute the root of a common signaling pathway”. *Plant & Cell Physiology* 49.11 (2008), pp. 1659–1671.

[295] Griesmann M et al. “Phylogenomics reveals multiple losses of nitrogen-fixing root nodule symbiosis”. *Science* 361.6398 (2018).

[296] Amor BB et al. “The NFP locus of Medicago truncatula controls an early step of Nod factor signal transduction upstream of a rapid calcium flux and root hair deformation”. *The Plant Journal* 34.4 (2003), pp. 495–506.

[297] Smit P et al. “NSP1 of the GRAS protein family is essential for rhizobial Nod factor-induced transcription”. *Science* 308.5729 (2005), pp. 1789–1791.

[298] Cerri MR et al. “Medicago truncatula ERN transcription factors: regulatory interplay with NSP1/NSP2 GRAS factors and expression dynamics throughout rhizobial infection”. *Plant Physiology* 160.4 (2012), pp. 2155–2172.

[299] Battaglia M et al. “A nuclear factor Y interacting protein of the GRAS family is required for nodule organogenesis, infection thread progression, and lateral root growth”. *Plant Physiology* 164.3 (2014), pp. 1430–1442.

[300] Campbell N. “Rewiring the circuitry”. *Nature Reviews. Genetics* 5.1 (2004), pp. 7–7.

[301] Shubin N et al. “Deep homology and the origins of evolutionary novelty”. *Nature* 457.7231 (2009), pp. 818–823.

[302] Marbach D et al. “Wisdom of crowds for robust gene network inference”. *Nature Methods* 9.8 (2012), pp. 796–804.

[303] Stolovitzky G et al. “Dialogue on reverse-engineering assessment and methods: the DREAM of high-throughput pathway inference”. *Annals of the New York Academy of Sciences* 1115 (2007), pp. 1–22.

[304] De Smet R and Marchal K. “Advantages and limitations of current network inference methods”. *Nature reviews. Microbiology* 8.10 (2010), pp. 717–729.

[305] Obayashi T et al. “ATTED-II in 2018: A Plant Coexpression Database Based on Investigation of the Statistical Property of the Mutual Rank Index”. *Plant & Cell Physiology* 59.1 (2018), e3.

[306] Kulkarni SR et al. “TF2Network: predicting transcription factor regulators and gene regulatory networks in Arabidopsis using publicly available binding site information”. *Nucleic Acids Research* 46.6 (2018), e31.

[307] Kaufmann K and Mueller-Roeber B. *Plant Gene Regulatory Networks: Methods and Protocols*. Ed. by Kaufmann K and Mueller-Roeber B. Humana Press, New York, NY, 2017.

[308] Berger B et al. “Computational solutions for omics data”. *Nature Reviews. Genetics* 14.5 (2013), pp. 333–346.

[309] Moerman T et al. “GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks”. *Bioinformatics* (2018).

[310] Verdier J et al. “A regulatory network-based approach dissects late maturation processes related to the acquisition of desiccation tolerance and longevity of Medicago truncatula ...” *Plant* (2013).

[311] Burks DJ and Azad RK. “Identification and Network-Enabled Characterization of Auxin Response Factor Genes in Medicago truncatula”. *Frontiers in plant science* 7 (2016), p. 1857.

[312] Guttikonda SK et al. “Whole genome co-expression analysis of soybean cytochrome P450 genes identifies nodulation-specific P450 monooxygenases”. *BMC plant biology* 10 (2010), p. 243.

[313] Jin J et al. “An Arabidopsis Transcriptional Regulatory Map Reveals Distinct Functional and Evolutionary Features of Novel Transcription Factors”. *Molecular biology and evolution* 32.7 (2015), pp. 1767–1773.

[314] Cerri MR et al. “The ERN1 transcription factor gene is a target of the CCaMK/CYCLOPS complex and controls rhizobial infection in Lotus japonicus”. *The New Phytologist* 215.1 (2017), pp. 323–337.

[315] Ariel F et al. “Two direct targets of cytokinin signaling regulate symbiotic nodulation in Medicago truncatula”. *The Plant Cell* 24.9 (2012), pp. 3838–3852.

