Mathematical Modeling and Evolutionary Analyses of Cell-Surface Signaling in Plants

Inaugural dissertation

for the attainment of the title of doctor

in the Faculty of Mathematics and Natural Sciences

at the Heinrich Heine University Düsseldorf

presented by

Sarah Rose Richards

from Detroit, Michigan, USA

Düsseldorf, 24 April 2018

from the institute for Population Genetics

at the Heinrich Heine University Düsseldorf

Published by the permission of the

Faculty of Mathematics and Natural Sciences at

Heinrich Heine University Düsseldorf

Supervisor: Prof. Dr. Laura Rose

Co-supervisor: Prof. Dr. Oliver Ebenhöh

Date of examination: 28 May 2018

For Joe Ely, Brian Hart, Andy James, Pam Graves, and Jean-Pierre Trías, the teachers who most raised my expectations of myself.

Statement of authorship

This dissertation is the result of my own work. No other person's work has been used without due acknowledgement. This dissertation has not been submitted in the same or similar form to other institutions. I have not previously failed a doctoral examination procedure.

Sarah Rose Richards

Table of Contents

1 Summary

Over the past century, advanced quantitative methods have provided ways to organize and evaluate ever-increasing amounts of data in various scientific fields, including Evolution and Genetics. Bioinformatic tools have become vital in the age of genomics, and mathematical modeling approaches have enabled detailed descriptions of complex biological systems. The work detailed in this thesis is an application of mathematical models and evolutionary analyses to networks of cell-surface-receptor-mediated genetic regulation and their components.

At the tip of the main root in model plant *Arabidopsis thaliana* lie cell layers characterized as having different fates: QC cells, columella stem cells, and columella cells. A series of mathematical models of these cell fates was able to provide insights into the network of transcription factors and peptides regulating cell fate decisions. The models were based on outcomes of experiments on wild type plants and various mutants and overexpression lines. The first model, simulating a single cell with no communication with other cells, was meant to predict cell fate when given initial concentrations of known cell fate regulators: *WOX5*-derived signal and peptide CLE40. It failed to capture the three cell fates. The fact that a second model, simulating a cell column, was able to capture the experimental results highlighted the importance of intercellular communication in achieving robust patterning of long-lived stem cells. After a new experiment was conducted and showed that *WOX5* and its derived signal were not strictly necessary to maintain these stem cells, a third model was generated and showed the plausibility of the existence of another cell fate regulator fulfilling a similar role to *WOX5*.

Statistical models developed to answer population genetics-related questions were used to infer the effects of different kinds of natural selection on wild tomato sequences, specifically those of LysM-RLKs (LYKs/LYRs). This family of genes code for key receptors important to both plant defense and symbiosis. Through phylogenetic analyses, tests of natural selection, and measures of conservation between compared groups of sequences, *Solanum lycopersicum* LYK3 (*Sl*LYK3) set itself apart as a particularly interesting candidate for further study. A bioinformatic analysis found that orthologs of *Sl*LYK8 have intact kinase domains in two wild tomato species, though the kinase is truncated in cultivated tomato. A phylogenetic analysis of the three clades of LysM-RLKs resolved the ambiguous relationships reported in previous phylogenies. The third clade – Group III – was previously assumed to have representatives in only two closely-related species, and further analysis has found them throughout diverse Rosid species.

2 Zusammenfassung

Während des letzten Jahrhunderts haben fortgeschrittene quantitative Methoden es ermöglicht, in diversen wissenschaftlichen Disziplinen kontinuierlich zunehmende Datenmengen zu organisieren und zu bewerten, eingeschlossen der Evolutionsforschung und Genetik. Bioinformatische Methoden sind im Zeitalter der Genomik unersetzlich geworden, und mathematische Modelle haben die detaillierte Beschreibung komplexer biologischer Systeme ermöglicht. Die in dieser Dissertation beschriebene Arbeit ist eine Anwendung mathematischer Modelle und evolutionstheoretischer Analysen auf Netzwerke von Zelloberflächenrezeptor-gesteuerter Genregulation und deren Komponenten.

An der Spitze der Hauptwurzel im Modellorganismus *Arabidopsis thaliana* befinden sich Zellschichten, die entsprechend ihrer Zellschicksale eingeteilt werden: QC Zellen, Columella-Stammzellen und Columella-Zellen. Eine Reihe von mathematischen Modellen der Zellschicksale ermöglichte Einsichten in das Netwerk von Transkriptionsfaktoren und Peptiden, welche Zellschicksal-Entscheidungen regulieren. Die Modelle basierten auf experimentellen Ergebnissen über Wildtyp-Pflanzen und etlichen Mutanten sowie Überexpressions-Linien. Das erste Modell, welches eine einzige Zelle ohne Kommunikation mit anderen Zellen simulierte, war dazu gedacht, Zellschicksale vorherzusagen wenn anfängliche Konzentrationen der folgenden Zellschicksal-Regulatoren gegeben waren: *WOX5*-abgeleitetes Signal und Peptid CLE40. Dieses Modell konnte nicht die drei Zellschicksale beschreiben. Die Tatsache dass ein zweites Modell, welches eine Säule von Columella-Zellen simulierte, die experimentellen Ergebnisse wiedergeben konnte, hebt die Wichtigkeit interzellulärer Kommunikation zur Erlangung robuster Muster von langlebigen Stammzellen hervor. Nachdem ein neues Experiment durchgeführt worden war, welches zeigte dass *WOX5* und das abgeleitete Signale nicht notwendig für das Erhalten des Stammzellschicksals notwendig ist, wurde ein Drittes Modell erzeugt. Dieses zeigte die Plausibilität der Existenz eines weiteren Stammzellschicksal-Regulators, der eine ähnliche Rolle wie *WOX5* annimmt.

Statistische Modelle zur Beantwortung populationsgenetischer Fragen wurden benutzt, um die Effekte verschiedener Arten natürlicher Selektion in den Sequenzen wilder Tomatenpflanzen zu untersuchen, insbesondere von LysM-RLKs (LYKs/LYRs). Die Gene in dieser Familie kodieren für Rezeptoren, die eine Schlüsselrolle sowohl für die Pflanzenabwehr als auch für Symbiose spielen. Durch phylogenetische Analysen, Tests auf natürliche Selektion und einem Maß für die Varianz in verglichenen Gruppen von Sequenzen, stach *Solanum lycopersicum* LYK3 (*Sl*LYK3) als ein besonders interessanter Kandidat für weitere Untersuchungen hervor. Eine bioinformatische Analyse zeigte, dass Orthologe von *Sl*LYK8 eine intakte Kinasedomäne in zwei Spezies wilder Tomaten haben, obwohl diese in kultivierten Tomaten verkürzt ist. Eine phylogenetische Analyse von drei Kladen von LysM-RLKs löste die mehrdeutige Beziehung, die zwischen den Kladen in bisherigen Phylogenien gefunden worden war. Von der dritten

Klade – Gruppe III – war vorher angenommen worden, dass sie nur Vertreter in zwei eng verwandten Spezies hatte, aber weitere Analysen haben sie auch in diversen Rosiden-Spezies nachgewiesen.

3 List of publications

3.1 Publications included in this thesis

1) Richards S, Wink RH, Simon R. Mathematical modelling of *WOX5*-and *CLE40*-mediated columella stem cell homeostasis in Arabidopsis. Journal of experimental botany. 2015 May 26;66(17):5375-84.
2) Richards S, Rose L. Selection on LysM-RLKs (LYKs/LYRs) in wild tomatoes and phylogenetic analysis in Angiosperms. BMC evolutionary biology. 2018 (in review)

4 Introduction

4.1 Indicators of Neutrality and Natural Selection

Biodiversity and adaptation exist due to mutations accrued by living organisms and passed down to their offspring. When a mutation happens at a single nucleotide within a DNA sequence, it is called a point mutation[1 p.25]. Evidence of their occurrence can frequently be found between sets of closely-related sequences, such as those of corresponding (orthologous) genes from different individuals of the same species. Point mutations are the driving force behind single nucleotide polymorphism (SNP), the variation within a set of sequences that occurs at one nucleotide position[1][2]. The positions (or sites) of SNPs are called segregating sites [3 p.2], and, for the work presented here, each unique sequence in a set of orthologous genes is defined as a haplotype. Systematic analyses of the SNPs throughout a set of orthologous gene sequences can yield a great deal of information about the evolutionary relationships of the individuals that supplied the sequences and about the pressure natural selection has exerted on the sequences.

Genes are sequences of DNA that form a blueprint for something that serves a function, such as a protein. In this case, the gene's DNA is transcribed into mRNA, which is then translated into an amino acid sequence, which then results in a functional protein[2] [4 pp.306,309]. Some point mutations change the DNA sequence of a gene but not the resulting protein's amino acid sequence (synonymous substitutions) [5 p.10]. The protein encoded by the DNA then remains unchanged, though the mRNA and the codon (the 3-letter code used to translate from mRNA to protein) does change. Alternatively, non-synonymous mutations do result in a change in the amino acid sequence [5 p.11]. When a change happens at the protein level, it is more likely to cause a significant change to protein function, which in turn is more likely to have consequences for the organism [1 p.119]. Consequently, relationships between the frequencies of non-synonymous and synonymous polymorphisms can provide indications for certain kinds of selection in some circumstances. π, the nucleotide diversity, is a standard measure of the polymorphism within a set of sequences [1 p.59]; it can be calculated for all polymorphisms at once or separately for synonymous and non-synonymous polymorphisms. If a gene or gene region has an equal measure of π for synonymous and non-synonymous differences (π_s and π_a respectively), it implies that the amino acid changes are neutral [5 p.51]; they cause neither harm nor benefit and occur at the same rate as polymorphisms that do not change the amino acid sequence. If π_a is less than π_s, this indicates that the change to the protein is more likely to be harmful; this happens e.g. when the structure of a protein is important for function and depends heavily on the amino acid sequence. This kind of selection is called purifying selection, because the individuals that possessed the non-functional (or less-functional) proteins have been weeded out through natural selection, and the non-synonymous mutations that ocurred are now missing from the data based on living individuals [5 p.51]. If π_a is larger

1: SNPs are often defined to occur in at least 1% of the population, but no arbitrary cut-off is implied here.
2: Small proteins are called peptides. The (somewhat arbitrary) cutoff is around 50 amino acids.

than π_s, that indicates that the non-synonymous mutations are beneficial (adaptive selection); this happens during adaptation, when changes in a protein make it more efficient or give it a new purpose [5 p.51]. The further the ratio π_a/π_s is away from one, the more confident one can be that purifying or adaptive selection has ocurred. All of these results rely on the idea that synonymous mutations are neutral. But these kinds of mutations can increase or decrease fitness (e.g. some codons are rare and less efficient in translation for highly expressed genes) [5 p.13]. Use of these principles must be accompanied by either a caveat or an assurance that the condition of synonymous mutation neutrality is met.

Another method that relies on the neutrality of synonymous mutations is the McDonald-Kreitman test, which tests for violation of the hypothesis of neutral evolution by comparing sequences from two populations (perhaps two different species or individuals from the same species living in two separate places) [6]. The segregating sites are sorted according to whether they are fixed differences (the two populations do not have the same nucleotides in common at a site) or polymorphisms (the differences occur within a population). The fixed differences and polymorphisms are then further sorted into synonymous or non-synonymous differences, and these are tallied. A G-test (similar to a χ^2 test) or similar statistical test can then be used to determine if the ratios of non-synonymous to synonymous fixed differences are significantly different from the ratio of non-synonymous to synonymous polymorphisms. These ratios are expected to be equal for genes which have evolved under neutral conditions. If the fixed difference ratio is higher, this implies that one of the populations has undergone adaptive evolution (or that both populations have, but in different ways). If the polymorphism difference is higher, this can be result of balancing selection, where the presence of two or more variants of a gene and the resulting high probability of heterozygosity are beneficial (as in the case of gene variants responsible for both malaria resistance and sickle cell anemia) [7]. Like nucleotide diversity analysis, this test should be used with caution; violation of the test's assumptions can lead to false positives [8].

4.2 Phylogenetics

Phylogenetic trees (or phylogenies) are a common way to explore and present the evolutionary history of genes or other sequences in terms of their relationships to one another. On an accurate phylogeny, genes which shared a common ancestor more recently and are therefore more similar are grouped closer together in the phylogeny[1]. A standard method of inferring phylogenies is the maximum likelihood method, which attempts to find the phylogeny that fits the following condition: assuming the phylogeny is correct, the probability of the sequence data occurring is maximized[1p.198;9]. The probability of the sequence data occurring depends on substitution matrices, organized collections of mutation probabilities from one nucleotide or amino acid to another[5 pp.27-28,35]. These

1: Before an assessment of similarity is made, a process called alignment is often necessary. Sections of related sequences can sometimes be deleted or inserted over time, and alignment creates spaces in the sequences so that related nucleotides or amino acids are placed at the same positions.

probabilities are not necessarily equal. For instance, a mutation from amino acid Aspartic acid (DNA sequence: GAT or GAC) to Glutamic acid (GAA or GAG) requires only one amino acid change in the third position, while a mutation from Aspartic acid (GAT or GAC) to Arginine (AGG or AGA) requires all three nucleotides to change. The first mutation is much more likely, and this must be taken into account when calculating the probability of the sequence data occurring. Various methods exist to find these probabilities, because there are different models of mutation, and some models fit better than others depending on species or sequence type [5 p.29].

Once the tree that best describes the data has been found, it can then be tested for reliability using a sampling method called bootstrapping. Bootstrapping involves using a subset of the positions from the alignment to make new trees referred to as bootstrap replicates [1 p.209]. If two genes are virtually identical along their entire sequence, they will group together on each of the bootstrap trees. If, however, there are three genes with the first sharing parts of its sequence with a second and other parts with a third, the sampling will result in the first gene sometimes being paired with the second gene and other times with the third. In this case, some bootstrap replicates will group genes one and two together while others will group genes one and three together. The results of the bootstrap analysis are summarized in the bootstrap values: numerical labels on each branch showing what percentage of the bootstrap trees match the original tree [1 p.209]. The higher the bootstrap value and number of bootstrap replicates calculated, the more sure one can be that the original tree is correct [1 p.211].

4.3 Mathematical Modeling of Genetic Regulatory Networks

In addition to providing a blueprint for protein production, genes are part of a complex network of interactions called gene regulatory networks. Within these networks, gene products – proteins or mRNA transcripts – affect the expression of other genes. Mathematical models of the networks can identify processes leading to optimization of functions controlled by the network or incomplete or false hypotheses. The models are typically derived from principles of chemical interactions or based on simplifications of those principles or already-existing models. When the models are solved, they provide the concentrations or activity of the gene products. A standard method derives ordinary differential equations (ODEs) from Michaelis-Menten kinetics and conservation laws to provide a system of nonlinear ODEs; changes in gene product concentration over time are given by Hill functions in the simplest cases [10,11 p.13]. Solving these equations can be labor-intensive, with time required to find a solution depending on precision and network size. Both simpler and more complicated approaches have been used regularly, the simpler ones for larger networks where the labor cost is unacceptable or data is scarce and the complicated ones when it is known that stochastic effects may make a substantial difference and every interaction needs to be modelled separately [12]. The ODE approach is suitable

when some detail is desirable and the time taken to solve the equations is acceptable, which is more often the case with small networks.

In any case, all models have parameters, values that are not necessarily interesting to the researcher but that are required by the model. The variables in the ODE models mentioned above are the concentrations of the gene products as a function of time: these are generally the desired values and what the solution to the ODE system provides, but those values depend on the parameters, such as maximum production rates and degradation rates [11 p.13]. If these values are not known, they must be measured with biochemical experiments, inferred from known concentrations of the gene products, or guessed. The values of these parameters can have drastic effects on the solution or virtually none at all. Typically, any given model has some important parameters and some unimportant ones [13]. Since a model can make large errors in solving for gene concentration when the more important parameters are badly estimated, it is important to perform a sensitivity analysis to determine which variables need special attention. Sensitivity analyses measure how sensitive the model is to change in any given parameter, in other words, how much the model's solutions change when a parameter takes on a different value. A very simple sensitivity analysis can be done by picking a fixed set of values for the parameters and then changing each parameter's value one at a time (local analysis), which works well if the parameters are already known with some precision. Ideally, the parameters should be changed together in groups as well so that combinatoric effects are not overlooked, and several fixed sets of parameter values should be used to cover the entire plausible range of the parameters (global analysis) if the parameters are not known [14]. How the model reacts to changes in parameter values will depend on the nature of the model, with negative feedback loops lending stability to the network and positive feedback loops causing instability [15].

4.4 Plant-Microbe Interactions

Plants are bombarded by an assortment of other organisms, some of which can infect and harm them. Some fungi, bacteria, nematodes, and oomycetes penetrate the plant's outer surface and use the plant for nourishment, either by colonizing it to take resources or by killing the plant's cells and ingesting the degrading plant matter [16,17]. The plant's first line of defense is to prevent further intrusion by strengthening cell walls [18], closing channels between cells [19], or producing toxins [20]. Different modes of defense work better on some invaders than others, so the plant should recognize the offending microbe and cue an appropriate response. Fortunately for plants, these microbes often shed recognizable compounds – or molecular "patterns" - such as chitin, that the plant can then detect through extracellular receptors [16], often receptor-like kinases (RLKs) [21]. This defense response is appropriately named pattern-triggered immunity (PTI) [22]. Infection and defense, however, are categorized by some researchers [23] as an orderly arms race, and some pathogens have evolved to

prevent the PTI responses through the use of secreted molecules called effectors [22]. In yet another step in these interactions, plants have developed ways to recognize the effectors. The resulting response, called effector-triggered immunity (ETI), is considered generally more extreme than PTI and can involve hypersensitive response and induced cell death [22].

Not all microbes that try to enter a plant are necessarily harmful, however. Under certain conditions, some plants can benefit from hosting a symbiont. Plants are sometimes colonized by mycorrhizae or rhizobia, fungi and bacteria respectively, that can aid the plant in the uptake of nutrients [21]. In this case, rather than strengthening defenses and sacrificing cells to cut the microbe off, the plant is better off if it aids the symbiont in its colonization. But this approach also requires recognition of the symbiont and triggering of the appropriate response, just like in the case of pathogens [16].