[316] Hirsch S et al. “GRAS proteins form a DNA binding complex to induce gene expression during nodulation signaling in Medicago truncatula”. *The Plant Cell* 21.2 (2009), pp. 545–557.

[317] Kang H et al. “A MYB coiled-coil transcription factor interacts with NSP 2 and is involved in nodulation in L otus japonicus”. *The New Phytologist* 201.3 (2014), pp. 837–849.

[318] McInnes L et al. “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction” (2018). arXiv: `1802.03426 [stat.ML]`.

[319] Haque S et al. “Computational prediction of gene regulatory networks in plant growth and development”. *Current Opinion in Plant Biology* 47 (2019), pp. 96–105.

[320] Brooks MD et al. “Network Walking charts transcriptional dynamics of nitrogen signaling by integrating validated and predicted genome-wide interactions”. *Nature Communications* 10.1 (2019), p. 1569.

[321] Fiers MWEJ et al. “Mapping gene regulatory networks from single-cell omics data”. *Briefings in functional genomics* 17.4 (2018), pp. 246–254.

[322] Krouk G et al. “Gene regulatory networks in plants: learning causality from time and perturbation”. *Genome Biology* 14.6 (2013), p. 123.

[323] Bartel DP. “MicroRNAs: Genomics, Biogenesis, Mechanism, and Function”. *Cell* 116.2 (2004), pp. 281–297.

[324] Whitmarsh AJ and Davis RJ. “Regulation of transcription factor function by phosphorylation”. *Cellular and molecular life sciences: CMLS* 57.8 (2000), pp. 1172–1183.

[325] Song L and Crawford GE. “DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells”. *Cold Spring Harbor protocols* 2010.2 (2010), db.prot5384.

[326] Bartlett A et al. “Mapping genome-wide transcription-factor binding sites using DAP-seq”. *Nature Protocols* 12.8 (2017), pp. 1659–1672.

[327] Aibar S et al. “SCENIC: single-cell regulatory network inference and clustering”. *Nature Methods* 14.11 (2017), pp. 1083–1086.

[328] Shulse CN et al. “High-Throughput Single-Cell Transcriptome Profiling of Plant Cell Types”. 2019.

[329] Bray NL et al. “Near-optimal probabilistic RNA-seq quantification”. *Nature Biotechnology* 34.8 (2016), p. 888.

[330] Köster J and Rahmann S. “Snakemake – a scalable bioinformatics workflow engine”. *Bioinformatics* 28.19 (2012), pp. 2520–2522.

[331] Buchfink B et al. “Fast and sensitive protein alignment using DIAMOND”. *Nature Methods* 12.1 (2015), pp. 59–60.

[332] Dongen S van. “A cluster algorithm for graphs”. *Information Systems* R 0010 (2000).

[333] Jin J et al. “PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants”. *Nucleic Acids Research* 45.D1 (2017), pp. D1040–D1045.

[334] Holmer R et al. “GeneNoteBook, a collaborative notebook for comparative genomics”. *Bioinformatics* 35 (2019), pp. 4779–4781.

[335] Skinner ME et al. “JBrowse: a next-generation genome browser”. *Genome Research* 19.9 (2009), pp. 1630–1638.

[336] Buels R et al. “JBrowse: a dynamic web platform for genome visualization and analysis”. *Genome Biology* 17 (2016), p. 66.

[337] Lee E et al. “Web Apollo: a web-based genomic annotation editing platform”. *Genome Biology* 14.8 (2013), R93.

[338] Smith RN et al. “InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data”. *Bioinformatics* 28.23 (2012), pp. 3163–3165.

[339] Krishnakumar V et al. “Araport: the Arabidopsis information portal”. *Nucleic Acids Research* 43.Database issue (2015), pp. D1003–9.

[340] Gonzales MD et al. “The Legume Information System (LIS): an integrated information resource for comparative legume biology”. *Nucleic Acids Research* 33.Database issue (2005), pp. D660–5.

[341] Stein L et al. “WormBase: network access to the genome and biology of Caenorhabditis elegans”. *Nucleic Acids Research* 29.1 (2001), pp. 82–86.

[342] Cheng CY et al. “Araport11: a complete reannotation of the Arabidopsis thaliana reference genome”. *The Plant Journal* 89.4 (2017), pp. 789–804.

[343] Velzen R van et al. “A Resurrected Scenario: Single Gain and Massive Loss of Nitrogen-Fixing Nodulation”. *Trends in Plant Science* 24.1 (2019), pp. 49–57.

[344] Katoh K and Standley DM. “MAFFT multiple sequence alignment software version 7: improvements in performance and usability”. *Molecular Biology and Evolution* 30.4 (2013), pp. 772–780.

[345] Grüning B et al. “Bioconda: sustainable and comprehensive software distribution for the life sciences”. *Nature Methods* 15.7 (2018), pp. 475–476.

[346] Dupin SE et al. “The Non-Legume Parasponia andersonii Mediates the Fitness of Nitrogen-Fixing Rhizobial Symbionts Under High Nitrogen Conditions”. *Frontiers in Plant Science* 10 (2019), p. 1779.

[347] Chomicki G et al. “The Evolution of Mutualistic Dependence”. *Annual Review of Ecology, Evolution, and Systematics* 51.1 (2020), pp. 409–432.

[348] Werner GDA et al. “Symbiont switching and alternative resource acquisition strategies drive mutualism breakdown”. *Proceedings of the National Academy of Sciences of the United States of America* 115.20 (2018), pp. 5229–5234.

[349] Shen D et al. “A Homeotic Mutation Changes Legume Nodule Ontogeny into Actinorhizal-Type Ontogeny”. *The Plant Cell* 32.6 (2020), pp. 1868–1885.

[350] Parniske M. “Uptake of bacteria into living plant cells, the unifying and distinct feature of the nitrogen-fixing root nodule symbiosis”. *Current Opinion in Plant Biology* 44 (2018), pp. 164–174.

[351] Parniske M. “Intracellular accommodation of microbes by plants: a common developmental program for symbiosis and disease?” *Current Opinion in Plant Biology* 3.4 (2000), pp. 320–328.

[352] Kvon EZ et al. “Progressive Loss of Function in a Limb Enhancer during Snake Evolution”. *Cell* 167.3 (2016), 633–642.e11.

[353] Albalat R and Cañestro C. “Evolution by gene loss”. *Nature Reviews. Genetics* 17.7 (2016), pp. 379–391.

[354] Rutten L et al. “Duplication of Symbiotic Lysin Motif Receptors Predates the Evolution of Nitrogen-Fixing Nodule Symbiosis”. *Plant Physiology* 184.2 (2020), pp. 1004–1023.

[355] Buendia L et al. “LysM receptor-like kinase and LysM receptor-like protein families: An update on phylogeny and functional characterization”. *Frontiers in Plant Science* 9 (2018), p. 1531.

[356] Gibelin-Viala C et al. “The Medicago truncatula LysM receptor-like kinase LYK9 plays a dual role in immunity and the arbuscular mycorrhizal symbiosis”. *The New Phytologist* 223.3 (2019), pp. 1516–1529.

[357] Bu F et al. “Mutant analysis in the nonlegume Parasponia andersonii identifies NIN and NF-YA1 transcription factors as a core genetic network in nitrogen-fixing nodule symbioses”. *The New Phytologist* 226.2 (2020), pp. 541–554.

[358] Kumar A et al. “Nodule Inception Is Not Required for Arbuscular Mycorrhizal Colonization of Medicago truncatula”. *Plants* 9.1 (2020).

[359] Rich-Griffin C et al. “Single-Cell Transcriptomics: A High-Resolution Avenue for Plant Functional Genomics”. *Trends in Plant Science* 25.2 (2020), pp. 186–197.

[360] West MAL et al. “Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in Arabidopsis”. *Genetics* 175.3 (2007), pp. 1441–1450.

[361] Nijveen H et al. “AraQTL - workbench and archive for systems genetics in Arabidopsis thaliana”. *The Plant Journal* 89.6 (2017), pp. 1225–1235.