Further complicating the situation, some organisms are difficult to classify neatly into any described category. Even organisms classified as symbionts can "cheat" in their symbiosis and become more like pathogens to their host [24]. For this reason, it may be more realistic to describe microbes as having a position on a spectrum from pathogenic to symbiotic [16], depending on the effect they have on plant fitness. This effect and resulting classification on the spectrum can be different for individuals within a species or even for the same individual over time and in diverse circumstances [24]. Because they need to deal with these complications, plants have undergone a great deal of pressure to develop a complex network of receptors that cue the appropriate response after processing information about what is attempting to gain entry [21].

4.5 LysM-Receptor-like Kinases

The RLKs are a major class of receptors used by plants in PTI and symbiosis initiation [16,25]. RLKs are proteins with an extracellular domain (that detects e.g. the molecular "patterns" given off by pathogens), an intracellular (or kinase) domain[1] (which passes signals to the interior of the cell), and a transmembrane domain (which passes through the cell membrane to connect the other domains) [16,25]. The extracellular region detects the molecular "patterns" when they bind to them, and this induces signaling between the kinase domain and regulators of defense and symbiosis within the cell. The LysM-RLKs (LYKs and LYRs) are RLKs containing three LysM motifs [26] (short recurring sequences) within their extracellular domains. Members of this gene family are known to be involved in several processes in both PTI and symbiosis [25]. Some of the individual genes are known to serve multiple roles, regulating both defense and symbiosis [27].

LysM-RLKs have been described with three major clades on phylogenies of their sequences [26]. Genes in Group I have kinase domains which share conserved amino acids with known functional kinase

1: Note that "kinase domain" is an ambiguous term; it can refer to most of the intracellular part of the protein sequence, or to a highly conserved sequence covering a small portion of this larger sequence. Each whole intracellular domain contains several of these highly conserved, smaller sequence regions. The meaning of "kinase domain" must be determined from context.

domains, and they have 10-12 exons (translated sections of the gene, which are separated by untranslated parts of the gene called introns). Group II genes have several mutated and presumably non-functional kinase domains [26]; each Group II LysM-RLK in model plant *Arabidopsis thaliana* and cultivated tomato *Solanum lycopersicum*, for instance, is missing its Glycine-rich loop [28]. Their kinase domains may not be active as described above [26]. In addition, they have one or two exons. Group III genes have a combination of these features: they have fewer exons like the genes in Group II, but they have classically conserved kinase domains like those in Group I [26]. Genes belonging to Groups I and II are found throughout diverse land plant species, but Group III has only been previously described in *Lotus japonicus* [29] and *Medicago truncatula* [26], two closely related species within the Order Fabales (a clade of the dicots) [30]. Researchers who discovered the Group III LysM-RLKs in *M. truncatula* have suggested that Group III may have arisen from the fusion of a gene region containing the LysM domains of a Group II LysM-RLK with a gene region containing the kinase domain of a protein outside of the LysM-RLK family; *M. truncatula* LYR5 and LYR6, both members of Group III, are 59% identical to WAK-like proteins from *A. thaliana* [26]. The group that discovered them in *L. japonicus* has the same hypothesis [29]. Phylogenies of the LysM-RLKs which include members of Group III, however, have not had strong bootstrap support for the closer relationship between Group II and Group III genes [26,29].

S. lycopersicum (cultivated tomato) has especially little genetic variation, and applied geneticists are regularly trying to unravel the effects of wild tomato genes on phenotype to develop methods for controlling disease and pathogen resistance in cultivated tomato [31]. The availability of new mapped transcriptomes of several wild tomato species [32] provides an opportunity to mine information about the evolutionary history of orthologs to known LysM-RLKs using population genetic techniques and phylogenetics and make suggestions about which genes have been instrumental in the evolutionary history of wild tomato species' defensive tactics.

4.6 The *WOX5*/*CLE40* Columella Stem Cell Regulatory Network

Some effectors of plant parasites mimic regulatory elements of plant development to proliferate cell types beneficial to them and increase their supply of nutrients, as in the case of nematodes secreting mimics of members of the development-regulating *CLE* gene family [16,33]. *CLE40*, a member of this family, produces a peptide that signals through an RLK ACR4 to encourage differentiation (cell specialization) in the primary (main) root tip of *A. thaliana*. It has a mutually antagonistic relationship with a transcription factor (protein controlling the expression of another gene) WOX5. *WOX5* and *CLE40* expression is known to affect columella cell (CC) and columella stem cell (CSC) fate. CCs, which lie at the root tip, help to detect gravity and provide protection for the rest of the developing part of the root. The CSCs divide to provide a source for these cells (fully differentiated CCs do not divide). The source of the CSCs themselves is a small, undifferentiated group of cells called quiescent center (QC)

cells. These cells divide, and their progeny form each of the different kinds of tissues in the root. There is typically one layer of columella stem cells between the QC and CCs, although variation does occur. The consistent occurrence of this layer of stem cells is referred to as stem cell homeostasis. *WOX5* is expressed in the QC and is thought to be necessary and sufficient for CSC maintenance, but there has been doubt about whether or not WOX5 itself travels to the CSCs to keep them from differentiating into CCs. In either case, the actual regulator of CSCs would be a *WOX5*-dependent signal, either WOX5 itself or some mobile protein or regulator of CSC fate that requires *WOX5* for its production or activity. In contrast to *WOX5*, *CLE40* is expressed in the CCs and encourages the CSCs to differentiate into CCs. In summary, *WOX5* is expressed from the QC and its signal encourages CSCs to remain CSCs, while *CLE40* is expressed from the CCs and encourages CSCs to differentiate instead [34].

A mathematical model has been successfully applied to the corresponding developmental system in the above-ground parts of the plant. It highlighted the robustness of the network to signaling noise, the need for spatial separation of the cell layers, and the usefulness of a combination of positive and negative feedback loops in maintaining the stem cell patterning in spite of parameter value fluctuations [35]. In addition, root growth has been extensively modeled as a function of auxin regulation [36,37]. Mathematical models have yielded insights into the field of developmental genetics, but they have not yet been applied to the *WOX5*/*CLE40* network.

5 Aims of this thesis

Intercellular signaling through cell-surface receptors permeates every major interaction plants have with their surrounding environment and controls many aspects of the coordination of cell activity within the plant. Studies concerning this kind of signaling are abundant, but knowledge about interaction partners and other details is known to be missing. The research included in this thesis was done in order to synthesize what is known about different kinds of signaling pathways and to provide insight and suggestions for future research obtained using a variety of modeling, bioinformatic and other advanced quantitative methods. To this end, select components of two intercellular signaling pathways involving cell-surface receptors were studied:

1) a network composed of CLE40 and *WOX5*-dependent signal, which affects columella stem cell homeostasis, to determine which characteristics of the network result in long-lived stem cells.
2) a versatile family of cell-surface receptors, the LysM-RLKs, to determine the effects of natural selection and sequence differences on their ability to detect pathogenic and symbiotic organisms and elicit immunological and host responses in wild tomatoes.

6 Summary of the obtained results

6.1 Publication I

To gain insight into primary root meristem genetic regulation, I employed mathematical models of known processes governing columella stem cell fate. The aim of the research was to discover principles of regulation (such as feedback loops and robustness) in this network and provide suggestions that would help experimentalists to discover new proteins or other regulators affecting stem cell fates.

Using stem cell layer counts as a proxy for mutant phenotype, several models were conceived, evaluated, and rejected or verified, revealing important aspects of columella stem cell regulation. The first model, that of a single cell with fate governed by a *WOX5*-dependent signal and CLE40 alone, failed to capture the three distinct fates (QC,CSC,CC) found in nature, indicating that multi-cell coordination and diffusion of regulatory proteins may play a vital role in the process of cell fate determination. This model performed as a switch between one or two cell fates, and the number of fates was highly sensitive to some of the parameters. A second model, which included a cell column through which the proteins could freely diffuse, fared better; diffusion provided a limited stabilizing effect, and this model was able to capture the phenotypes of each of the previously published *wox5* and *cle40* mutants as well as *WOX5* and *CLE40* overexpression lines. However, results from a new experiment with *wox5*/*cle40* double mutants could not be explained by this model due to the crucial role thought to be played by *WOX5* in stem cell fate. It had previously been assumed that long-lived stem cells were entirely dependent on *WOX5*: if *WOX5* was not functional, there would not consistently be a layer of columella stem cells. The *wox5*/*cle40* double mutants showed that this was not the case, since they lacked *WOX5* functionality and had, on average, one layer of stem cells. Taken together, this established that the hypothesis used to inform this model was wrong, and that some other regulator(s) of stem cell fate remains undiscovered. A new model, which included another regulator of stem cell fate playing a similar role to *WOX5*, was able to describe all of the mutant phenotypes and overexpression lines (double mutants included), indicating that such a regulatory network could plausibly be acting on the columella stem cells.

The work I contributed to this research highlighted the importance of intercellular signaling in stem cell homeostasis and verified the possibility of another WOX5-like protein affecting columella stem cell fate. It also supported the general principle of positive feedback loops and their resulting instability being an important part of a cell fate decision-making.

6.2 Publication II

RLKs play many vital roles in cellular signaling. To learn more about the evolutionary history of those involved in extracellular signal perception, a key group of RLKs involved in plant-microbe interactions – the LysM-RLKs – was selected for analysis. Recently sequenced transcriptomes of wild tomato species provided an excellent resource for sequence data and enabled population genetic analyses of the LysM-RLKs for these species. The aim of this research was to perform population genetic analyses and uncover aspects of the evolutionary history of these genes, with a focus on genes from *Solanum* species.

An analysis of synonymous and non-synonymous mutations in intracellular and extracellular domains separately revealed an interesting phenomenon in wild tomato *Sl*LYK3 orthologs: unlike most *Sl*LYK orthologs, its extracellular domain was subject to at least as strong purifying selection as its intracellular domain. This implies that the extracellular part of the amino acid sequence of *Sl*LYK3 (which detects ligands shed by pathogens or symbionts) is more important to its function than those of the other LysM-RLKs. In addition, the intracellular domain of *Sl*LYK8 orthologs was found to have undergone purifying selection, which was a surprising result due to the presumed lack of functionality of *Sl*LYK8. Further analysis revealed that, unlike *Sl*LYK8, some of its orthologs in wild tomatoes have intact intracellular domains. To place the evolution of *Solanum* LysM-RLKs in a broader context and match phylogeny to known functions, evolutionary analyses were performed on LysM-RLKs from *S. lycopersicum*, *A. thaliana*, *O. sativa*, *M. truncatula*, and *L. japonicus*. These analyses revealed that the Group III and Group II clades of LysM-RLKs very likely share an ancestor that is more recent than those between Group I and Group II. A BLAST search of the four previously known Group III LysM-RLKs identified a further 88 protein sequences from 24 genera as putative Group III LysM-RLKs. An evolutionary analysis of the individual LysM domains of each LysM-RLK gene from *S. lycopersicum*, *A. thaliana*, *O. sativa*, *M. truncatula*, and *L. japonicus* revealed that the first two LysM domains of *Sl*LYK3 belong to clades of closely-related LysM domains which do not align well with the other domains. The other genes whose domains belong to these clades cover a wide range of genera and are also closely related according to the evolutionary analysis of whole LysM-RLK genes. This result implies that *Sl*LYK3 and its close relatives have distinct LysM domains that have been preserved on a long timescale.

My work on this research provided support for the hypothesis that Group III LysM-RLKs are more closely related to Group II LysM-RLKs than to Group I LysM-RLKs. Additionally, it recommended *Sl*LYK3 as an especially interesting candidate for further study, due to characteristics of its extracellular domain.

7 Conclusions

There are plenty of opportunities to apply advanced quantitative methods to biological subjects, but these tools are most useful when there is an abundance of information that cannot be easily evaluated without a systematic quantitative method such as a model, statistical test, or phylogeny. However, limitations in experimental measurements are the norm, and under some circumstances, models can still be informative without much input. Sometimes they point out principles that seem obvious once they are suggested. This was the case for the first model in Publication I. *WOX5* and *CLE40* both repress each other's expression; essentially, *WOX5* expression is part of a positive feedback loop. Each model was unsurprisingly (after the fact) sensitive to *WOX5*-derived signal production and degradation rates, resulting in vastly different outcomes depending on these parameters. Still, the single-cell model helped to connect the positive feedback instability phenomenon to this particular network, and the C/W multi-cell model was falsifiable despite the sensitivity of the model to unmeasurable parameters. Further, the C/W/X model was able to show the plausibility of a hypothesis that another regulator of cell fate exists and promotes CSC fate. In the case of the phylogenetic study in Publication II, the bootstrap values of the LysM-RLK whole protein sequence phylogeny were a convincing piece of evidence that Group III LysM-RLKs are more closely related to Group II LysM-RLKs. This conclusion, based on 49 sequences with variation at nearly every amino acid site, would have been impossible to achieve without a reliable phylogeny. And the ability to pick the best fitting of a variety of substitution matrices using RAxML resulted in a much clearer picture of this development than what was available before. The abundance of data generated from several different wild tomato individuals allowed statistical testing for selection on the intracellular *Sl*LYK8 gene, which eventually led to the conclusion that some wild tomato orthologs had significant differences to their cultivated tomato relative in essential parts of their sequences. Statistical tests of neutrality, phylogenies of individual LysM domains, and a method for organizing the results of many alignments all pointed independently to the *Sl*LYK3 extracellular domain as a likely conserved, distinct, and important perceiver of microbial signals, a prime candidate for further exploration.

8 Publications associated with this thesis

8.1 Publication I: Mathematical modelling of *WOX5*- and *CLE40*-mediated columella stem cell homeostasis in Arabidopsis.

Authors: Sarah Richards, Rene Wink, Rüdiger Simon.

This article was published in the *Journal of Experimental Botany* in the year 2015.

Contribution of Sarah Richards:

Major

- conceived the models
- conducted the simulations
- drafted the manuscript
- created and edited all figures except Figure 7 and the root outline used in Figures 1, 6 and 9

Supplementary material can be accessed via the publisher's website:

https://academic.oup.com/jxb/article/66/17/5375/541103

Journal of Experimental Botany, Vol. 66, No. 17 pp. 5375–5384, 2015
doi:10.1093/jxb/erv257 Advance Access publication 26 May 2015
This paper is available online free of all access charges (see http://jxb.oxfordjournals.org/open_access.html for further details)

RESEARCH PAPER

Mathematical modelling of *WOX5*- and *CLE40*-mediated columella stem cell homeostasis in *Arabidopsis*

Sarah Richards[1], Rene H. Wink[1,†] and Rüdiger Simon[1,*]

[1] Institute of Developmental Genetics, Heinrich Heine University, 40225 Düsseldorf, Germany

* To whom correspondence should be addressed. E-mail: ruediger.simon@uni-duesseldorf.de
† Present address: Institute of Transformative Bio-Molecules (WPI-ITbM), Nagoya University, Nagoya 464-8602, Japan

Received 2 March 2015; Revised 22 April 2015; Accepted 28 April 2015

Editor: Thomas Dresselhaus

Abstract

The root meristem of *Arabidopsis thaliana* harbours a pool of stem cells, which divide to give rise to the differentiated cells of the various root tissues. Regulatory networks of inter-cellular signalling molecules control the homeostasis of stem cell number and position so that both stem and differentiated cells are consistently available. This work focuses on the transcription factor *WUSCHEL-RELATED HOMEOBOX 5* (*WOX5*), the signalling peptide *CLAVATA3/EMBRYO-SURROUNDING REGION 40* (*CLE40*) and the feedback loops involving them, which maintain the columella stem cells (CSCs). WOX5 signals from the quiescent centre (QC) to promote stem cell fate, while CLE40 is secreted from the differentiated columella cells (CCs) to promote differentiation. Our analyses of mathematical models of this network show that, when cell fate is controlled primarily by antagonistic factors, homeostasis requires a spatial component and inter-cellular signalling. We have also found that WOX5 contributes to, but is not absolutely necessary for, CSC maintenance. Furthermore, our modelling led us to postulate an additional signalling molecule that promotes CSC maintenance. We propose that this WOX5-independent signal originates in the QC, is targeted by CLE40 signalling and is capable of maintaining CSCs.

Key words: Columella cells, CLE40, gene regulatory networks, peptide signalling, root development, stem cell homeostasis, WOX5.

Introduction

The root stem cell niche of *Arabidopsis thaliana* is a collection of undifferentiated cells which divide to give rise to the many different root cell types. In the centre of the niche is the quiescent centre (QC), a group of four cells which maintain the identity of the stem cells and utilize marginal cell division activity to replenish the stem cell supply. The stem cells proximal to the QC are the vascular initials, and the stem cells lateral to the QC are initials for the endodermis, epidermis and lateral root cap. Those distal to the QC are the columella cell initials, also called columella stem cells (CSCs). Their descendants, the columella cells (CCs), are located distal to the CSCs and they detect the direction of gravity, store energy by accumulating starch and provide a protective layer for the stem cell niche.

The stem cell niche is made up of the QC cells and one layer of adjacent stem cells surrounding it. The cells in this layer serve as initials for all of the cell types in the various root tissues; when a stem cell divides, the cell in contact with the QC remains a stem cell while the other enters a differentiation pathway. The differentiated cell then fulfills particular

tasks necessary for the plant's development and function. Stem cell homeostasis (a steady number of stem cells) is necessary so that there is a supply of both the differentiated cells and their stem cell initials (Stahl and Simon, 2005).

If the QC is ablated, the stem cells around it differentiate. The QC then reforms in a location proximal to the original QC site and a new stem cell niche is formed. This suggests that the surrounding stem cells are maintained by short-range signalling from the QC (van den Berg *et al.*, 1997). The number and position of the CSCs are regulated by signals from both the QC and CCs. The QC expresses the transcription factor *WUSCHEL-RELATED HOMEOBOX 5* (*WOX5*), which promotes CSC fate, while the CCs express and secrete the signalling peptide CLAVATA3/EMBRYO-SURROUNDING REGION 40 (CLE40) to encourage differentiation into CCs (Sarkar *et al.*, 2007; Stahl *et al.*, 2009). A typical wild-type *Arabidopsis* root has one layer of CSCs in the first layer distal to the QC (D1). The cells from D2 to the root tip have starch granules, the trait used to distinguish between CC and CSC identity (Fig. 1). It has been shown that constitutive expression of *WOX5* results in massive accumulation of CSCs, while *wox5-1* loss-of-function mutants typically have starch in the D1 layer. This suggested that WOX5 is both necessary and sufficient for CSC maintenance (Sarkar *et al.*, 2007).