[362] Zou F et al. “Brain Expression Genome-Wide Association Study (eGWAS) Identifies Human Disease-Associated Variants”. *PLoS Genetics* 8.6 (2012), e1002707.

[363] Park PJ. “ChIP–seq: advantages and challenges of a maturing technology”. *Nature Reviews. Genetics* 10.10 (2009), pp. 669–680.

[364] Soyano T et al. “A shared gene drives lateral root development and root nodule symbiosis pathways in Lotus”. *Science* 366.6468 (2019), pp. 1021–1023.

[365] Schiessl K et al. “NODULE INCEPTION Recruits the Lateral Root Developmental Program for Symbiotic Nodule Organogenesis in Medicago truncatula”. *Current Biology* 29.21 (2019), 3657–3668.e5.

[366] Liu J and Bisseling T. “Evolution of NIN and NIN-like Genes in Relation to Nodule Symbiosis”. *Genes* 11.7 (2020).

[367] Lin JS et al. “Author Correction: NIN interacts with NLPs to mediate nitrate inhibition of nodulation in Medicago truncatula”. *Nature Plants* 4.12 (2018), p. 1125.

[368] Konishi M and Yanagisawa S. “Arabidopsis NIN-like transcription factors have a central role in nitrate signalling”. *Nature Communications* 4.1 (2013), p. 1617.

[369] Suzuki W et al. “The evolutionary events necessary for the emergence of symbiotic nitrogen fixation in legumes may involve a loss of nitrate responsiveness of the NIN transcription factor”. *Plant Signaling & Behavior* 8.10 (2013).

[370] Quilbé J et al. “Genetics of nodulation in Aeschynomene evenia uncovers mechanisms of the rhizobium-legume symbiosis”. *Nature Communications* 12.1 (2021), p. 829.

[371] Chandler MR. “Some Observations on Infection of Arachis hypogaea L. by Rhizobium”. *Journal of Experimental Botany* 29.3 (1978), pp. 749–755.

[372] Billault-Penneteau B et al. “Dryas as a Model for Studying the Root Symbioses of the Rosaceae”. *Frontiers in plant science* 10 (2019), p. 661.

[373] Sohn JI and Nam JW. “The present and future of de novo whole-genome assembly”. *Briefings in Bioinformatics* 19.1 (2018), pp. 23–40.

[374] Ruan J and Li H. “Fast and accurate long-read assembly with wtdbg2”. *Nature Methods* 17.2 (2020), pp. 155–158.

[375] König S et al. “Simultaneous gene finding in multiple genomes”. *Bioinformatics* 32.22 (2016), pp. 3388–3395.

[376] Dunne MP and Kelly S. “Erratum to: OrthoFiller: utilising data from multiple species to improve the completeness of genome annotations”. *BMC Genomics* 18.1 (2017), p. 488.

[377] Shen D et al. “The BOP-type co-transcriptional regulator NODULE ROOT1 promotes stem secondary growth of the tropical Cannabaceae tree Parasponia andersonii”. *The Plant Journal* tpj.15242 (2021).

[378] Lähnemann D et al. “Eleven grand challenges in single-cell data science”. *Genome Biology* 21.1 (2020), pp. 1–35.

[379] Zeng T et al. “Host- and stage-dependent secretome of the arbuscular mycorrhizal fungus Rhizophagus irregularis”. *The Plant Cell* 94 (3 2018), pp. 411–425.

[380] Bakker FT et al. “Herbarium Genonomics, Skimming and Plastome Sequencing”. *Tropical Plant Collections: Legacies from the Past? Essential Tools for the future?* The Royal Danish Academy of Sciences and Letters, 2017.

[381] Zhao T et al. “Phylogenomic synteny network analysis of MADS-Box transcription factor genes reveals lineage-specific transpositions, ancient tandem duplications, and deep positional conservation”. *The Plant Cell* 29 (2017), pp. 1278–1292.

[382] Bakker FT et al. “Herbarium genomics: plastome sequence assembly from a range of herbarium specimens using an Iterative Organelle Genome Assembly pipeline”. *Biological Journal of the Linnean Society* 117 (2015), pp. 33–43.

Summary

Given that nodulating plants do not require exogenous nitrogen fertilizer, engineering nodulation in non-nodulating crops has been a longstanding ambition. From an evolutionary perspective, identifying the genetic changes that led to nodulation can provide key engineering targets. The occurrence of nitrogen-fixing root nodule symbiosis with rhizobium or *Frankia* bacteria is limited to ten plant lineages in four orders: Fagales, Fabales, Rosales, and Cucurbitales. These four orders together form a clade, referred to as the nitrogen-fixing clade. The scattered phylogenetic distribution of nodulating lineages previously led to the hypothesis that nodulation evolved independently multiple times, possibly preceded by a predisposition event at the root of the nitrogen-fixing clade. This thesis presents comparative genomic and transcriptomic analyses to identify genetic changes leading to the evolution of nodulation, as well as innovations in the computational tools required for such analyses.

Chapter 2 consists of a review of known molecular mechanisms in two plant-bacteria symbioses (with rhizobia and *Frankia*) and in two plant-fungus symbioses (arbuscular mycorrhizae and ectomycorrhizae). Specifically, I explore to what extent molecular mechanisms are shared between these four symbioses. The two main commonalities in symbiotic signalling are (1) rhizobium and *Frankia* symbioses are known to co-opt various elements of arbuscular mycorrhizal symbiotic signalling, and (2) plant-secreted flavonoids and strigolactones act as attractants to the symbiont in all four symbioses. Placing the known symbiotic molecular mechanisms in a comparative context will provide a targeted approach at studying the molecular evolution of nodulation.

In **chapter 3** I studied the molecular evolution of nodulation in the only lineage outside of the legumes that engages in rhizobium symbiosis – *Parasponia* – and its non-nodulating sister lineage *Trema*, both from the Cannabaceae family in the order Rosales. I started with assembling and annotating reference genomes for *Parasponia andersonii* and *Trema orientalis* using newly generated data, taking care to avoid lineage specific annotation errors. Using targeted and untargeted approaches, I performed a comparative genomic analysis to identify genetic changes that correlate with the nodulation trait. Following the multiple gain hypothesis, I expected to find evidence for a gain of nodulation in *Parasponia*. However, such evidence could not be found. Instead, I found pseudogenes *Trema* genomes of genes known to be essential for nodulation, which indicated a loss of the nodulation trait in *Trema* species. In an extended evolutionary perspective, I analyzed public data for several non-nodulating lineages in the Rosales, revealing consistent gene loss of *NOD FACTOR*

PERCEPTION 2 (*NFP2*), *NODULE INCEPTION* (*NIN*), and *RHIZOBIUM-DIRECTED POLAR GROWTH* (*RPG*). Combined with the identification of 290 conserved genes that are transcriptionally upregulated in nodules of *Parasponia andersonii* and the legume *Medicago truncatula* (order Fabales), I conclude that the evolutionary origin of nodulation lies at least at the root of the Rosales, and that the trait was subsequently lost multiple times in non-nodulating lineages within the nitrogen-fixing clade.

As genes do not function in isolation, **Chapter 4** extends the perspective of chapter 3 to the evolution of transcriptional networks in nodulation. In model legumes, multiple transcriptional regulators are known to be crucial for nodulation, including *NIN*. However, it is not known if the transcriptional networks controlled by these transcriptional regulators are conserved among nodulating species. Furthermore, since *NIN* is lost in most non-nodulating lineages in the nitrogen-fixing clade, it is likely that the transcriptional networks involving *NIN* are different between nodulating and non-nodulating lineages. I develop a bioinformatic stategy to compare predicted transcriptional networks from RNA sequencing data across multiple species. Unfortunately however, a critical inspection of the accuracy of the predicted networks revealed a false-positive rate of 90%-99%, rendering a comparative analysis infeasible. As is, this chapter provides a complete framework to study the evolution of transcriptional networks, once it is technically feasible to identify such networks on a genome-wide scale.