WOX5 is a homologue of the transcription factor WUSCHEL (WUS) which is expressed in the organizing centre of the shoot apical meristem (Sharma *et al.*, 2003; Sarkar *et al.*, 2007). WUS is a mobile protein which regulates cell fate non-cell-autonomously and expression of WUS in place of WOX5 can functionally replace WOX5, suggesting that both proteins perform very similar functions at opposite ends of the plant (Yadav *et al.*, 2011). However, how *WOX5* acts from the QC to regulate CSC fate in the neighbouring distal cells and whether WOX5 is also a mobile protein is not yet known. It is assumed that either *WOX5* itself, or a gene transcriptionally controlled by WOX5, gives rise to a signal that moves from the QC to the CSCs to maintain their cell fate. This WOX5-dependent signal (W) could be the WOX5 protein moving from the QC through plasmodesmata to promote stem cell fate in adjacent cells (as WUS does in the shoot), or it could be a mobile signal generated by WOX5 that diffuses to the next cell layer (Sarkar *et al.*, 2007).

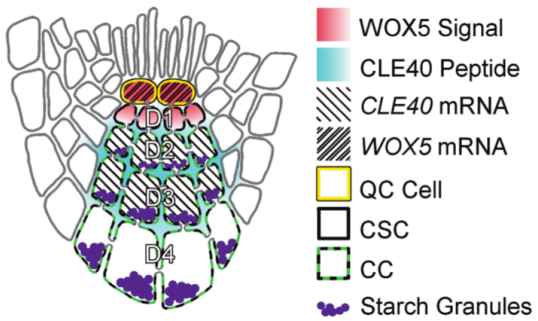


Fig. 1. WOX5 signal and CLE40 peptide locations and their effect on CSC fate. *WOX5* is expressed in the QC and signals to maintain CSCs distal to the QC. *CLE40* is expressed in CCs, the differentiating daughters of the CSCs, and secreted into the apoplast, where it can act through plasma membrane-localized receptor-like kinases and inhibit the WOX5 signal. CCs are identifiable by their stainable starch granules.

Acting antagonistically to the *WOX5*-mediated promotion of CSCs, *CLE40* is expressed in the CCs themselves to promote differentiation. CLE40 is secreted into the inter-cellular space and interacts with receptor-like kinases (RLKs) ARABIDOPSIS CRINKLY4 (ACR4) and CLAVATA1 (CLV1), embedded in the plasma membrane of neighbouring cells, to limit *WOX5* expression. Intriguingly, both ACR4 and CLV1 preferentially localize to plasmodesmata, where they may directly regulate the mobility of QC-derived stemness factors in a CLE40-dependent manner (Stahl *et al.*, 2013). In *cle40-2* mutants the *WOX5* expression domain is expanded, while the differentiation of CSC descendants is delayed, often resulting in the maintenance of two stem cell layers. The addition of synthetic CLE40 peptide (CLE40p) decreased the number of stem cells in a dose-dependent manner. Addition of CLE40p to wild-type roots promoted differentiation to CCs. There is a substantial difference in the phenotypes of wild type and *wox5-1* mutants when sufficient CLE40p is added; with the same dosage of CLE40p, the *wox5-1* mutants have a higher frequency of starch granules in the QC position. These results together indicate that CLE40 promotes differentiation via two pathways, one independent of WOX5 and the other via WOX5 (Stahl *et al.*, 2009).

WOX5 and CLE40 are part of a subnetwork within the larger network of QC and CSC fate-governing factors. *WOX5* expression is downstream of several pathways involving the auxin maximum around the QC (Ding and Friml, 2010). *WOX5* expression also relies on a transcription factor *SCARECROW (SCR)*, which, similarly to *WOX5*, is expressed in the QC, specifies QC identity and maintains stem cells. Both WOX5 and SCR can function redundantly to maintain the cortex initials proximal to the QC (Sarkar *et al.*, 2007). Several other transcription factors have been reported to control CSC abundance, including auxin response factors (ARF) ARF10 and ARF16 (Wang *et al.*, 2005) and a regulatory feedback loop between the NAC-domain proteins FEZ and SOMBRERO (SMB; Willemsen *et al.*, 2008). Whereas ARF10 and ARF16 were suggested to restrict CSC fate in a parallel pathway to WOX5, it was hypothesized that SMB could negatively regulate WOX5 via the CLE40/ACR4 receptor module (Bennett *et al.*, 2014). The auxin responsive protein IAA17 was also shown to indirectly regulate CSC fate through mediation of the auxin response in the QC, which is crucial for WOX5 activity (Tian *et al.*, 2014).

Here we used mathematical modelling to test whether the current information available on WOX5 and CLE40 and their known interactions can explain in sufficient detail the observed cell fates in different mutant backgrounds and upon experimental changes of the amount of individual components. We have developed three mathematical models for the CSC fate-governing regulatory network consisting of WOX5 and CLE40. We determined that a single-cell model of this network, which lacked signals from other cells, was incapable of describing the observation of long-lived CSCs. Although

a first multi-cell model of the network was sufficient to simulate the majority of biological tests, only a modified multi-cell model introducing an additional stem cell promoting factor into our network was able to describe all experimental results.

Materials and methods

Model and simulations

All model equations were solved using Matlab function ode45, which provides solutions for ODEs at discrete time points. Parameter values were scaled by hand.

For the multi-cell models, solutions were obtained for time t from 0 to 100, where values of the variables were compared at t=99 and t=100. If the values differed by more than 1e-5, the simulations were run for a longer time. The values of the variables at the last time point were used to determine cell fate.

Plant accessions

Arabidopsis thaliana ecotype *Columbia* (Col-0) was used as wild type. Col-0 was the background for all mutant seeds. *wox5-1* mutant seeds (SALK_038262) were obtained from the Nottingham *Arabidopsis* Stock Centre (NASC, UK) and were described in Sarkar *et al.* (2007). *cle40-2* mutants were previously described (Stahl *et al.*, 2009). Homozygous *cle40-2*/*wox5-1* double mutants were generated via crossing and verified by genotyping.

Plant growth conditions

Plant growth conditions were previously described in Stahl *et al.* (2009).

Starch staining and microscopy

Starch granules were stained with the mPS-PI method described in Truernit *et al.* (2008) and imaged with a Zeiss LSM 510 confocal microscope.

Results and discussion

Spatial protein distribution is important for CSC patterning

The analysis of a network with several components and relationships requires at least some known parameter values. While a mathematical model can be used to predict the values of variables (e.g. concentration of CLE40), parameters are set ahead of time and remain constant throughout simulations. Examples of parameter values include production and degradation rates of CLE40. A model may exhibit very different behaviours based on small changes in the value of a parameter. It is possible to fit a model of a network to data and determine likely values of some of the parameters from the fitting (Gutenkunst *et al.*, 2007). To model CSC regulation, it would be useful to find the concentrations of WOX5, CLE40 and the other factors affecting CSC fate, but measuring the concentrations of WOX5 and CLE40 *in vivo* has proven to be a technical challenge. Consequently in this case, it would be best to use the simplest model with the fewest possible regulatory factors to elucidate the roles of WOX5 and CLE40 in cell fate determination. This way, the number of parameters is manageable, and it is possible to analyse the model output for many different sets of parameter values. This leads to knowledge about the model's general behaviour and, consequently, a possibility of insight into the function of the network emulated by the model.

To that end, we first developed a simple single-cell model, where the fate of a cell is controlled entirely by the interactions and different concentrations of a *WOX5*-derived signal (W) and the signalling peptide CLE40 (C). To keep the model simple, we assumed that everything necessary for the function of W and C (e.g. RLKs necessary for signal transmission) was abundant. In order to incorporate QC, CSC and CC fates into a mathematical model, numerical representatives of cell fate were required. These representatives could then be used to specify production rates of the two factors W and C, allowing us to require C to be produced by CCs, but not by CSCs or the QC. The representatives could also determine the fate of a cell, so that cells with insufficient concentrations of W differentiate into CCs as they do under experimental conditions. We designated variables F_{CC} and F_{QC} as the representatives, with high values of F_{CC} representing CC fate and high values of F_{QC} representing QC fate. Both F_{CC} and F_{QC} vary between 0 and 1 (Fig. 2) as a function of W. Parameters W_a and W_b are the half-maximum values of F_{CC} and F_{QC}, respectively. The value of F_{QC} increases with W concentration, simulating the WOX5-dependent aspects of QC fate. A cell with $W>W_b$ would be categorized as a QC cell. Low values of W yield higher values of F_{CC}, simulating differentiation to CC fate when W concentration is insufficient for CSC maintenance. We would categorize a cell with $W<W_a$ as having CC identity and a cell with any value of W between W_a and W_b as having CSC identity.

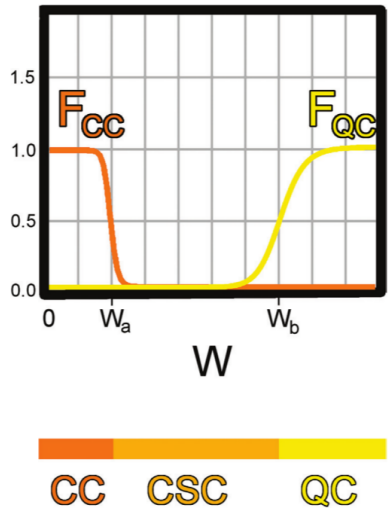


Fig. 2. Values of W determine cell fate, represented by F_{CC} and F_{QC}. The numerical representatives of CC and QC fate (F_{CC} and F_{QC} respectively), were modelled as Hill functions dependent on W. Parameters W_a and W_b determine their half-maximum values. Cells with values of $W<W_a$, $W_a<W<W_b$, and $W>W_b$ would be classified as CC, CSC and QC cell, respectively.

F_{QC} as the representative of QC fate determines the production rate of W. To simulate production of W only in QC cells, the W production rate increases as F_{QC} increases (blue in Figs 3 and 4). Likewise, F_{CC} determines the production rate of C. The production rate of C increases as F_{CC} increases to simulate CLE40 being produced only in CCs (teal in Figs 3 and 4). A network diagram of the interactions is shown in Fig. 3 with the corresponding equations.

A table of parameter symbols and definitions can be found in Table 1. The parameter values have units, but the individual parameter values are not reliable for comparison with measurements. Determining appropriate parameter values would require numerical data, and the experimental results used to calibrate the model are qualitative in nature.

To explore what would happen to cells with various initial values of W and C, the model given by the equations in Fig. 4 was implemented in a Matlab program (Supplementary File S1). It includes a graphical user interface so that a user can easily edit the parameter values, give initial values for W and C in the cell and see the model output. We compiled several results from different sets of initial values (Fig. 4A). The same was done for a different set of parameter values (Fig. 4B), which illustrates how the model predictions can change when parameter values are altered. The solutions are shown on top of a vector diagram. The vectors in the vector diagram point in the direction of change of W and C. Values of W and C start at the given initial value (blue squares in Fig. 4) and change over time in the direction of the arrows until they reach a stable fixed point (black dots in Fig. 4). The stable fixed point comprises the coordinates of W and C values at a stable equilibrium solution (i.e. cell fate). The results of this model yield at most two stable fixed points (or fates) for the cell, one with no W and a fixed amount of C (CC fate), and another with a small amount of C and a large amount of W (QC fate). When these two stable fixed points are present, there is always an unstable fixed point as well within the CSC domain (grey dot in Fig. 4A). However, we do not count the unstable fixed point as CSC fate. In order for a fixed point to represent a cell fate, values of W and C must be able to settle on that fixed point. Unlike the stable fixed points, which have arrows pointing toward them, unstable fixed points only have arrows pointing away; solutions that start close to a fixed point always flow away from it over time, never toward it (green box in Fig. 4). Therefore, it is highly unlikely for a solution to settle on an unstable fixed point, because only a cell with initial values of W and C exactly on that fixed point would remain there over time. A cell with values of W and C even slightly different from those values (or upon random fluctuations, which have not been integrated into the model) would settle instead on one of the stable fixed points. In summary, there are three fixed points including one representing CC fate, one representing QC fate and one that is unstable and cannot represent a cell fate; CSC fate is therefore not represented by our single-cell model.

A C/W multi-cell model achieves stem cell homeostasis

A regulatory network like the one involving W and C, where there are two network components acting antagonistically to each other, is called a switch. A switch forces a decision between two states, because it causes an increase in concentration of one component to result in a decrease in the other, with the eventual outcome that one of them wins (Alon, 2007). In a single cell, this network forces a decision between two cell fates, and it cannot maintain stem cell homeostasis (which requires three cell identities). This behaviour is evident in the solutions of the single-cell model, where there is a fine line between the values of W and C that lead to differentiation and those that lead to a QC fate with no stable region for CSCs between (green box in Fig. 4). This behaviour is fundamentally different from the more robust behaviour of models of the shoot network, because a negative feedback loop controls WUS expression through CLV3, while a mutual repression acts between WOX5 and CLE40 (Brand *et al.*, 2000; Schoof *et al.*, 2000; Stahl *et al.*, 2009; Yadav *et al.*, 2011).

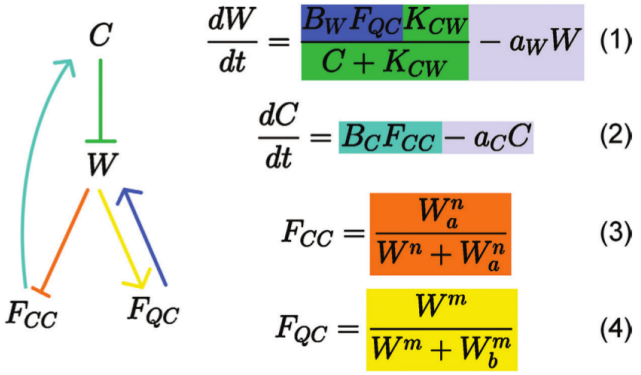


Fig. 3. Graphical and mathematical presentation of the relationships between C, W, F_{CC} and F_{QC}. Equations (1) and (2) determine the rate changes of W and C values, respectively, while equations (3) and (4) detail the Hill functions governing cell fate dependence on W. C inhibits W (green), while W promotes production of F_{QC} (yellow) and represses F_{CC} (orange). F_{QC} provides positive feedback to W (blue), and F_{CC} promotes C (teal). W and C are degraded at a constant rate (grey). B_w, K_{cw}, a_w, B_c, a_c, W_a, W_b, m and n are parameters.

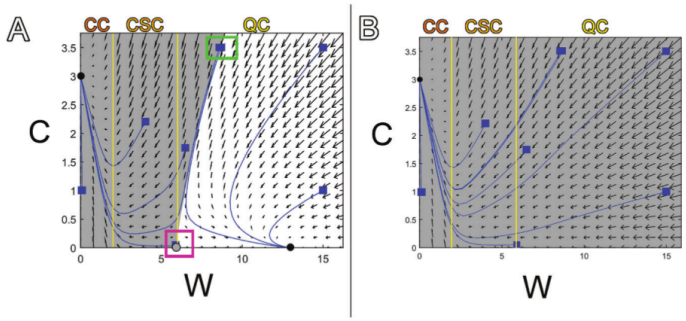


Fig. 4. Vector diagram of single-cell model results. (A) Fixed points are represented by black and grey filled circles, stable and unstable respectively. Solutions (blue lines) over time were obtained from various initial values (blue squares). Initial values chosen close to the unstable fixed point (like the initial value in the pink box) yield solutions that flow away from that fixed point. The basin of attraction for the CC fixed point at coordinates (0, 3) is shown in grey, and the basin of attraction for the QC fixed point close to (13, 0) is shown in white. Two initial values close to one another but lying in separate basins of attraction (green box) lead to different solutions over time. Vertical yellow lines mark changes in cell fate (where $W=W_a$ and $W=W_b$). (B) A different set of parameter values (identical to the values used to produce Fig. 1A, but with decreased maximum production rate of W) yields only one fixed point (black dot), which is stable. All values of W and C are within its basin of attraction (grey).

Table 1. *Parameter, variable, and index symbols and definitions*

This is a comprehensive list for all models presented in this paper, so specification of which model uses each parameter, variable and index is provided in the Models column.

Symbol	Definition	Models
W, W_i	Concentration of WOX5-derived signal, concentration of W in cell i	All
C, C_i	Concentration of CLE40 peptide, concentration of C in cell i	All
X_i	Concentration of X in cell i	W/C/X
$F_{CC}, F_{CC,i}$	Numerical representative of CC fate, F_{CC} in cell i	All
F_{QC}	Numerical representative of QC fate	Single-cell
S_i	Sum of W_i and X_i	W/C/X
B_W, B_C, B_X	Maximum production rates of W, C, and X	All
a_W, a_C, a_X	Degradation rates of W, C, and X	All
K_{CW}	Value of C where W production rate is halved	All
K_{CX}	Value of C where X production rate is halved	W/C/X
W_a	Value of W between CC and CSC fate	All
W_b	Value of W between CSC and QC fate	Single-cell
S_a	Value of S_i needed to maintain CSC fate	W/C/X
n	Hill co-efficient which controls steepness of S_1 curve	All
m	Hill co-efficient which controls steepness of S_2 curve	Single-cell
$D_{w,i}, D_{c,i}, D_{x,i}$	Proportional to diffusion rate of W, C, X between cell i-1 and cell i	Multi-cells
$D_{w,i+1}, D_{c,i+1}, D_{x,i+1}$	Proportional to rate of diffusion of W, C, X between cell i and cell i+1	Multi-cells
i	Cell index	Multi-cells

Despite the inability of this model to capture all three cell fates, we wanted to keep the model as small as possible and refrain from adding regulatory components until we had exhausted other options. We were also interested to see if a network with only W and C was capable of describing cell fate. We therefore considered the possibility that this network could still describe stem cell homeostasis if it were implemented in several cells. Each cell could make a cell fate decision as in the single-cell model, but the spatial component could result in a CSC habitable zone where the constant flow of W and C from other cells would result in a tie for W and C regardless of the switch-like nature of the regulatory network. Sophisticated models of the shoot have used 2- and 3-D templates for shoot architecture (Heisler and Jönsson, 2007), as observation of the WUS maximum at the centre of the meristem requires at least two dimensions. In the root, the WOX5, CLE40 and cell fate gradients run in one dimension along the proximal-distal axis. This may make it possible to describe the 3-D effects of WOX5 and CLE40 on cell fate using a one-dimensional model.