In **chapter 5** I present GeneNoteBook, a web-based genome-browser for comparative genomics studies in model and non-model organisms. The GeneNoteBook user interface is optimized to facilitate browsing genes and to query genes based on a variety of metadata, including protein domains, orthogroups, and gene ontology terms. Furthermore, GeneNoteBook enables users to edit and add custom notes and attributes to individual genes. This editing feature was used extensively in the targeted analysis of chapter 3, where a team of several domain experts manually curated homologs of known symbiosis genes in the genomes of *Parasponia* and *Trema*. As such, the *Parasponia* GeneNoteBook represents an accessible single source of genomic and transcriptomic data for the *Parasponia-Trema* comparative study system. In general, the GeneNoteBook architecture is not confined to the *Parasponia* / *Trema* system, but can be applied to a wide range of comparative genomics studies.

In conclusion, this thesis represents a milestone in the study of the evolutionary origin of nodulation, and the accompanying molecular changes. As a result of the widespread loss of essential nodulation genes in the

Rosales, a single evolutionary origin of nodulation has now become most likely. This new evolutionary hypothesis raises a variety of novel questions on the molecular mechanisms involved in nodulation, which can be experimentally verified. As such, this thesis is a key example of hypothesis generation through data-driven bioinformatics research.

Curriculum Vitae

About the Author

Rens Holmer was born on the 18th of september, 1987 in Eck en Wiel, the Netherlands. Staying close to home, he completed the BSc and MSc Biology at Wageningen University in 2012 and 2014 respectively. During that time, he acquired an interest in all computational aspects of biological research and frequently worked as student assistent in various courses, including Introduction to Statistics, and Flora & Fauna of the Netherlands. He did a BSc thesis modelling heterogeneous rates of evolution in *Lentibulariaceae* and an MSc thesis modelling evolution of wood anatomical characteristics of *Annonaceae* lianas with dr. Lars Chatrou, an internship on targeted assembly of chloroplast genomes with dr. Freek T. Bakker, and an MSc thesis on assembling genomes of nitrogen-fixing symbiotic rhizobium bacteria with dr. René Geurts. At the end of his MSc studies, Rens enrolled in the graduate program for talented MSc students of the Graduate School Experimental Plant Sciences, where he was supervised by dr. René Geurts to write a research proposal that culminated in the work presented in this thesis.

Publications

Shen D, Holmer R, Kulikova O, Mannapperuma C, Street NR, Yan Z, Maden T van der, Bu F, Zhang Y, Geurts R, and Magne K. “The BOP-type co-transcriptional regulator NODULE ROOT1 promotes stem secondary growth of the tropical Cannabaceae tree Parasponia andersonii”. *The Plant Journal* tpj.15242 (2021)

Lähnemann D, Köster J, Szczurek E, McCarthy DJ, Hicks SC, Robinson MD, Vallejos CA, Campbell KR, Beerenwinkel N, Mahfouz A, Pinello L, Skums P, Stamatakis A, Attolini CSO, Aparicio S, Baaijens J, Balvert M, Barbanson Bd, Cappuccio A, Corleone G, Dutilh BE, Florescu M, Guryev V, Holmer R, Jahn K, Lobo TJ, Keizer EM, Khatri I, Kielbasa SM, Korbel JO, Kozlov AM, Kuo TH, Lelieveldt BPF, Mandoiu II, Marioni JC, Marschall T, Mölder F, Niknejad A, Raczkowski L, Reinders M, Ridder Jd, Saliba AE, Somarakis A, Stegle O, Theis FJ, Yang H, Zelikovsky A, McHardy AC, Raphael BJ, Shah SP, and Schönhuth A. “Eleven grand challenges in single-cell data science”. *Genome Biology* 21.1 (2020), pp. 1–35

Holmer R, Velzen R van, Geurts R, Bisseling T, Ridder D de, and Smit S. “GeneNoteBook, a collaborative notebook for comparative genomics”. *Bioinformatics* 35 (2019), pp. 4779–4781

Velzen R van, Holmer R, Bu F, Rutten L, Zeijl A van, Liu W, Santuari L, Cao Q, Sharma T, Shen D, Roswanjaya Y, Wardhani TAK, Kalhor MS, Jansen J, Hoogen J van den, Güngör B, Hartog M, Hontelez J, Verver J, Yang WC, Schijlen E, Repin R, Schilthuizen M, Schranz ME, Heidstra R, Miyata K, Fedorova E, Kohlen W, Bisseling T, Smit S, and Geurts R. “Comparative genomics of the nonlegume Parasponia reveals insights into evolution of nitrogen-fixing rhizobium symbioses”. *Proceedings of the National Academy of Sciences of the United States of America* 115.20 (2018), E4700–E4709