To determine if the single-cell model implemented in several cells could describe experimental results, a modified version of the model was implemented in a virtual cell column, with the most proximal cell defined as the W-producing QC cell and the rest allowed to take on either CSC or CC fate. The fate of each cell would still depend on W via F_{CC} as in the single-cell model. With this C/W multi-cell model, W and C could diffuse through the cells, emulating inter-cellular communication. The value of C at the QC determines the production rate of W. The network diagram for the C/W multi-cell model is shown in Fig. 5. We also implemented an alternative model (Supplementary File S1), where values of C at each cell decreased the mobility of WOX5 through that cell, and the results of the alternative model matched the results of the C/W multi-cell model distal to the QC. The alternative C/W multi-cell model, where C regulates W mobility rather than production, is described by the following equations, where the B_W term in equation 11 (shown in Fig. 8) only applies in the QC (cell i=1). The equation for $F_{CC,i}$ is identical to equation 7 shown in Fig. 5.

$$\frac{dW_i}{dt} = B_w + \frac{K_{CW}}{C_{i-1} + K_{CW}} D_{W,i}\left(W_i - W_{i-1}\right) + \frac{K_{CW}}{C_i + K_{CW}} D_{W,i+1}\left(W_i - W_{i+1}\right) - a_w W_i$$

$$\frac{dC_i}{dt} = B_C F_{CC,i} + D_{C,i}\left(C_i - C_{i-1}\right) + D_{C,i+1}\left(C_i - C_{i+1}\right) - a_C C_i$$

Parameter values were adjusted by hand, and solutions were obtained for time values until equilibrium was reached. The values of W and C at equilibrium were used to evaluate cell fate in terms of F_{CC}. Parameter values were changed to simulate wild-type roots as well as *wox5*, *cle40*, and *wox5/cle40* full loss-of-function mutants, constitutively expressed WOX5, and externally applied CLE40p to wild-type roots. For example, to simulate *wox5* mutants, the production rate of W was set to 0. To simulate constitutive WOX5 expression, the production rates of W in every cell were set to the user-defined maximum W production rate. Addition of CLE40p was simulated by keeping the value of C at a user-defined constant at the QC, where C has its effect on W production. Since CLE40p is applied ectopically and usually in relatively large doses so that the root is flooded, a constant supply of C is more realistic than an increase in production or diffusion rate for C. We assume that the user specifies a high enough number that the level of C at the QC is significantly higher than wild type.

Several initial values of W and C in each cell were tested with the multi-cell models, and this did not change the equilibrium values for any of the realistic parameter sets (all production rates should be greater than or equal to zero, all other parameter values should be greater than 0; m and n should be 1 or greater; and W_a should be less than W_b).

C/W multi-cell model predictions are shown in Fig. 6. It was possible to find a range of parameter values that emulated the most common phenotypes of wild type, *wox5* and *cle40* mutants, WOX5 constitutive expression, and addition of CLE40p to wild type.

The C/W multi-cell model fails to explain wox5-1/cle40-2 *double mutants*

The predictive capabilities of a model can be assessed by a validation experiment, a test to see if the model will predict the outcome of an experiment that was not considered in the derivation of the model. To assess the reliability of the C/W multi-cell model, we performed an experiment on *wox5-1/cle40-2* double mutants and compared the results to the model prediction for *wox5/cle40* double mutants.

Surprisingly, a partial rescue of the *wox5-1* phenotype was observed in the *wox5-1/cle40-2* double mutants (Fig. 7). Since the C/W multi-cell model assumes that W, a WOX5-dependent signal, is necessary for CSC maintenance, it is impossible for the model to correctly predict this result; if there is no WOX5, there is no W, and according to the model there are no CSCs present. The failure of the model indicates that a basic assumption was wrong: WOX5 is not absolutely necessary for CSC maintenance.

Since loss of CLE40 restores CSC fate in *wox5-1* mutants, CLE40 must have an effect on CSC fate that is independent of WOX5. Since biologically active CLE40 is localized to the intercellular space, it cannot directly regulate the processes in the nuclei of target cells to affect their fate; there must be at least one additional component in the CSC regulatory network that affects stem cell fate and is affected by CLE40. Since CLE40 represses CSC fate through this unknown component X, there are two possibilities: either CLE40 promotes expression of X while X represses CSC fate, or CLE40 represses expression of X while X promotes CSC fate from the QC. Neither hypothesis contradicts any experimental results, but the second is simpler. For the first possibility, an ablated QC would result in less WOX5, more differentiation, more CLE40, more X, and further repression of CSC fate by X. If the second possibility were the case, an ablated QC would result both in less WOX5 and less X. Since the latter is the simpler explanation, we tested it further.

A C/W/X multi-cell model is plausible

We designed another model, the C/W/X multi-cell model, where X was included with the same role and relationships as W in the C/W multi-cell model (Fig. 8). This C/W/X multi-cell model is able to describe the most common outcomes of all of the experimental results to which we have access, including that of *wox5-1/cle40-2* mutants (Fig. 9). We then used the C/W/X model to predict the phenotype of *wox5/x* mutants. Since the model assumes that either W or X is necessary to maintain CSC fate, those roots are expected to have no CSCs. The model predicts that the number of CSC layers would depend on the rate of *WOX5* expression, so that increasing the rate of W production while still keeping W expression restricted to the QC would result in more stem cell layers due to a higher flux of W to distal cells. In the C/W/X multi-cell model, X functions redundantly with WOX5, perhaps protecting the pluripotent nature of the QC and keeping it from differentiating when WOX5 levels are low. The phenotypes of *wox5-1* mutants suggest that X should be less abundant or less effective at CSC maintenance than the WOX5-dependent signal (W) in the presence of CLE40, since *wox5-1* mutants have less CSCs than *wox5-1/cle40-2* mutants. We simulated this in the model by making the production rate of X less than that of W. Despite the smaller direct effect of X on CSC fate, it had a significant impact on sensitivity of the model to changes in parameters controlling W (Table 2). Due to the buffering activity of X, the model was less sensitive to the production and diffusion rates of W. We conclude that the activity of X allows the network controlling CSC fate to be more robust to perturbations in the levels of WOX5. Starch granules can be found in the QC infrequently in *wox5-1* mutants, but frequently when a substantial amount of synthetic CLE40p is added to *wox5-1* mutants. The model would suggest that the CLE40p is negatively affecting X in this case.

We now conclude that the roles of WOX5 and CLE40, given their mutually antagonistic nature, are to function as

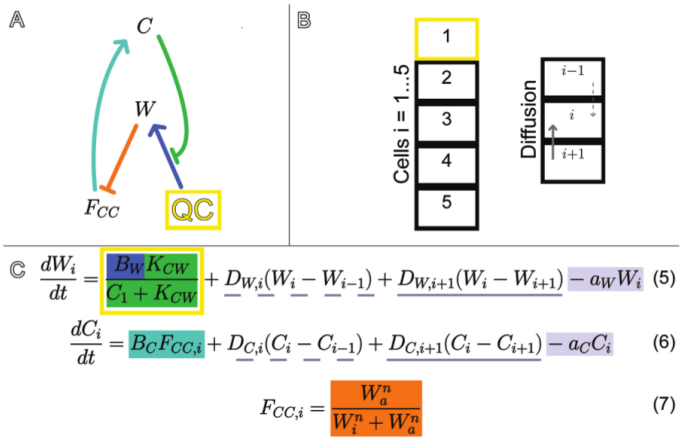


Fig. 5. Graphical and mathematical representations of the C/W multi-cell model of cell fate. (A) C represses W (green), which represses F_{CC} (orange). F_{CC} promotes C (teal). W production is confined to the QC cell (yellow). (B) The model simulates a cell column. The cell with index i=1 is the QC (yellow cell), while the model determines the fates of those distal to it. W and C can both diffuse through the cell column and the value of each in cell i are denoted W_i and C_i. (C) Equations 5, 6, and 7. The W production term in the yellow box only applies in the QC cell and W production is restricted by the amount of C signalling to the QC (green). Diffusion terms track the fluxes of W and C between cell i and its proximal neighbour (dotted underline) and between cell i and its distal neighbour (solid underline). A diffusion term is omitted if an index is outside of 1 through 5; there is no flow between the QC and the cell proximal to it and no flow between the 5th cell and the cell distal to it. The value of F_{CC} in cell i ($F_{CC,i}$, orange) determines the fate of cell i and the production rate of C (teal) in cell i. W and C are degraded at constant rates (grey).

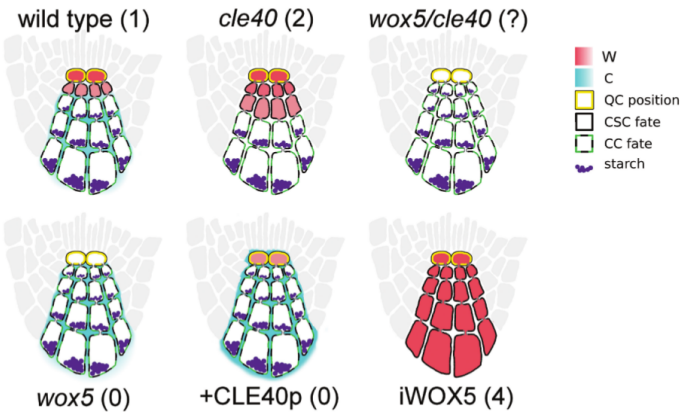


Fig. 6. Representation of C/W multi-cell model predictions of CSC fate and W and C localization. The number of expected CSC rows, based on experimental results is shown in parentheses. The model is capable of emulating the results of root phenotypes of wild-type roots, *wox5* mutant roots *cle40* mutant roots, roots with additional synthetic CLE40 (CLE40p) added, and roots expressing *WOX5* in all cells through constitutively expressed *WOX5* with inducible WOX5 function (iWOX5). Results of a *wox5/cle40* double mutant phenotype were also predicted so that the model could be tested later for predictive ability.

triggers for switches in cell fate. Their interaction does not maintain the robustness of CSC patterning, but discourages cells from staying in a state between CSC and CC fate. As evident from the results of the single-cell model, CSC fate may not be possible without the inter-cellular signalling conducted by WOX5 and CLE40. This result parallels that of the model of the shoot meristem by Yadav *et al.* (2011), which showed the importance of WUS mobility and its resulting gradient to the regulation of stem cell number. Furthermore, it was previously suggested that WOX5 was necessary for CSC fate, so that in *wox5* mutants, CSCs could exist only transiently just after a QC cell division (Sarkar *et al.*, 2007). Using the results of an experiment on *wox5-1/cle40-2* mutants and the inability of the C/W multi-cell model to emulate those results, we have determined that stem cells can be maintained without WOX5 in the absence of CLE40 signalling and, equivalently, that WOX5 is not absolutely necessary for CSC homeostasis. Using the C/W/X multi-cell model, we determined that the existence of another stem-cell promoting factor within the WOX5/

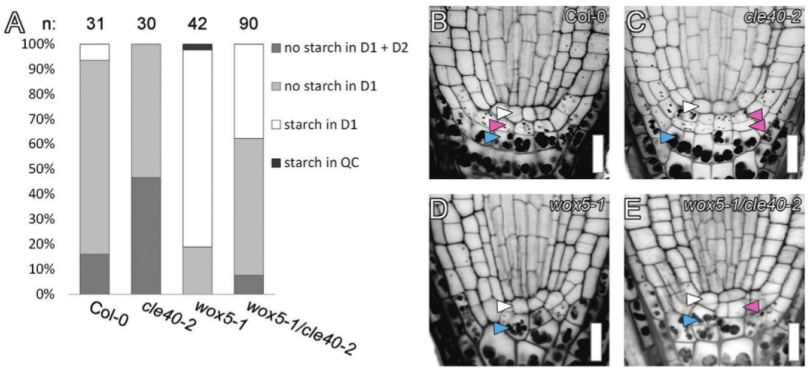


Fig. 7. CSC maintenance in Col-0 and *wox5-1*, *cle40-2* and *wox5-1/cle40-2* mutants. (A) Wild-type (Col-0) roots typically maintain stem cells in the D1 layer. Roots lacking CLE40 accumulate more stem cells than wild type, as *cle40-2* mutants more frequently lack starch in D2 as well as D1. Roots lacking WOX5 accumulate less stem cells than the wild type, with a higher frequency of starch in D1. Roots lacking both WOX5 and CLE40 partially rescue the *wox5-1* mutant phenotype. Typical phenotypes are shown for (B) Col-0, and mutants (C) *cle40-2*, (D) *wox5-1* and (E) *wox5-1/cle40-2*. White arrows indicate the QC position, pink indicates starch-free cell layers, and blue indicates the first layer with starch. Bars, 30 μm.

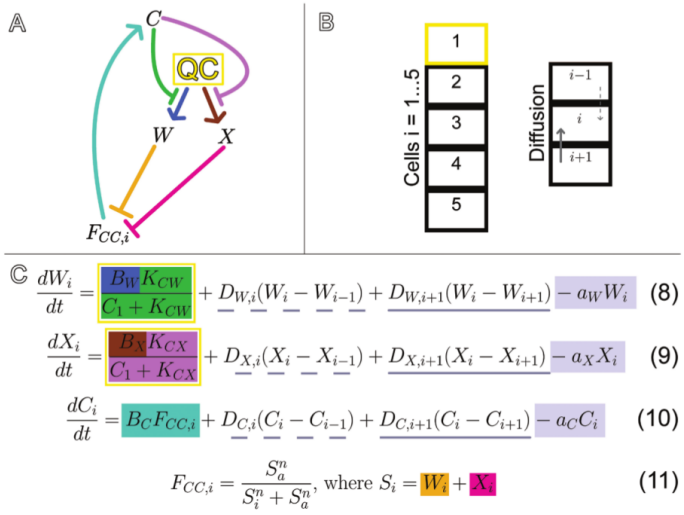


Fig. 8. Graphical and mathematical representations of the C/W/X multi-cell model. The C/W/X model was constructed by adding X to the C/W multi-cell model and giving it the same role as W. (A) C represses W (green) and X (purple). W represses F_{CC} (orange), and X represses F_{CC} (pink). F_{CC} promotes C (teal). Production of W and X is confined to the QC cell (yellow). (B) The model simulates a cell column. The cell with index i=1 is the QC (yellow cell), while the model determines the fates of those distal to it. W, X and C can all diffuse through the cell column and the value of each in cell i are denoted W_i, X_i and C_i. (C) Equations 8, 9, 10, and 11. The production terms of W and X in the yellow box only apply in the QC cell and production of W and X is restricted by the amount of C signalling to the QC (green). Diffusion terms track the fluxes of W, X and C between cell i and its proximal neighbour (dotted underline) and between cell i and its distal neighbour (solid underline). A diffusion term is omitted if the value of i is outside of 1 through 5; there is no flow between the QC and the cell proximal to it and no flow between the 5th cell and the cell distal to it. The value of F_{CC} in cell i ($F_{CC,i}$) is determined by the sum of W_i and X_i. $F_{CC,i}$ determines the fate of cell i and the production rate of C (teal) in cell i. W and C are degraded at constant rates (grey).

CLE40 regulatory network of CSC maintenance is plausible. In the C/W/X model, stem-cell promoting X functions redundantly with WOX5. It would be interesting to see if CLE40 affects SCR expression. Like X, *SCR* expression is independent of WOX5. SCR is known to function redundantly with WOX5 in stem cell maintenance, though to date this has only been demonstrated in the initials proximal to the QC; lack of CSCs is a phenotype of *scr-1* mutants, but this can be explained by lack of WOX5 (Sarkar *et al.*, 2007). Whether X can be found among already-known components of the root stem cell regulatory network (Bennett *et al.*, 2014) like SCR, or among known transcriptional targets of CLE40 (Pallakies and Simon, 2014) remains to be experimentally determined.

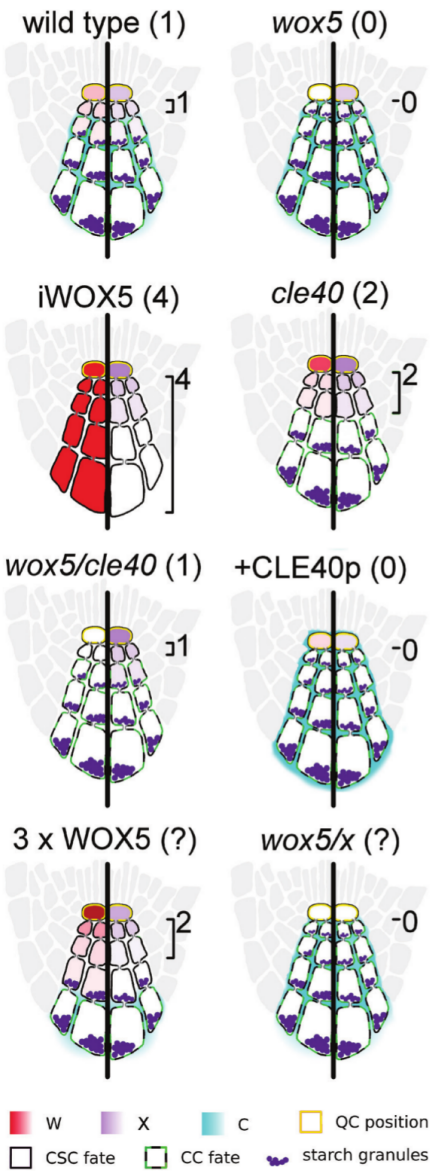


Fig. 9. Representation of C/W/X multi-cell model predictions of CSC fate and C/W/X localization. Gradients of W as predicted by the model are shown on the left sides of roots, while X gradients are shown on the right. Expected number of CSC rows, based on experimental results, is shown in parentheses, while those based on model predictions are shown next to the root diagram. The model outcomes can emulate the expected number of rows for wild type, *wox5* mutants, *cle40* mutants, *wox5/cle40* double mutants, constitutively expressed *WOX5* (iWOX5), and the addition of sufficient amounts of CLE40p (+CLE40p). We have also used the model to predict results of experiments that have not yet been completed to aid in later validations of this model. Tripling the W production rate (3× WOX5) is expected to yield more layers of stem cells. In the case of these parameter values, it is expected to result in two layers.

Table 2. *Parameter values used for simulations*

Dashes indicate where a parameter was absent from a particular model. Parameters can be perturbed by the amount specified (with all of the others held at the given value), without changing the number and stability of fixed points for the single-cell model or the number of rows of CSCs in each experimental condition for the multi-cell models.

Parameter	1-cell	C/W	C/W/X
B_w	13±31%	34±35%	34±47%
B_c	3±67%	30±67%	45±67%
B_x	-	-	25±60%
k_{cw}	2±50%	2±50%	2±50%
k_{cx}	-	-	2±50%
a_w	1±50%	2±15%	2±13%
a_c	1±50%	2±25%	2±25%
a_x	-	-	2±13%
n	10±90%	2±50%	2±50%
m	10±90%	-	-
W_a	2±75%	1±25%	-
W_b	6±58%	-	-
S_a	-	-	1±20%
D_w	-	1.2±8%	1±50%
D_c	-	1±50%	1±25%
D_x	-	-	1±50%

Supplementary data

Supplementary data is available at JXB online.

Supplementary File S1. Text file containing a Matlab program and instructions for its use. The program displays a graphical user interface (GUI), where parameter values can be adjusted and representations of the model output can be acquired.

Supplementary Files S2–S6. Image files required to run the Matlab program.

Acknowledgements

We thank Achim Schädle for helpful discussions during model development, Georg Jansing for tips on writing code for the interactive Matlab GUI, the Center of Advanced Imaging (CAi) at Heinrich Heine University for microscope maintenance, Cornelia Gieseler, Silke Winters and Carin Theres for technical support, and Frédéric Boyer, Nadia Heramvand, Nima Abedpour, Christopher Blum, Barbara Berckmans, Avantika Jakati and Yvonne Stahl for critical discussions of the manuscript. RS conceived the model interactions to be studied. RS and RHW designed the experiments, and RHW carried them out. SR derived and evaluated the model equations, and SR and RS wrote the paper. All authors discussed the results and the manuscript. This work was supported by the Deutsche Forschungsgemeinschaft (DFG) (grant number GRK1525 to RS).

References

Alon U. 2007. Network motifs: theory and experimental approaches. *Nature Reviews Genetics* **8,** 450–461.

Bennett T, van den Toorn A, Willemsen V, Scheres B. 2014. Precise control of plant stem cell activity through parallel regulatory inputs. *Development* **141,** 4055–4064.

Ding Z, Friml J. 2010. Auxin regulates distal stem cell differentiation in *Arabidopsis* roots. *Proceedings of the National Academy of Sciences* **107,** 12046–12051.

Brand U, Fletcher JC, Hobe M, Meyerowitz EM, and Simon R. 2000. Dependence of stem cell fate in *Arabidopsis* on a feedback loop regulated by CLV3 activity. *Science* **289,** 617–619.

Gutenkunst RN, Waterfall JJ, Casey FP, Brown KS, Myers CR, Sethna JP. 2007. Universally sloppy parameter sensitivities in systems biology models. *PLoS Computational Biology* doi: 10.1371/journal.pcbi.0030189

Heisler MG, Jönsson H. 2007. Modelling meristem development in plants. *Current Opinion in Plant Biology* **10,** 92–97.

Pallakies H, Simon R. 2014. The CLE40 and CRN/CLV2 signaling pathways antagonistically control root meristem growth in *Arabidopsis*. *Molecular Plant* **7,** 1619–1636.

Sarkar AK, Luijten M, Miyashima S, Lenhard M, Hashimoto T, Nakajima K, Scheres B, Heidstra R, Laux T. 2007. Conserved factors regulate signaling in *Arabidopsis thaliana* shoot and root stem cell organizers. *Nature* **446,** 811–814.

Schoof H, Lenhard M, Haecker A, Mayer KF, Jürgens G, Laux T. 2000. The stem cell population of Arabidopsis shoot meristems is maintained by a regulatory loop between the CLAVATA and WUSCHEL genes. *Cell* **100,** 635–644.

Sharma VK, Carles C, Fletcher JC. 2003. Maintenance of stem cell populations in plants. *Proceedings of the National Academy of Sciences* **100,** 11823–11829.

Stahl Y, Simon R. 2005. Plant stem cell niches. *International Journal of Developmental Biology* **49,** 479–489.

Stahl Y, Wink RH, Ingram GC, Simon R. 2009. A signaling module controlling the stem cell niche in *Arabidopsis* root meristems. *Current Biology* **19,** 909–914.

Stahl Y, Grabowski S, Bleckmann A, *et al*. 2013. Moderation of *Arabidopsis* root stemness by CLAVATA1 and ARABIDOPSIS CRINKLY5 receptor kinase complexes. *Current Biology* **23,** 362–371.

Tian H, Wabnik K, Niu T, *et al*. 2014. WOX5-IAA17 feedback circuit mediated cellular auxin response is crucial for the patterning of root stem cell niches in *Arabidopsis*. *Molecular Plant* doi: 10.1093/mp/sst118

Truernit E, Bauby H, Dubreucq B, Grandjean O, Runions J, Barthélémy J, Palauqui JC. 2008. High-resolution whole-mount imaging of three-dimensional tissue organization and gene expression enables the study of phloem development and structure in *Arabidopsis*. *The Plant Cell* **20,** 1494–1503.

van den Berg C, Willemsen V, Hendriks G, Weisbeek P, Scheres B. 1997. Short-range control of cell differentiation in the *Arabidopsis* root meristem. *Nature* **390,** 287–289.

Wang JW, Wang LJ, Mao YB, Cai WJ, Xue HW, Chen XY. 2005. Control of root cap formation by microRNA-targeted auxin response factors in *Arabidopsis*. *The Plant Cell Online* **17,** 2204–2216.

Willemsen V, Bauch M, Bennett T, Campilho A, Wolkenfelt H, Xu J, Scheres B. 2008. The NAC domain transcription factors FEZ and SOMBRERO control the orientation of cell division plane in *Arabidopsis* root stem cells. *Developmental Cell* **15,** 913–922.

Yadav R, Perales M, Gruel J, Girke T, Jönsson H, Reddy G. 2011. WUSCHEL protein movement mediates stem cell homeostasis in the *Arabidopsis* shoot apex. *Genes & Development* **25,** 2025–2030.

8.2 Publication II: Natural selection on LysM-RLKs (LYKs/LYRs) in wild tomatoes and phylogenetic analysis in Angiosperms

Authors: Sarah Richards, Laura Rose

This manuscript was submitted to *BMC Evolutionary Biology* in the year 2018.

Contribution of Sarah Richards:

Major

- Obtained the sequences (except the wild tomato LysM-RLK orthologs)
- performed the phylogenetic and population genetic analyses
- developed most of the hypotheses
- tested the hypotheses
- drafted the manuscript
- created and edited the figures

Title Natural selection on LysM-RLKs (LYKs/LYRs) in wild tomatoes and phylogenetic analysis in Angiosperms

Authors
Sarah Richards[1] (sarah.richards@hhu.de), Laura Rose (laura.rose@hhu.de)[1,2,3*]
*: Corresponding Author
Address
1: Institute of Population Genetics, Heinrich Heine University, Duesseldorf, Universitaetsstr. 1, 40225 Duesseldorf, Germany
2: iGRAD-Plant Graduate School, Heinrich Heine University, Duesseldorf, Universitaetsstr. 1, 40225 Duesseldorf, Germany
3: CEPLAS, Cluster of Excellence in Plant Sciences, Heinrich Heine University, Duesseldorf, Universitaetsstr. 1, 40225 Duesseldorf, Germany

Abstract:

Background: The LysM receptor-like kinases (LysM-RLKs) are vital to both plant defense and symbiosis. Previous studies described three clades of LysM-RLKs: LysM-I/LYKs (10+ exons per gene, conserved classical kinase domains), LysM-II/LYRs (1-5 exons per gene, lacking conserved kinase domains), and LysM-III (two exons per gene, conserved classical kinase domains unlike other LysM-RLK kinase domains, restricted to legumes). LysM-II gene products are presumably not functional as conventional receptor kinases, but several are known to operate in complexes with other LysM-RLKs. The first aim of our study was to utilize recently mapped wild tomato transcriptomes to uncover evidence of natural selection on LysM-RLKs in wild tomato populations. The second was to put this information into a functional context using a combination of phylogeny and known functions of better-studied LysM-RLKs.

Results: We discovered new genes within the LysM-III clade in diverse Rosid species, including *Eucalyptus grandis*. Our maximum likelihood phylogeny of LysM-RLKs from diverse plant species supports a previously postulated closer relationship between LysM-II and LysM-III clades. We found intact kinase domains in *Solanum lycopersicum* LYK8 orthologs in two wild tomato species. A small clade within LysM-I has distinct LysM domains that do not align well with those of other LysM-RLKs. The clade includes *Sl*LYK3, whose orthologs in wild tomatoes showed signs of strong purifying selection in the extracellular domain (where the LysM domains are found), unlike the other wild tomato LysM-RLK orthologs.

Conclusions: The LysM-III genes originated before the divergence of *Eucalyptus* from other Rosids. Due to evidence of selection, its position in a clade of LysM-RLKs with distinct LysM domains, and its close phylogenetic relationship to the versatile *Arabidopsis thaliana* LYK3, *Sl*LYK3 is an especially interesting candidate for further study.

Key Words:
phylogenetics, population genetics, Solanum, plant immunity, symbiosis

Background

Plants are regularly targeted for colonization by organisms ranging from pathogenic to symbiotic. Pathogenic organisms infect the plant and use it for nourishment, eventually causing damage or reduced host fitness. Pathogens often cause changes in the host plant's genetic regulation to maximize the amount of nutrients that can be accessed and to avoid detection and subsequent host defense responses. Symbiotic organisms also use the host plant for nourishment but offer benefits for the plant in exchange (e.g. better uptake of water and nutrients). Plants that can differentiate between the two improve their chances of survival and reproduction. Detection of the presence of these organisms involve extracellular receptors, often receptor-like kinases (RLKs), which bind proteins or other molecular "patterns" produced by the colonizing organisms and trigger phosphorylation and downstream signaling cascades within the cell [1,2]. The signaling cascades then cue the appropriate defense or symbiosis response [2].

Genes containing the LysM motif, including the family of LysM-RLKs, have been implicated in the detection of both plant-symbiotic and -pathogenic organisms. In the case of symbiotic organisms, some LysM-RLKs are an integral part of the signaling necessary for the plant's cooperative activity with the symbiont to allow and encourage colonization. Other LysM-RLKs are necessary for the detection of pathogens and signaling for defense responses, which involves detection of molecules, such as chitin, which are shed by the invading pathogen [3]. LysM-RLKs sometimes function together as heterodimers, with the presence of the colonizing pathogen or symbiont detected by the extracellular region of one LysM-RLK, and the symbiotic or defense response mediated by the kinase domain of another [3,4]. Some LysM-RLKs, such as *Oryza sativa* CERK1 (*Os*CERK1), also function as dual-purpose detectors of both pathogenic and symbiotic organisms [5].

Shiu *et al.* describe two main clades of LysM-RLKs: LysM-I and LysM-II [6], with LysM-RLKs in Group II lacking conserved kinase domain sequences; the Glycine-rich loop is missing from the kinase domains of all *Arabidopsis thaliana* and *Solanum lycopersicum* LysM-RLKs in Group II [7,8]. In addition, LysM-I RLKs have ten or more exons, while LysM-II RLKs typically have one or two [8]. A group of two LysM-RLK genes each from *Medicago truncatula* and *Lotus japonicus*, which contain two exons and classically conserved kinase domain sequences, cannot be classified easily into either Group I or Group II, and their clade has been named Group III [9,10]. Phylogenetic analyses of LysM-RLKs have been conducted on a wide variety of plant species, but comprehensive phylogenetic analyses of this family have not included genes from tomato, and much has been discovered in the meantime about the functions of many individual LysM-RLKs.

Here we synthesize the currently known information about function and phylogenetic relationships of LysM-RLKs, including those from tomato, and show a closer relationship between Group II and Group III LysM-RLKs than was previously shown. Our work is based on maximum likelihood analysis of entire sequences and results in higher bootstrap support than previously published phylogenies. Newly discovered Group III LysM-RLKs are present outside of *M. truncatula* and *L. japonicus*. Wild tomato *Sl*LYK3 orthologs show evidence of strong purifying selection, and, although the kinase domain of *Sl*LYK8 is truncated in cultivated tomato, we find that this is not the case for orthologs in wild tomato species.

Results

Group III LysM-RLKs reliably cluster with Group II LysM-RLKs

Plants and colonizing microbes evolve together with LysM-RLKs functioning as key regulators of microbe detection. We aimed to uncover new insights from the synthesis of sequence and functional data. To this end, we constructed a phylogeny of LysM-RLK protein sequences from *A. thaliana*, *S. lycopersicum*, *L. japonicus*, *O. sativa*, and *M. truncatula* and combined it with the known functions of the proteins (Fig. 1) [4,7,10-24]. We differentiated between function implicated by gene regulation and function implicated by mutant phenotype, since – as in the case of *Lj*LYS11 – it is possible for a gene to be regulated by symbiosis or defense without the gene necessarily playing a role in symbiosis or defense [22]. While we do recover multiple small clades of closely related sequences reported to fulfill similar functions, in most cases, major clades do not appear to be strictly associated with a specific form of microbe discrimination, suggesting that orthologs of the same gene can function differently in different organisms. However, it should be mentioned that most of the genes have not been tested for each of the functions listed, and functions in the best-studied functional process - Rhizobia symbiosis – dominate the tree. This bias in functional characterization may limit our power to detect a correlation between the type of microbe recognition and phylogenetic position.

This phylogeny generally agrees with previously published phylogenies based on entire coding regions or kinase domains only, including the separation of the LysM-RLKs into two major clades: Group I and Group II and one smaller one: Group III. Group I is separated into two clades, in which the *L. japonicus* gene placements agree with Lohmann *et al.*'s findings [10] of the existence of a microexon in one clade and not in the other. Interestingly, the Group III LysM-RLKs are within a clade containing the Group II LysM-RLKs in all of the 500 bootstrap replicates. Furthermore, Group I LysM-RLKs formed a separate clade from Group II and III LysM-RLKs in every bootstrap replicate. This indicates that the sequences in Group II and III share strong sequence similarity. We further observed that Group III LysM-RLKs are grouped together in each of the bootstrap replicates, but the 45% bootstrap value on the branch subtending Group II LysM-RLKs makes it clear that the entire Group III clade often clusters within the Group II clade. A review of the bootstrap replicates shows that the Group III clade has no consistent placement at a specific place on the tree. Taken together, this strongly supports a closer relationship between Group II and Group III LysM-RLKs than either clade has with the Group I LysM-RLKs, and it does not support the idea that Group III forms its own clade outside of Group II.

Group III LysM-RLKs exist in diverse Rosid species

Previous publications report Group III LysM-RLKs only in *L. japonicus* and *M. truncatula*. To elucidate the evolutionary history of the Group III LysM-RLKs, we found putative homologs to the

known Group III LysM-RLKs (*Lj*LYS20, *Lj*LYS21, *Mt*LYR5, and *Mt*LYR6) using BLAST searches [25] of each of these genes against NCBI's non-redundant protein sequences database [26]. The group of putative homologs is well-represented in the Rosids, and a phylogeny of the genes generally agrees with major clades, with genes from Fabids and Malvids grouped largely according to their species' evolutionary history [27] (Fig. 2 and Figure S1). The notable exceptions are the genes from species within the Malpighiales order. These species share a closer relationship with the Fabids than the Malvids, but their LysM-RLKs were grouped with low bootstrap support with the Malvids in our phylogeny, indicating an ambiguous relationship.

This analysis resulted in two to five potential homologs per species, with one exception: *Eucalyptus grandis* (*E. grandis*), whose sequences form their own clade of fifteen members distinct from all of the other putative homologs, matching the species phylogeny (Fig. 2). Some of these sequences are isoforms of the same gene, while others, reported under different naming schemes (see Table 1), have high similarity to other conspecific sequences. This may indicate detection of different isoforms of the same gene by different gene-finding algorithms. With four to five duplicate pairs of genes in the set of sequences, this would mean that *E. grandis* has ten or eleven unique genes represented here. One whole genome duplication occurred in the *Eucalyptus* lineage that did not occur in the *Citrus* and *Theobroma* lineages [28], so we would expect *E. grandis* to have four to ten putative homologs (double the two to five found in the other species). This makes the larger number of putative *E. grandis* homologs consistent with what would be expected.

The *E. grandis* putative homologs had approximately 55% identity to the query sequences, which was lower than the identity found for putative homologs in other species. It was possible that these *E. grandis* genes were a more distantly-related set of genes, and that they do not belong to Group III. To verify the *E. grandis* genes' position, we constructed a new tree from the *E. grandis* and LysM-RLK sequences described above from *S. lycopersicum*, *A. thaliana*, *L. japonicus*, *M. truncatula*, and *O. sativa* (Figure S2). The *E. grandis* sequences are included with strong bootstrap support within the Group III clade, just outside of the *L. japonicus* and *M. truncatula* Group III LysM-RLKs. In addition, all of the *E. grandis* sequences possess the Glycine-rich loop. Theirs are, however, distinct from and more varied than the glycine rich loops of the Group III LysM-RLKs from *L. japonicus* and *M. truncatula*. *L. japonicus* and *M. truncatula* Group III LysM-RLKs all have QGGY as their Glycine-rich loop, while the *E. grandis* sequences have EGGF, HGGF, or QGGF. The most-numerous, QGGF, is also common in Group I LYKs. We conclude that these *E. grandis* genes and the rest of those in our BLAST search are Group III LysM-RLKs.

*Sl*LYK3, *At*LYK3, *Lj*LYS4, and *Lj*LYS5 LysM domains are distinct from those of other LysM-RLKs

The ability of a LysM-RLK to detect ligands depends on its three extracellular LysM domains, and relationships between LysM-RLKs with similar functions may be better revealed by a phylogeny of their LysM domains than by a phylogeny of only their kinase domains or entire sequences. To reconstruct the history of the LysM domain sequences, we constructed a maximum likelihood tree of the individual LysM domain sequences (LysM1, LysM2, and LysM3) from each known LysM-RLK gene of *S. lycopersicum*, *A. thaliana*, *L. japonicus*, *M. truncatula*, and *O. sativa* (Fig. 3 and Figure S3). The small sequence lengths and substantial variation between the individual LysM domain sequences led to a generally unreliable tree featuring many bootstrap values in the single digits. Since the three-LysM-domain structure is ancient [3], we expected a tree with three distinct clades consisting of sequences from the first, second and third domains. However, this phylogeny did not show three monophyletic clades according to domain position as expected. Instead, there is substantial mixing of the first and third domain sequences, which is unlikely to represent the true evolutionary history of these sequences. Further suggesting potential problems with the phylogeny, branches leading to some groups of domains are long compared to the rest. Four sequences stand out with all three sets of LysM domains clustered together with bootstrap values near 90%. The LysM domains of *Sl*LYK3, *At*LYK3, *Lj*Lys4, and *Lj*LYS5 are found grouped together with long branches leading to clades of their first two LysM domains, indicating that these sequences are substantially different from the rest, but similar enough to each other to warrant reliable clustering based upon only the 40-66 nucleotides of sequence used to build the alignment and phylogeny.