Zeng T, Holmer R, Hontelez J, Lintel-Hekkert B te, Marufu L, Zeeuw T de, Wu F, Schijlen E, Bisseling T, and Limpens E. “Host- and stage-dependent secretome of the arbuscular mycorrhizal fungus Rhizophagus irregularis”. *The Plant Cell* 94 (3 2018), pp. 411–425

Holmer R, Rutten L, Kohlen W, Velzen R van, and Geurts R. “Commonalities in Symbiotic Plant-Microbe Signalling”. *Advances in Botanical Research*. Vol. 82. Elsevier, 2017, pp. 187–221

Bakker FT, Lei D, and Holmer R. “Herbarium Genonomics, Skimming and Plastome Sequencing”. *Tropical Plant Collections: Legacies from the Past? Essential Tools for the future?* The Royal Danish Academy of Sciences and Letters, 2017

Zhao T, Holmer R, Bruijn S de, Angenent GC, Burg H van den, and Schranz ME. “Phylogenomic synteny network analysis of MADS-Box transcription factor genes reveals lineage-specific transpositions, ancient tandem duplications, and deep positional conservation". *The Plant Cell* 29 (2017), pp. 1278–1292

Bakker FT, Lei D, Yu J, Mohammadin S, Wei Z, Kerke S van de, Gravendeel B, Nieuwenhuis M, Staats M, Alquezar-Planas DE, and Holmer R. "Herbarium genomics: plastome sequence assembly from a range of herbarium specimens using an Iterative Organelle Genome Assembly pipeline". *Biological Journal of the Linnean Society* 117 (2015), pp. 33–43

Acknowledgements

This thesis marks the end of a transformative period in my life, during which I have had the pleasure to encounter a wide variety of people. I feel that all of these people contributed to the completion of this work, either through direct scientific guidance or indirect moral support.

First I want to thank my daily supervisors and promotor for their support throughout the process. René, I am proud to have worked in your interdisciplinary team and grateful for the guidance you provided me in becoming an independent researcher. Your knowledge on symbiosis and ambition to solve the nodule-engineering puzzle have been inspiring. Sandra, your supervision has helped tremendously in my professional and personal development. Your attention to detail and technical knowledge are at the foundation of many things I have learned over the past few years. I have thoroughly enjoyed teaching together and am happy that we will continue working together as colleagues. Ton, your ability to dissect scientific arguments has made me a more critical thinker. Thank you for letting me work on my thesis in your group.

The *Parasponia* team, led by René, has been the cornerstone of the work in this thesis. Special thanks go out to Robin: working together has been educational and fun. Our discussions on the challenges and opportunities of doing computational research in a team of molecular biologists were always entertaining, and your patience and dedication are admirable. The other members of the team have always provided insightful discussions and generous ideas, and working together has been a pleasure. Thank you Arjan, Luuk, Fengjiao, Trupti, Titis, Yuda, Sidney, Jan V, Marijke, Kevin, Defeng, Elena.

Molecular Biology has been an inspirational and welcoming place. I am especially grateful to Wouter for incentivizing discussions and moral support during the final stages of finishing this thesis. Your work ethic and knowledge on hormonal regulation are exceptional. I am proud to see you are getting the recognition you deserve, and look forward to our ongoing collaboration. Maria and Marie-José, you were always there with kind words and administrative support, without which I would not have been able to finish this thesis. The friday borrels after the work-discussion were always a nice way to end the week, and often provided fresh insights for the following week. Thank you Henk, Joan, Rik, Tian, Olga, Jan Ho, Carolien, Guiling, Tingting, Huchen, Xu, Peng, Martinus, Jelle, Lucas, Jieyu, Wenkun, Zhichun.