We suspected that the long branches subtending this clade could indicate a problem with the alignment and tree, and that these sequences may have been difficult to align with the rest of the LysM-RLKs. A GUIDANCE [29] analysis of first, second, and third domains separately showed that the third domain sequences were generally aligned similarly regardless of alignment method. In contrast, the low-scoring sequences – the first domains of all of the proteins as well as the second domains of the previously described group *Sl*LYK3, *At*LYK3, *Lj*Lys4, and *Lj*LYS5 and another group *Os*LYK1, *Sl*LYK14, *Lj*LYS3, and *Mt*LYK10 - aligned unreliably, with bias in the alignment method strongly affecting the alignment and the trees upon which they are based. This means that the long branches we noted in the domain tree were caused by the nature of the sequences, and they probably do not accurately reflect the evolutionary history of those sequences. Logos of LysM domains of the two groups of problematically-aligned sequences as well as the rest of the LysM-RLKs shown separately demonstrate the issue visually (Fig. 4). Sequence similarity at conserved amino acids is clearly visible between these two groups and the rest of the genes in their third domain sequences, while these sets of sequences are simply very distinct in the first and second domains. The phylogeny of entire proteins (Fig. 1) shows that the two groups of proteins share a close relationship. They group together in 99% of the 500 bootstrap replicates.

If the analysis includes only the sequences which can be robustly aligned, more reliable inferences can be made about the sequences that remain (Figure S4). The bootstrap values are higher; there are no long branches leading to small clades; and the domains form monophyletic clades according to domain position. The last point is, however, trivial, since there are no longer first domain sequences included in the phylogeny.

Wild Tomato *Sl*LYK3 Orthologs Have Undergone Purifying and Possibly Directional Selection

To evaluate the evolutionary history of the LysM-RLKs on a finer scale, we investigated the patterns of polymorphism and divergence in LysM-RLK genes in *Solanum chilense* and *Solanum peruvianum*, two recently-diverged wild tomato species. We first calculated standard population genetic summary statistics for these genes (Table 2) in the species of interest, using LysM-RLK sequences from *Solanum ochranthum* and *Solanum lycopersicoides* as outgroups. After a stringent filtering step for coverage, we retained allelic sequences of the orthologs of six LysM-RLKs. The non-synonymous nucleic polymorphism at the LysM-RLK genes ranged from 0.11% (LYK3) to 1.19% (LYK6) in *S. chilense* and 0.08% (LYK3) to 1.4% (LYK6) in *S. peruvianum*. The synonymous nucleic polymorphism at the LysM-RLK genes ranged from 0.64% (LYK8) to 2.5% (LYK6) in *S. chilense* and 0.87% (LYK8) to 2.4% (LYK6) in *S. peruvianum*. These values are consistent with the overall non-synonymous (0.18%) and synonymous (1.27%) rates of polymorphism in *S. chilense* and non-synonymous (0.22%) and synonymous (1.69%) rates of polymorphism in *S. peruvianum* [30]. To determine whether selection had differential effects on the intracellular or extracellular domains, we calculated the summary statistics for each of these domains separately (Table 3). The ratio of non-synonymous (π_a) and synonymous (π_s) pairwise differences is often used to gauge the impact of natural selection on sequences. For most genes, π_a/π_s was higher in the extracellular domain than in the intracellular/kinase domain. Orthologs of *Sl*LYK3 stand out as the only set of LysM-RLK genes in our analysis with comparable or lower π_a/π_s in their extracellular domain sequences than those of the intracellular domain. This indicates that the sequences coding for the *Sl*LYK3 orthologs' extracellular domains have undergone strong purifying selection relative to those of other LysM-RLKs.

We then applied two standard tests of neutrality to determine if the pattern of genetic variation deviated from neutral expectations. According to Tajima's D, no deviation from neutrality could be detected at these six genes. For three loci, *Sl*LYK1, *Sl*LYK3, and *Sl*LYK8, the McDonald-Kreitman test indicated deviations from the neutral expectation in several comparisons (Table 2). For the SlLYK3 analysis comparing *S. peruvianum* and *S. ochranthum*, the ratio of non-synonymous to synonymous fixed differences was higher than that of the polymorphisms, indicating directional selection. However, after correcting for multiple testing, the corrected p-value exceeded a significance threshold of 0.05. After 22 tests, the Šidák correction [31] requires a p-value of 0.0023 or less for a 5% significance threshold. The lowest p-value for an individual test – *Sl*LYK3 with p=0.00507 - is

equivalent to a 10-11% p-value after this correction. This is a weak indication that *Sl*LYK3 orthologs have undergone directional selection between *S. peruvianum* and *S. ochranthum*.

*Sl*LYK8 orthologs with intact kinase domains exist in wild tomatoes

In our analysis of intracellular and extracellular π_a/π_s, we observed possible strong purifying selection in the partial kinase domains of wild tomato *Sl*LYK8 orthologs. This would be inconsistent with the *Sl*LYK8 orthologs having a truncated kinase like *Sl*LYK8 itself, which is missing vital sequences within its kinase domain and is presumably not functional. In that case, *Sl*LYK8 and its wild tomato orthologs might be expected to evolve as pseudogenes. However, selection pressure and the visibility of its hallmarks in the partial kinase domain would be more likely if the wild tomato orthologs had an intact kinase domain which is lacking in *S. lycopersicum*.

We therefore investigated whether *Sl*LYK8 orthologs in wild tomatoes also had truncated kinases. The transcriptome sequences, which were based on reads aligned to *S. lycopersicum*, would not show sequences that did not already exist in *S. lycopersicum*. Therefore, it was necessary for us to check for kinases trailing *Sl*LYK8 orthologs in *de novo* assemblies [30], which are not based on *S. lycopersicum*. To this end, *Sl*LYK8 and the corresponding positions of *Sl*LYK9 were used as query sequences in a nucleotide BLAST search against the *de novo* assembled transcriptomes. After filtering by sequence length and percent identity to *Sl*LYK8 and *Sl*LYK9, no single sequence was assigned to both queries. Amino acid translations of sequences which extended beyond the position of *Sl*LYK8 truncation in *S. lycopersicum* were aligned, and this alignment was used to build a rooted maximum likelihood tree (Fig. 5). Three sequences matched *Sl*LYK8 better than *Sl*LYK9 and had kinase domains extending past the position of *Sl*LYK8 truncation. Further inspection of the sequences revealed that each has an intact kinase including all positions of domains required for activity. One of the sequences was found in *S. chilense* and two were from *S. peruvianum*, which suggests that the truncation of *Sl*LYK8's kinase happened after the divergence between lineages leading to *S. lycopersicum* and the wild tomato species included here.

Discussion

We have found that Group III LysM-RLKs are more closely related to Group II LysM-RLKs than Group I LysM-RLKs. Group III LysM-RLKs are present in diverse Rosid species. *Sl*LYK3 is a member of a Group I clade with distinct first and second LysM domains, and its wild tomato orthologs have been subject to especially strong purifying selection on both their extracellular domains when compared to other wild tomato *Sl*LYK orthologs. *Sl*LYK3 wild tomato orthologs also show some indication of directional selection. Further, complete kinase domains are present in several *Sl*LYK8 wild tomato orthologs.

It is still unclear whether all wild tomatoes have *Sl*LYK8 orthologs with intact kinase domains, but its presence in both *S. peruvianum* and *S. chilense* shows that some wild tomatoes do. Other *Sl*LYK8 orthologs with kinases may have been below our cutoffs for percent identity or sequence length or not sufficiently expressed at the time samples were taken. If intact kinase domains are found in diverse wild tomato species, it would suggest that the truncation of the kinase is unique to *S. lycopersicum*. The potential for genetic drift in the *S. lycopersicum* genome due to bottlenecks introduced by selective breeding makes this a more likely scenario [32].

Overall, our study suggests that *Sl*LYK3 is an especially interesting candidate for further study. There is currently no data on the functionality of *Sl*LYK3, but its close relative in *A. thaliana* *At*LYK3 has functions in fungal and bacterial defense [17] and is essential for recognizing Rhizobia Nod factors [18]. Due to the indications of purifying selection on the extracellular domain and weak indications of directional selection pressure of *Sl*LYK3 orthologs in *S. chilense*, coupled with the distinctness of its LysM domains, it is an interesting candidate for further exploration of its ligands. Like *At*LYK3, it may detect several ligands and fulfill multiple roles.

It was previously postulated that LysM-II and LysM-III clades shared a more recent common ancestor than LysM-I and LysM-II due to both the lack of monocot genes in the LysM-III clade and the exon/intron structure of each clade [10]. Our analysis supports this idea, with strong indication via bootstrap support that LysM-II and LysM-III genes are more closely related to each other than either clade is to LysM-I. Despite expanding the known number of species containing LysM-III genes, we

still did not find any outside of the Rosids, a clade of the Eudicots. Zhang et al. noted that LysM-II genes in both *M. truncatula* and *O. sativa* lacked activation loops and conservation at residues necessary for activity [8], and we note that the same is true for the Glycine-rich loop; this sequence is missing in every LysM-II gene in our analysis, and multiple representatives of this clade occur in each species. Taken together, these results suggest that the most recent common ancestor of LysM-I diverged from the rest of the LysM-RLKs, and the LysM-III common ancestor originated prior to the divergence of *E. grandis* from the other Rosids.

Conclusions

The LysM-RLKs are a diverse family of proteins with many functions in plant symbiosis and defense and little correspondence between function and phylogenetic relationships. Here we provided an overview of the functions and phylogenetic relationships and found that the Group III LysM-RLKs share a closer relationship with those in Group II than those in Group I. Newly identified Group III LysM-RLKs were found in a variety of species in the Rosids. The kinase domain of *Sl*LYK8 homologs is intact in at least some species. We suggest that *Sl*LYK3 is a prime candidate for ligand-binding and functional analysis, owing to its close relationship to the multi-functional *At*LYK3, its distinct LysM domains, and signs that the extracellular domain of its gene sequence has undergone purifying selection in wild tomatoes.

Methods

Sources for Genomic Sequences

Amino acid sequences from *A. thaliana*, *S. lycopersicum*, *L. japonicus*, *O. sativa*, and *M. truncatula* were obtained from the sources listed in Table 4.

Transcriptome Data and Coverage Selection Criteria

Reads mapped to *S. lycopersicum* LysM-RLK genes in wild tomatoes were obtained from *S. chilense*, *S. peruvianum*, *S. ochranthum*, and *S. lycopersicoides* transcriptomes from Beddows *et al.* [30], except *Sl*LYK8 ortholog sequences, which were generated under the same conditions as the rest from [30], but with a minimum read depth of 5 (Supplemental File 5). Sequences were included in the population genetic analysis, provided they met the following conditions for sequence coverage:

- Individual sequences had <10% N-content (undetermined nucleotides) in the coding region
- A minimum of 8 sequences from *S. chilense* were required to satisfy the above condition in order for analysis to be done on the LysM-RLK orthologs from *S. chilense*. This condition was likewise upheld for *S. peruvianum*.

Accession LA0752 was included in the *S. chilense* sequence set. Sequences from LA1274, an accession described as *Solanum corneliomulleri*, were included in the *S. peruvianum* data set. LA2750, LA2884, LA0752, and LA2930 were excluded from analyses of the *Sl*LYK8 homologs, due to evidence of likely mismapping of sequences from *Sl*LYK9 homologs to the middle LysM2 domain.

Alignment Methods

Protein alignments were done in Mega7 [33] using the MUSCLE algorithm [34] with the following parameters: gap open -2.9; gap extend -0.01; hydrophobicity multiplier 1.2; max iterations 8; clustering UPGMB; lambda 24. Nucleotide alignments were done with the following parameters: gap open -400; gap extend 0; max iterations 8; clustering UPGMB; lambda 24.

Phylogenetic Analyses

Phylogenies based on protein sequences were built and tested in RAxML [35] with the protein substitution model that best fit the data (found using the PROTGAMMAAUTO function) and 500 bootstrap replicates. For DNA sequences, the GTRGAMMA function was used, and 500 bootstrap

replicates were generated. Seed values of 100 were chosen for reproducibility. Trees were rooted with MAD [36].

Population Genetic Analysis

All population genetic tests were performed in DnaSP [37] on the first haplotype of each sequence only. Significance for the McDonald-Kreitman test was determined by the G-test [38], except where this was not possible; otherwise, Fisher's exact test was used. Tajima's D was calculated using the total number of mutations [39].

BLAST Procedure for *Sl*LYK8 homologs

Two nucleotide BLAST searches [25] were performed against the *de novo* transcriptomes from Beddows et al. [30]: one with *Sl*LYK8 as the query and one with the corresponding positions of *Sl*LYK9 as the query. Percent identity was used to measure the quality of the hits to *Sl*LYK8 and *Sl*LYK9. All hits with at least 1000 nucleotides and 97% or greater identity to either *Sl*LYK8 or *Sl*LYK9 were used for further analysis. The sequences were translated in six frames and aligned together with *Sl*LYK8 and *Sl*LYK9. Translations which covered more than 40% of the *Sl*LYK8 domain and extended past the position of *Sl*LYK8 truncation were included in the alignment.

BLAST Procedure for Group III LysM-RLK homologs

*Lj*LYS20, *Lj*LYS21, *Mt*LYR5, and *Mt*LYR6 amino acid sequences were each used as queries in separate online protein BLAST searches [25] against NCBI's non-redundant protein sequences database [26]. The first 100 hits from each were compiled, and duplicates were removed before phylogenetic analysis.

List of Abbreviations

RLK: receptor-like kinase
LysM-RLK: lysin motif RLK
LYK: LysM-RLK with classically conserved kinase domain
LYR: LYK-related, or LysM-RLK without classically conserved kinase domain
LYS: LysM-RLKs in *Lotus japonicus*
CERK1: Chitin elicitor receptor-like kinase 1
At: *Arabidopsis thaliana*
Eg: *Eucalyptus grandis*
Lj: *Lotus japonicus*
Mt: *Medicago truncatula*
Os: *Oryza sativa*
Sl: *Solanum lycopersicum*

Declarations

Ethics approval and consent to participate:
Not applicable

Consent for publication:
Not applicable

Availability of data and material:
The *Sl*LYK8 ortholog sequences generated for this study can be found in the Supplemental Material for this paper. All other sequences analyzed in this study can be found in public repositories as indicated in the citations.

Competing interests:

The authors declare that they have no competing interests.

Funding:
Support for this research was provided by Deutsche Forschungsgemeinschaft (DFG) grants RO 2491/4-1 and RO 2491/5-2.

Authors' contributions:
LR and SR developed the hypotheses. SR tested the hypotheses and wrote the manuscript. LR edited the manuscript. Both authors approved of the final version of the manuscript.

Acknowledgements:
We would like to thank Christopher Blum, Sophie de Vries, and Jan de Vries for help editing the manuscript, Thorsten Klösges for providing assistance with the computer cluster, and the Rose lab for helpful suggestions and insights during the analysis.

References

[1] Oldroyd GED, Robatzek S. The broad spectrum of plant associations with other organisms. Curr Opin Plant Biol. 2011;14.4:347-350.

[2] Rajamuthiah R, Mylonakis E. Effector triggered immunity. Virulence. 2014;5.7:697-702.

[3] Gust AA, Willmann R, Desaki Y, Grabherr HM, Nürnberger T. Plant LysM proteins: modules mediating symbiosis and immunity. Trends Plant Sci. 2012;17.8:495-502.

[4] Cao Y, Liang Y, Tanaka K, Nguyen C, Jedrzejczak R, Joachimiak A et al. The kinase LYK5 is a major chitin receptor in Arabidopsis and forms a chitin-induced complex with related kinase CERK1. Elife. 2014;3:e03766.

[5] Zhang X, Dong W, Sun J, Feng F, Deng Y, He Z et al. The receptor kinase CERK1 has dual functions in symbiosis and immunity signaling. Plant J. 2014;81.2:258-267.

[6] Shiu SH, Karlowski WM, Pan R, Tzeng YH, Mayer KFX, Li WH. Comparative Analysis of the Receptor-Like Kinase Family in Arabidopsis and Rice. Plant Cell. 2004;16.5:1220-1234.

[7] Buendia L, Wang T, Girardin A, Lefebvre B. The LysM receptor-like kinase *Sl*LYK10 regulates the arbuscular mycorrhizal symbiosis in tomato. New Phytol. 2016;210.1:184-195.

[8] Zhang XC, Cannon SB, Stacey G. Evolutionary genomics of LysM genes in land plants. BMC Evol Biol. 2009;9.1:183.

[9] Arrighi JF, Barre A, Amor BB, Bersoult A, Soriano LC, Mirabella R, et al. The Medicago truncatula Lysine Motif-Receptor-Like Kinase Gene Family Includes NFP and New Nodule-Expressed Genes. Plant Physiol. 2006;142.1:265-279.

[10] Lohmann GV, Shimoda Y, Nielsen MW, Jørgensen FG, Grossmann C, Sandal N et al. Evolution and Regulation of the Lotus japonicus LysM Receptor Gene Family. Mol Plant Microbe Interact. 2010;23.4:510-521.

[11] Zeng L, Velásquez AC, Munkvold KR, Zhang J, Martin GB. A tomato LysM receptor-like kinase promotes immunity and its kinase activity is inhibited by AvrPtoB. Plant J. 2012 Jan 1;69(1):92-103.

[12] Radutoiu S, Madsen LH, Madsen EB, Felle HH, Umehara Y, Grønlund M, Sato S, Nakamura Y, Tabata S, Sandal N, Stougaard J. Plant recognition of symbiotic bacteria requires two LysM receptor-like kinases. Nature. 2003 Oct;425(6958):585.

[13] Limpens E, Franken C, Smit P, Willemse J, Bisseling T, Geurts R. LysM domain receptor kinases regulating rhizobial Nod factor-induced infection. Science. 2003 Oct 24;302(5645):630-3.

[14] Jones KM, Sharopova N, Lohar DP, Zhang JQ, VandenBosch KA, Walker GC. Differential response of the plant Medicago truncatula to its symbiont Sinorhizobium meliloti or an exopolysaccharide-deficient mutant. Proc Natl Acad Sci U S A. 2008 Jan 15;105(2):704-9.

[15] Miya A, Albert P, Shinya T, Desaki Y, Ichimura K, Shirasu K, Narusaka Y, Kawakami N, Kaku H, Shibuya N. CERK1, a LysM receptor kinase, is essential for chitin elicitor signaling in Arabidopsis. Proc Natl Acad Sci U S A. 2007 Dec 4;104(49):19613-8.

[16] Willmann R, Lajunen HM, Erbs G, Newman MA, Kolb D, Tsuda K, Katagiri F, Fliegmann J, Bono JJ, Cullimore JV, Jehle AK. Arabidopsis lysin-motif proteins LYM1 LYM3 CERK1 mediate bacterial peptidoglycan sensing and immunity to bacterial infection. Proc Natl Acad Sci U S A. 2011 Dec 6;108(49):19824-9.

[17] Paparella C, Savatin DV, Marti L, De Lorenzo G, Ferrari S. The Arabidopsis LYSIN MOTIF-CONTAINING RECEPTOR-LIKE KINASE3 regulates the cross talk between immunity and abscisic acid responses. Plant physiol. 2014 May 1;165(1):262-7

[18] Liang Y, Cao Y, Tanaka K, Thibivilliers S, Wan J, Choi J, ho Kang C, Qiu J, Stacey G. Nonlegumes respond to rhizobial Nod factors by suppressing the innate immune response. Science. 2013 Sep 20;341(6152):1384-7.

[19] Wan J, Tanaka K, Zhang XC, Son GH, Brechenmacher L, Nguyen TH, Stacey G. LYK4, a lysin motif receptor-like kinase, is important for chitin signaling and plant innate immunity in Arabidopsis. Plant physiol. 2012 Sep 1;160(1):396-406.

[20] Fliegmann J, Canova S, Lachaud C, Uhlenbroich S, Gasciolli V, Pichereaux C, Rossignol M, Rosenberg C, Cumener M, Pitorre D, Lefebvre B. Lipo-chitooligosaccharidic symbiotic signals are recognized by LysM receptor-like kinase LYR3 in the legume Medicago truncatula. ACS Chem Biol. 2013 Jul 5;8(9):1900-6.

[21] Hogekamp C, Arndt D, Pereira PA, Becker JD, Hohnjec N, Küster H. Laser microdissection unravels cell-type-specific transcription in arbuscular mycorrhizal roots, including CAAT-box transcription factor gene expression correlating with fungal contact and spread. Plant Physiol. 2011 Dec 1;157(4):2023-43.

[22] Rasmussen SR, Füchtbauer W, Novero M, Volpe V, Malkov N, Genre A, Bonfante P, Stougaard J, Radutoiu S. Intraradical colonization by arbuscular mycorrhizal fungi triggers induction of a lipochitooligosaccharide receptor. Sci Rep. 2016 Jul 20;6:29733.

[23] Rey T, Nars A, Bonhomme M, Bottin A, Huguet S, Balzergue S, Jardinaud MF, Bono JJ, Cullimore J, Dumas B, Gough C. NFP, a LysM protein controlling Nod factor perception, also intervenes in Medicago truncatula resistance to pathogens. New Phytol. 2013 May 1;198(3):875-86.

[24] Amor BB, Shaw SL, Oldroyd GE, Maillet F, Penmetsa RV, Cook D, Long SR, Dénarié J, Gough C. 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. Plant J. 2003 May 1;34(4):495-506.

[25] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403-10.

[26] O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2015 Nov 8;44(D1):D733-45.

[27] Sun M, Naeem R, Su JX, Cao ZY, Burleigh JG, Soltis PS, Soltis DE, Chen ZD. Phylogeny of the Rosidae: A dense taxon sampling analysis. Journal of systematics and evolution. 2016 Jul 1;54(4):363-91.

[28] Panchy N, Lehti-Shiu M, Shiu SH. Evolution of gene duplication in plants. Plant physiol. 2016 Aug 1;171(4):2294-316.

[29] Sela I, Ashkenazy H, Katoh K, Pupko T. GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. Nucleic Acids Research. 2015 Apr 16;43(W1):W7-14.

[30] Beddows I, Reddy A, Kloesges T, Rose LE. Population Genomics in Wild Tomatoes—The Interplay of Divergence and Admixture. Genome Biol Evol. 2017 Oct 24;9(11):3023-38.

[31] Šidák Z. Rectangular confidence regions for the means of multivariate normal distributions. J Am Stat Assoc. 1967 Jun 1;62(318):626-33.

[32] Bai Y, Lindhout P. Domestication and breeding of tomatoes: what have we gained and what can we gain in the future?. Ann Bot. 2007 Aug 23;100(5):1085-94.

[33] Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Mol Biology Evol. 2016;33.7:1870-1874.

[34] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32.5:1792-1797.

[35] Stamatakis A. RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. Bioinformatics. 2014;30.9:1312-1313.

[36] Tria FDK, Landan G, Dagan T. Phylogenetic rooting using minimal ancestor deviation. Nat Ecol Evol. 2017;1: s41559-017

[37] Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 2017;15.11:1451-1452.

[38] McDonald JH, Kreitman M. Adaptive protein evolution at the Adh locus in Drosophila. Nature. 1991;351.6328:652.

[39] Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989;123.3:585-95.

[40] Zhang XC, Wu X, Findley S, Wan J, Libault M, Nguyen HT et al. Molecular Evolution of Lysin Motif-Type Receptor-Like Kinases in Plants. Plant Physiol. 2007;144.2:623-636.

[41] Rhee SY, Beavis W, Berardini TZ, Chen G, Dixon D, Doyle A, et al. The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. Nucleic Acids Res. 2003;31.1:224.

[42] National Center for Biotechnology Information (NCBI). Bethesda, MD, USA. https://www.ncbi.nlm.nih.gov/. Accessed 21 Oct 2016.

[43] J. Craig Venter Institute MedicMine (JCVI MedicMine). Rockville, MD, USA. https://medicmine.jcvi.org/. Accessed 21 Oct 2016

[44] Kawahara Y, de la Bastide M, Hamilton JP, Kanamori H, McCombie WR, Ouyang S, et al. Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. Rice. 2013;6.1:4.

[45] Fernando-Pozo N, Menda N, Edwards JD, Saha S, Tecle IY, Strickler SR et al. The Sol Genomics Network (SGN) – from genotype to phenotype breeding. Nucleic Acids Res. 2015;43.D1:D1036-D1041

Tables

Table 1: Name conversions for Group III BLAST hits from Eucalyptus

Short name	NCBI non-redundant protein sequence
EgLYK1	gi|1091493514|ref|XP_010026290.2| PREDICTED: wall-associated receptor kinase-like 10
EgLYK2	gi|1091497044|ref|XP_010070127.2| PREDICTED: wall-associated receptor kinase-like 6
EgLYK3	gi|1091501059|ref|XP_018717605.1|| PREDICTED: wall-associated receptor kinase-like 10 isoform X1
EgLYK4	gi|1091501061|ref|XP_018717606.1| PREDICTED: wall-associated receptor kinase-like 10 isoform X2
EgLYK5	gi|629092723|gb|KCW58718.1| hypothetical protein EUGRSUZ_H013631, partial
EgLYK6	gi|629092727|gb|KCW58722.1| hypothetical protein EUGRSUZ_H01368
EgLYK7	gi|629092728|gb|KCW58723.1| hypothetical protein EUGRSUZ_H01369
EgLYK8	gi|629093021|gb|KCW59016.1| hypothetical protein EUGRSUZ_H01642
EgLYK9	gi|629093022|gb|KCW59017.1| hypothetical protein EUGRSUZ_H01643, partial
EgLYK10	gi|629093024|gb|KCW59019.1| hypothetical protein EUGRSUZ_H01646
EgLYK11	gi|629093026|gb|KCW59021.1| hypothetical protein EUGRSUZ_H01648
EgLYK12	gi|629093027|gb|KCW59022.1| hypothetical protein EUGRSUZ_H01649
EgLYK13	gi|702436539|ref|XP_010070126.1| PREDICTED: wall-associated receptor kinase-like 10
EgLYK14	gi|702453754|ref|XP_010026289.1| PREDICTED: protein LYK5
EgLYK15	gi|702510373|ref|XP_010040652.1| PREDICTED: wall-associated receptor kinase-like 1

Group III LysM-RLKs *Lj*LYS20, *Lj*LYS21, *Mt*LYR5, and *Mt*LYR6 were used as query sequences in a BLAST search against NCBI's non-redundant protein sequences database. There were fifteen hits in Eucalyptus, some of them isoforms of the same gene or nearly identical sequences with different naming schemes. The names given to these genes were shortened for Figure 2, and this table converts the shortened names to the original names with gi numbers.

Table 2: Summary statistics and tests of neutrality in wild tomato LYK orthologs

	No. Seqs	No. Sites	Haplotypes	π(non,syn,silent)	π_a/π_s	S	MK p-value
LYK1		1881					
chil v ochr	17 v 1		13	0.0027 (0.0016, 0.0059, 0.0059)	0.28	19	0.062
chil v lyco	17 v 1			0.0027 (0.0016, 0.0059, 0.0059)	0.28		0.026
peru v ochr	17 v 1		15	0.0049 (0.0026, 0.0118, 0.0117)	0.22	43	0.185
peru v lyco	17 v 1			0.0049 (0.0026, 0.0118, 0.0117)	0.22		0.146
LYK3		1992					
chil v ochr	17 v 1		16	0.0046 (0.0011, 0.0160, 0.0159)	0.07	37	0.036
chil v lyco	17 v 1			0.0046 (0.0011, 0.0152, 0.0151)	0.07		0.718
peru v ochr	16 v 1		15	0.0056 (0.0008, 0.0211, 0.0209)	0.04	51	0.005
peru v lyco	16 v 1			0.0056 (0.0008, 0.0215, 0.0214)	0.04		0.822
LYK4		1938					
chil v ochr	15 v 1		15	0.0080 (0.0061, 0.0138, 0.0137)	0.44	51	0.510
chil v lyco	15 v 1			0.0080 (0.0061, 0.0138, 0.0137)	0.44		0.760
peru v ochr	16 v 1		16	0.0089 (0.0054, 0.0202, 0.0201)	0.26	80	0.275
peru v lyco	16 v 1			0.0089 (0.0054, 0.0202, 0.0201)	0.26		0.336
LYK6		1599					
chil v ochr	15 v 1		13	0.0151 (0.0119, 0.0246, 0.0246)	0.48	108	0.108
chil v lyco	15 v 1			0.0151 (0.0119, 0.0245, 0.0245)	0.48		0.069
peru v ochr	8 v 1		8	0.0161 (0.0135, 0.0244, 0.0244)	0.55	87	0.118
peru v lyco	8 v 1			0.0161 (0.0135, 0.0244, 0.0244)	0.55		0.064
LYK8		1149					
chil v ochr	13 v 1		12	0.0048 (0.0044, 0.0064, 0.0064)	0.69	34	0.025
ceru v ochr	17 v 1		16	0.0052 (0.0042, 0.0087, 0.0087)	0.47	55	0.026
LYK9		1890					
chil v ochr	17 v 1		17	0.0073 (0.0042, 0.0174, 0.0176)	0.24	74	0.818
chil v lyco	17 v 1			0.0073 (0.0042, 0.0174, 0.0176)	0.24	73	0.604
peru v ochr	17 v 1		17	0.0068 (0.0045, 0.0135, 0.0146)	0.33	79	0.813
peru v lyco	17 v 1			0.0068 (0.0045, 0.0174, 0.0146)	0.33	77	0.649

Wild tomato orthologs for several LysM-RLK genes from *Solanum lycopersicum* were analyzed. Wild tomato species included were *Solanum chilense* (chil), *Solanum peruvianum* (peru), *Solanum ochranthum* (ochr), and *Solanum lycopersicoides* (lyco). Haplotypes apply to the first species listed.

Table 3: Summary statistics and tests of neutrality on extracellular and intracellular domains

	Extra-/ Intracellular	No. Seqs	No. Sites	Haplotypes	π(non,syn,silent)	π_a/π_s	S
LYK1							
chil v ochr	Extracellular	17 v 1	699	8	0.0026 (0.0025, 0.0032, 0.0032)	0.79	8
	Intracellular		1101	9	0.0028 (0.0011, 0.0079, 0.0078)	0.14	10
chil v lyco	Extracellular	17 v 1	699	8	0.0026 (0.0025, 0.0032, 0.0032)	0.80	8
	Intracellular		1101	9	0.0028 (0.0011, 0.0079, 0.0078)	0.14	10
peru v ochr	Extracellular	17 v 1	699	12	0.0063 (0.0042, 0.0129, 0.0129)	0.33	21
	Intracellular		1101	15	0.0043 (0.0018, 0.0117, 0.0116)	0.16	21
peru v lyco	Extracellular	17 v 1	699	12	0.0063 (0.0043, 0.0130, 0.0130)	0.33	21
	Intracellular		1101	15	0.0043 (0.0018, 0.0117, 0.0116)	0.16	21
LYK3							
chil v ochr	Extracellular	17 v 1	705	12	0.0046 (0.0005, 0.0176, 0.0176)	0.03	13
	Intracellular		1203	16	0.0044 (0.0009, 0.0158, 0.0156)	0.06	21
chil v lyco	Extracellular	17 v 1	705	12	0.0046 (0.0005, 0.0155, 0.0155)	0.03	13
	Intracellular		1203	16	0.0044 (0.0009, 0.0158, 0.0156)	0.06	21
peru v ochr	Extracellular	16 v 1	705	13	0.0086 (0.0008, 0.0328, 0.0328)	0.03	24
	Intracellular		1203	14	0.0042 (0.0008, 0.0155, 0.0153)	0.05	25
peru v lyco	Extracellular	16 v 1	705	13	0.0088 (0.0009, 0.0349, 0.0349)	0.02	24
	Intracellular		1203	14	0.0042 (0.0008, 0.0155, 0.0153)	0.05	25
LYK4							
chil v ochr	Extracellular	15 v 1	801	15	0.0089 (0.0074, 0.0136, 0.0136)	0.54	18
	Intracellular		1053	14	0.0074 (0.0053, 0.0139, 0.0137)	0.38	31
chil v lyco	Extracellular	15 v 1	801	15	0.0089 (0.0074, 0.0136, 0.0136)	0.54	18
	Intracellular		1053	14	0.0074 (0.0053, 0.0139, 0.0137)	0.38	31
peru v ochr	Extracellular	16 v 1	801	16	0.0106 (0.0087, 0.0161, 0.0089)	0.54	41
	Intracellular		1053	11	0.0075 (0.0029, 0.0236, 0.0233)	0.12	33
peru v lyco	Extracellular	16 v 1	801	16	0.0106 (0.0087, 0.0161, 0.0161)	0.54	41
	Intracellular		1053	11	0.0075 (0.0029, 0.0236, 0.0233)	0.12	33
LYK6							
chil v ochr	Extracellular	15 v 1	780	12	0.0146 (0.0132, 0.0198, 0.0198)	0.66	29
	Intracellular		741	13	0.0150 (0.0105, 0.0265, 0.0265)	0.39	31
chil v lyco	Extracellular	15 v 1	780	12	0.0146 (0.0132, 0.0198, 0.0198)	0.66	29
	Intracellular		741	13	0.0150 (0.0105, 0.0265, 0.0265)	0.39	31
peru v ochr	Extracellular	8 v 1	780	8	0.0152 (0.0149, 0.0166, 0.0166)	0.90	33
	Intracellular		741	8	0.0188 (0.0127, 0.0397, 0.0397)	0.31	30
peru v lyco	Extracellular	8 v 1	780	8	0.0152 (0.0149, 0.0165, 0.0165)	0.90	33
	Intracellular		741	8	0.0188 (0.0127, 0.0397, 0.0397)	0.31	30
LYK8							
chil v ochr	Extracellular	13 v 1	771	9	0.0042 (0.0053, 0.0006, 0.0006)	9.1	9
	Intracellular*		300	3	0.0022 (0.0000, 0.0116, 0.0116)	0.00	3
peru v ochr	Extracellular	17 v 1	771	13	0.0038 (0.0032, 0.0058, 0.0058)	0.55	16
	Intracellular*		300	9	0.0047 (0.0035, 0.0101, 0.0101)	0.34	11
LYK9							
chil v ochr	Extracellular	17 v 1	774	14	0.0086 (0.0068, 0.0145, 0.0145)	0.47	22
	Intracellular		1038	14	0.0060 (0.0020, 0.0199, 0.0202)	0.10	22
chil v lyco	Extracellular	17 v 1	774	14	0.0086 (0.0068, 0.0145, 0.0145)	0.47	22
	Intracellular		1038	14	0.0060 (0.0020, 0.0199, 0.0202)	0.10	22
peru v ochr	Extracellular	17 v 1	774	14	0.0053 (0.0040, 0.0092, 0.0092)	0.44	25
	Intracellular		1038	17	0.0082 (0.0051, 0.0168, 0.0188)	0.30	32
peru v lyco	Extracellular	17 v 1	774	14	0.0053 (0.0040, 0.0092, 0.0092)	0.44	25
	Intracellular		1038	17	0.0082 (0.0052, 0.0169, 0.0189)	0.30	32

Wild tomato orthologs for intracellular and extracellular domains of several LYK genes from *Solanum lycopersicum* were analyzed. Wild tomato species included were *Solanum chilense* (chil), *Solanum peruvianum* (peru), *Solanum ochranthum* (ochr), and *Solanum lycopersicoides* (lyco). Haplotypes apply to the first species listed. *: LYK8 has a truncated kinase domain