With my desk in one of the open workspaces, the Bioinformatics Group has been my second scientific home while working on my thesis. I consider myself lucky to have found such a group of likeminded people, and have enjoyed seeing it grow over the years. I cherish the memory of sharing desks with Martijn, Ehsan, Mehmet, Janani, Miguel, and Siavash. Pushing back our chairs to talk science, programming, or politics has lifted my mood more times than I can remember. My other fellow bioinformatics PhD-students and postdocs have been fun and inspiring to work with. Thank you Carlos, Eef, Raul, Sander, Serina, Barbara, Sina, Victoria, Satria, Vittorio, Xiaowen, Ben, Hannah, Hernando, Lotte, Judith, Margi, Mohammad, Zach, Sevgin. The bioinformatics staff go through great lengths to make the group an open and welcoming place, and I am excited to join the team as a researcher. Thank you for all the support Harm, Marnix, Justin, Anne, Aalt-Jan, Dick.

I have learnt the basics of molecular evolution, and the beauty of natural diversity, during my BSc and MSc theses at the Biosystematics Group of Wageningen University. Thank you Eric, Lars, and Freek, for being such supportive mentors and inspiring naturalists. A special thank you to Freek, for suggesting bioinformatics as a specialization I might be interested in. Without that spark, this thesis would not exist.

Wat begon als De Geheime Club 500 en is geevolueerd in De Zulikarners en De Nieuwe Buurtvereniging vertegenwoordigt een grote en diverse groep mensen waarvan ik het geluk heb dat ik ze mijn vrienden kan noemen. Spelletjesavonden met eindeloos veel Catan zorgden voor gezelligheid tijdens het begin van het werken aan mijn proefschrift. Zeilweken, vogeltrips en fietsweekenden hebben voor broodnodige afleiding gezorgd. Belangrijker nog, hoewel we elkaar ondertussen niet meer zo regelmatig zien, weet ik dat ik altijd op deze mensen kan rekenen voor steun en vriendschap. Dankjewel, Niels, Arieke, Josse, Thijs, Charlotte, Martijn, Rinske, Nienke, Thomas, Annelies, Lodewijk, Johannes, Johanna, Marloes, Wouter.

Mijn studie Biologie in Wageningen heeft me naast een diploma ook een groep vrienden met dezelfde fascinatie voor de natuur om ons heen opgeleverd. Onze vogel- en botaniseertripjes – gespekt met schunnige grappen – gebeuren niet vaak genoeg, maar herinneren me altijd aan waarom ik biologie ben gaan studeren en waarom jullie mijn vrienden zijn. Dankjewel Rutger, Tis, Erik-Jan, Marjolein, Jeike, Andre.

Dankjewel Bas en Pola. Jullie horen natuurlijk bij alle eerder genoemde groepjes: studievrienden, fietsvrienden, natuurvrienden. Onze onvoorwaardelijke vriendschap heeft me door moeilijk periodes geholpen en meer bijzondere herinneringen bezorgd dan ik hier kan opnoemen. Het is fantastisch geweest om jullie in Schotland te kunnen opzoeken tijdens jullie promotieonderzoek, jullie doorzettingsvermogen is een belangrijke inspiratie geweest om mijn eigen proefschrift af te ronden.

Mijn vader Henk en mijn schoonouders Hetty en Gerrit hebben mijn reis van afstuderen, verhuizen, verbouwen, vader worden, en nu promoveren met veel betrokkenheid meegemaakt. Zonder hun steun was ik niet waar ik nu ben, ik ben trots om ze dit boekje te kunnen laten zien.

Erika, je bent mijn zielsverwant, mijn maatje, mijn lievelingspersoon. Ik kan je onmogelijk genoeg bedanken voor alles wat je voor me hebt gedaan. Ik kijk op tegen wat je hebt bereikt en geniet van wie je bent. Ons jonge gezinnetje heeft me door de afrondingsfase van mijn proefschrift gesleept, dankzij jou. Dankjewel!

Sanne, ooit zal ik je uitleggen wat er in dit boek staat, dankjewel dat je er bent.

This work was performed with financial support from the European Research Council (ERC-2011-AdG294790) and the Dutch Research Council NWO (865.13.001).