Table 4: Sources for protein sequences used in LysM-RLK phylogeny analysis

Name	Alt Name	Locus ID	GenBank	Source	Database
AtLYK1	AtCERK1	At3g21630		[40]	TAIR10 [41]
AtLYK2		At3g01840		[40]	TAIR10 [41]
AtLYK3		At1g51940		[40]	TAIR10 [41]
AtLYK4		At2g23770		[40]	TAIR10 [41]
AtLYK5		At2g33580		[40]	TAIR10 [41]
LjNFR1			CAE02589.1	[10]	[42]
LjNFR5			CAE02597.1	[10]	[42]
LjLYS1			AB503681	[10]	[42]
LjLYS2			AB503682	[10]	[42]
LjLYS3			AB503683	[10]	[42]
LjLYS4			AB503684	[10]	[42]
LjLYS5			AB503686	[10]	[42]
LjLYS6			AB503687	[10]	[42]
LjLYS7			AB503688	[10]	[42]
LjLYS11			AB503689	[10]	[42]
LjLYS12			AB503690	[10]	[42]
LjLYS13			AB503691	[10]	[42]
LjLYS14			AB503692	[10]	[42]
LjLYS15			AB503693	[10]	[42]
LjLYS16			AB503694	[10]	[42]
LjLYS20			AB503695	[10]	[42]
LjLYS21			AB503696	[10]	[42]
MtNFP	MtNFR5	Medtr5g019040.1		[9]	Mt4.0v2 [43]
MtLYK1		Medtr5g086540.1		[9]	Mt4.0v2 [43]
MtLYK2		Medtr5g086310.1		[9]	Mt4.0v2 [43]
MtLYK3		Medtr5g086130.1		[9]	Mt4.0v2 [43]
MtLYK4		Medtr5g086120.1		[9]	Mt4.0v2 [43]
MtLYK5		Medtr5g086090.1		[9]	Mt4.0v2 [43]
MtLYK6		Medtr5g086040.1		[9]	Mt4.0v2 [43]
MtLYK7		Medtr5g086030.1		[9]	Mt4.0v2 [43]
MtLYK8		Medtr2g024290.1		[9]	Mt4.0v2 [43]
MtLYK9		Medtr3g080050.1		[9]	Mt4.0v2 [43]
MtLYK10		Medtr5g033490.1		[9]	Mt4.0v2 [43]
MtLYR1		Medtr8g078300.1		[9]	Mt4.0v2 [43]
MtLYR2		Medtr1g021845.1		[9]	Mt4.0v2 [43]
MtLYR3		Medtr5g019050.1		[9]	Mt4.0v2 [43]
MtLYR4		Medtr5g085790.1		[9]	Mt4.0v2 [43]
MtLYR5		Medtr7g079350.1		[9]	Mt4.0v2 [43]
MtLYR6		Medtr7g079320.1		[9]	Mt4.0v2 [43]
MtLYR7		Medtr3g080170.1		[9]	Mt4.0v2 [43]
OsLYK1		LOC_Os01g36550		[40]	TIGR7 [44]
OsLYK2		LOC_Os06g41980		[40]	TIGR7 [44]
OsLYK3		LOC_Os06g41960		[40]	TIGR7 [44]
OsLYK4		LOC_Os02g09960		[40]	TIGR7 [44]

OsLYK5		LOC_Os03g13080		[40]	TIGR7 [44]
OsLYK6		LOC_Os11g35330		[40]	TIGR7 [44]
SlLYK1	Bti9	Solyc07g049180.2.1		[11]	ITAG2.4 [45]
SlLYK2		Solyc02g094010.1.1		[11]	ITAG2.4 [45]
SlLYK3		Solyc03g121050.2.1		[11]	ITAG2.4 [45]
SlLYK4		Solyc02g089900.1.1		[11]	ITAG2.4 [45]
SlLYK6		Solyc12g089020.1.1		[11]	ITAG2.4 [45]
SlLYK7		Solyc02g089920.1.1		[11]	ITAG2.4 [45]
SlLYK8		Solyc09g083200.2.1		[11]	ITAG2.4 [45]
SlLYK9		Solyc09g083210.2.1		[11]	ITAG2.4 [45]
SlLYK10		Solyc02g065520.1.1		[11]	ITAG2.4 [45]
SlLYK11		Solyc02g081040.2.1		[11]	ITAG2.4 [45]
SlLYK12		Solyc02g081050.2.1		[11]	ITAG2.4 [45]
SlLYK13		Solyc01g098410.2.1		[11]	ITAG2.4 [45]
SlLYK14		Solyc06g069610.1.1		[11]	ITAG2.4 [45]
SlLYK15		Solyc11g069630.1.1		[7]	ITAG2.4 [45]

All amino acid sequences analyzed in this study were obtained from publicly available databases as indicated in the database column.

Figures

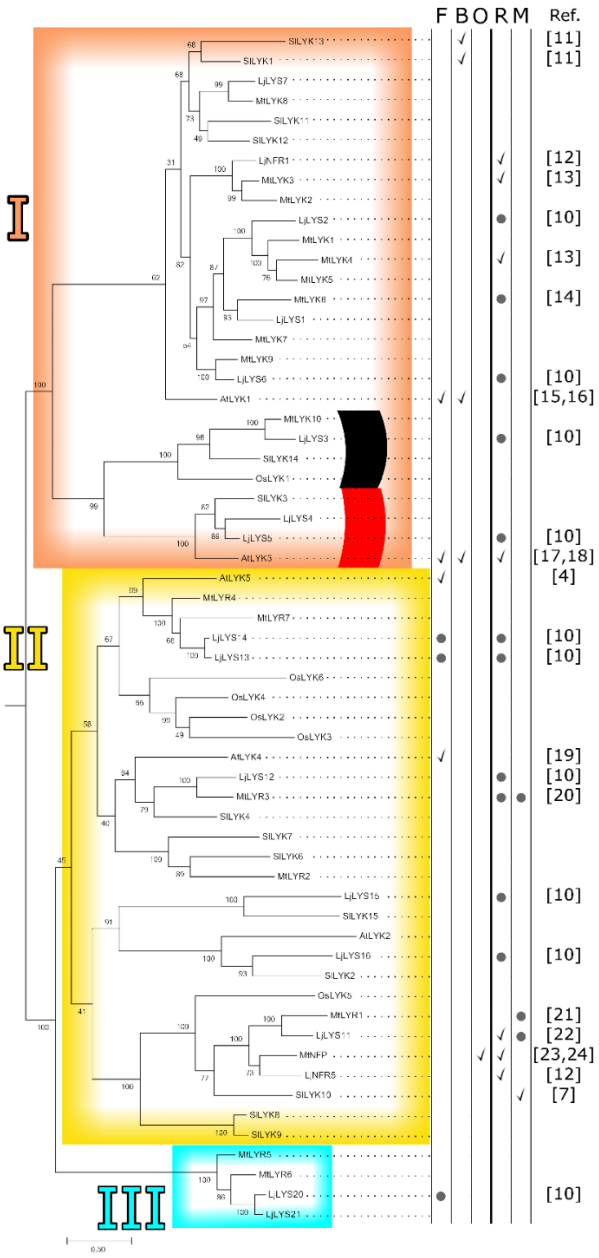


Fig. 1 LysM-RLK phylogeny and functions

Maximum likelihood phylogeny and functions of LysM-RLKs from *Solanum lycopersicum* (*Sl*), *Arabidopsis thaliana* (*At*), *Lotus japonicus* (*Lj*), *Oryza sativa* (*Os*), and *Medicago truncatula* (*Mt*). Gene functions are indicated: defense against fungi (F), bacteria (B), and oomycetes (O) and symbiosis with rhizobia (R) and myccorhiza (M). LysM-RLKs form three clades. Red and black arcs indicate groups of proteins with distinct LysM domain sequences. Functions verified by mutation phenotypes are indicated by check marks. Functions inferred by differential expression are indicated by gray circles. Citations for sources of functional information are shown in brackets.

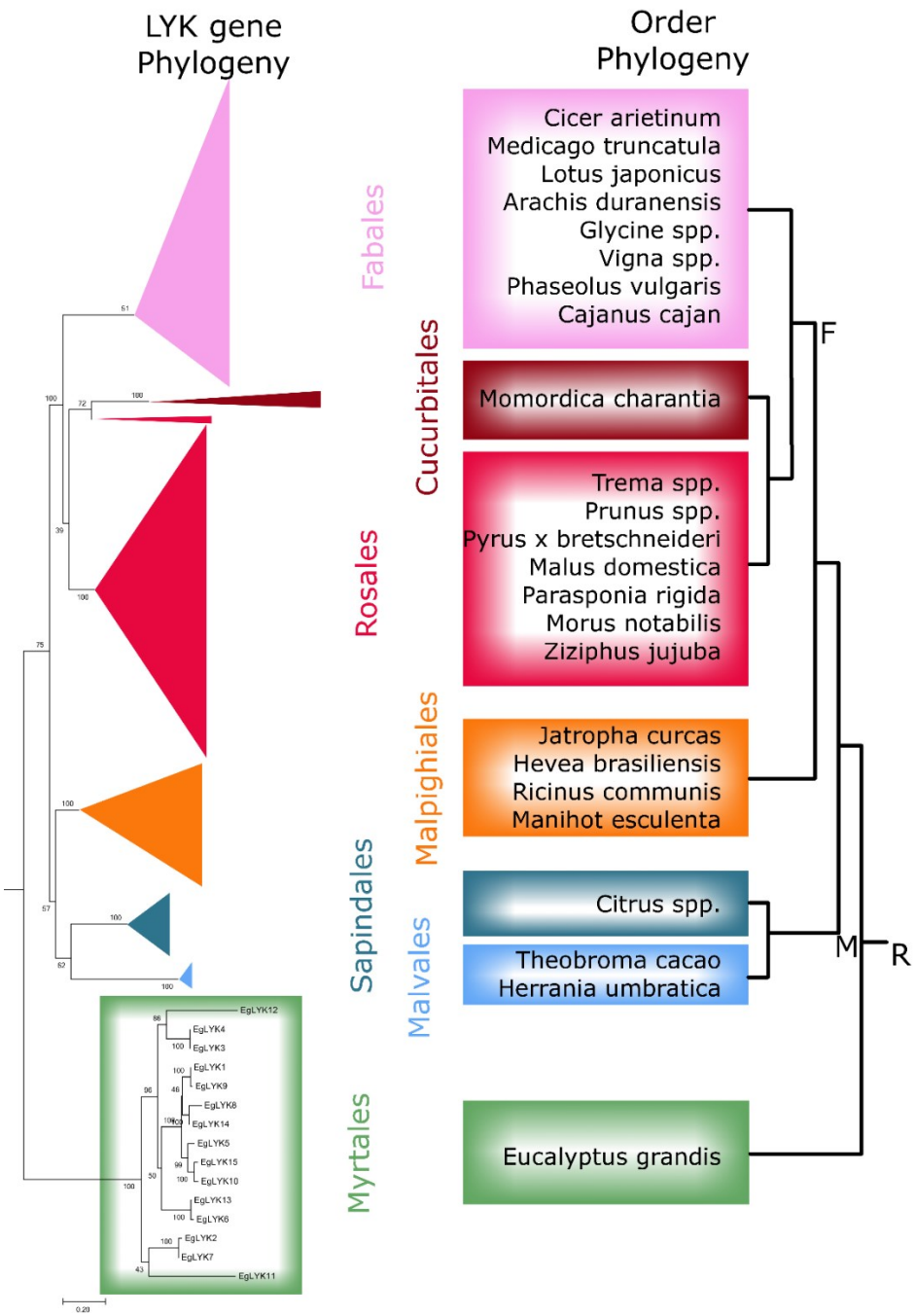


Fig. 2 Condensed Group III LysM-RLK BLAST hit phylogeny and Order phylogeny

Homologs of known Group III LysM-RLKs are found in several species throughout the Rosids (R). The vertical length of the triangles corresponds to number of genes found in the clade (e.g. three in the Malvales clade and two in the Cucurbitales clade). Each species had two to five BLAST hits each, but fifteen were found in *Eucalyptus grandis*. Hits from Malpighiales and orders within the Malvids (M) have an ambiguous relationship, and one sequence from *Trema orientalis* reliably groups with genes from *Momordica charantia*, a species in a closely related order within the Fabids (F). Order phylogeny is based on Sun *et al.* [27].

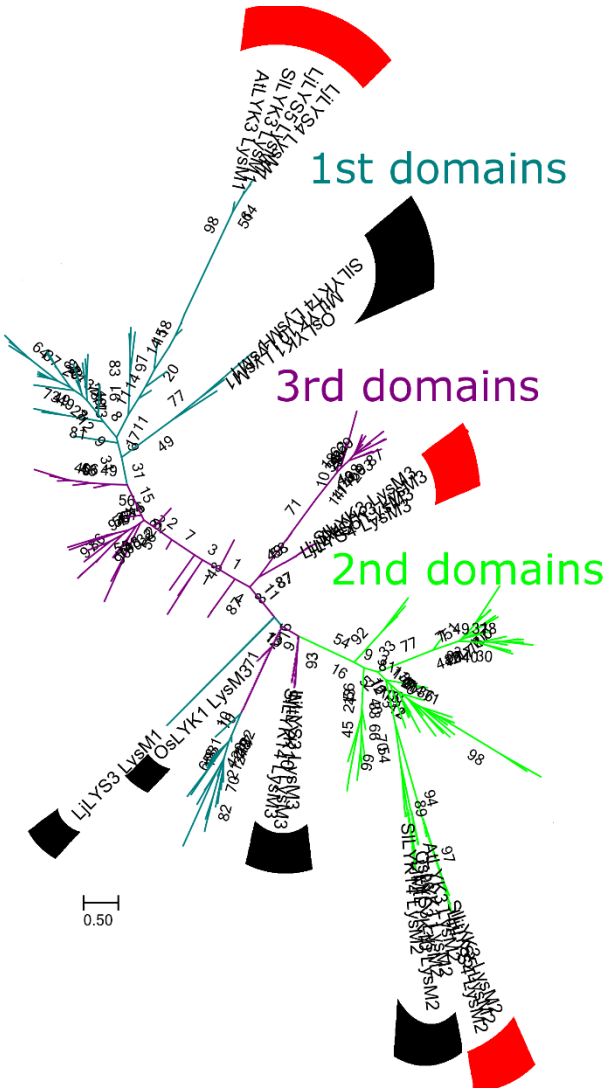


Fig. 3 Unrooted phylogeny of individual LysM-RLK domains

Phylogeny of amino acid sequences of each of the three LysM-RLK protein domains from each of the LysM-RLKs of *Solanum lycopersicum* (*Sl*), *Arabidopsis thaliana* (*At*), *Lotus japonicus* (*Lj*), *Oryza sativa* (*Os*), and *Medicago truncatula* (*Mt*). First, second, and third domain sequences generally cluster in clades with others of the corresponding domain, but the first and third domains do not form separate clades. Especially long branches lead to the first and second domains of two groups of genes of interest. Domain sequences from the first group are highlighted in red: *At*LYK3, *Sl*LYK3, *Lj*LYS4, and *Lj*LYS5. The second group is highlighted in black: *Os*LYK1, *Mt*LYK10, *Sl*LYK14, and *Lj*LYS3. The third domain of *Os*LYK1 and first domain of *Lj*LYS3 are separated from the corresponding domains of the second group.

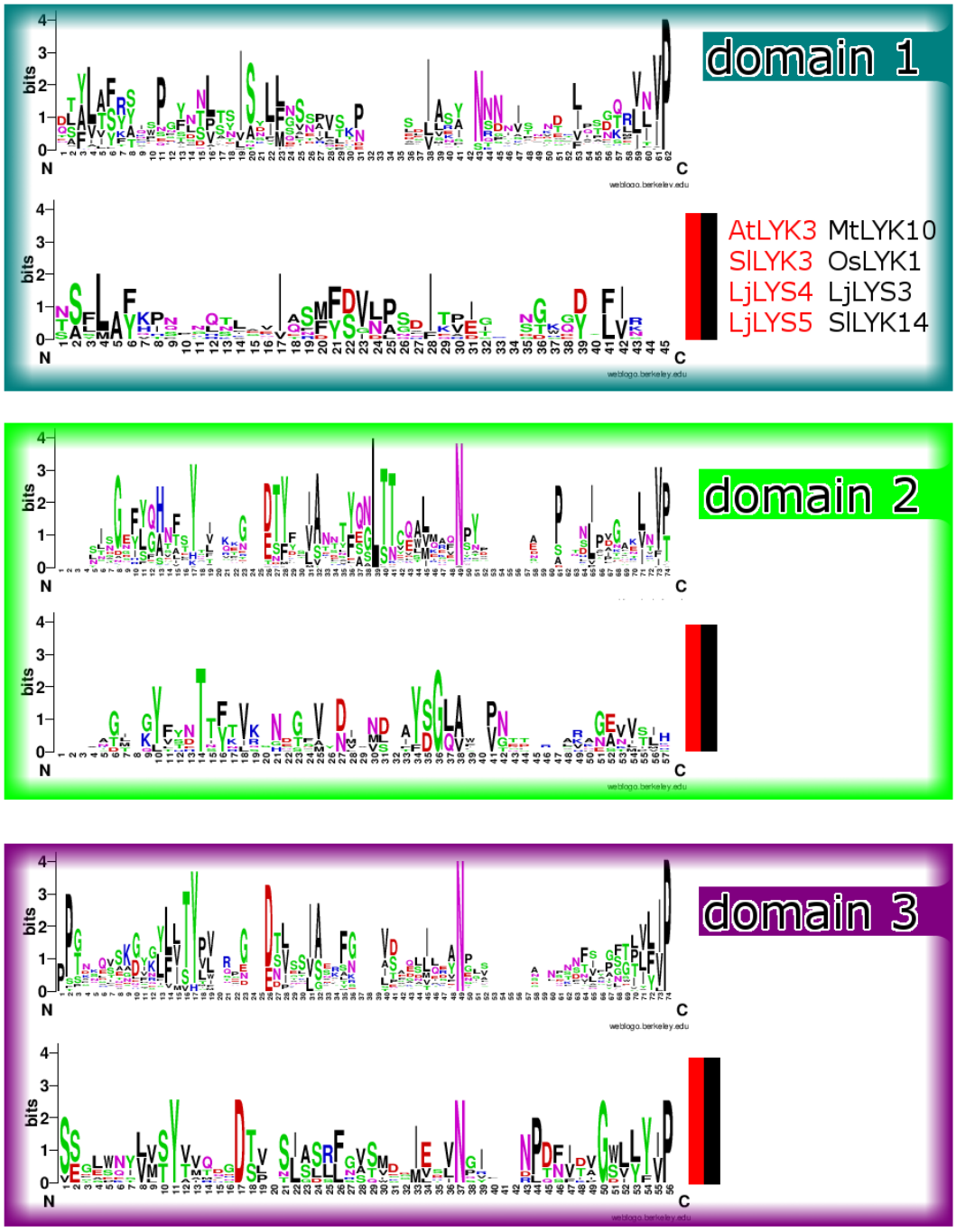


Fig. 4 Amino acid logos of LysM-RLK domains

Logos of LysM-RLK LysM domains, with those of *At*LYK3, *Sl*LYK3, *Lj*LYS4, *Lj*LYS5, *Mt*LYK10, *Os*LYK1, *Lj*LYS3, and *Sl*LYK14 logos computed separately. The third domains of both sets of sequences share conserved amino acids, while first and second domains of the two sequence sets share few conserved amino acids.

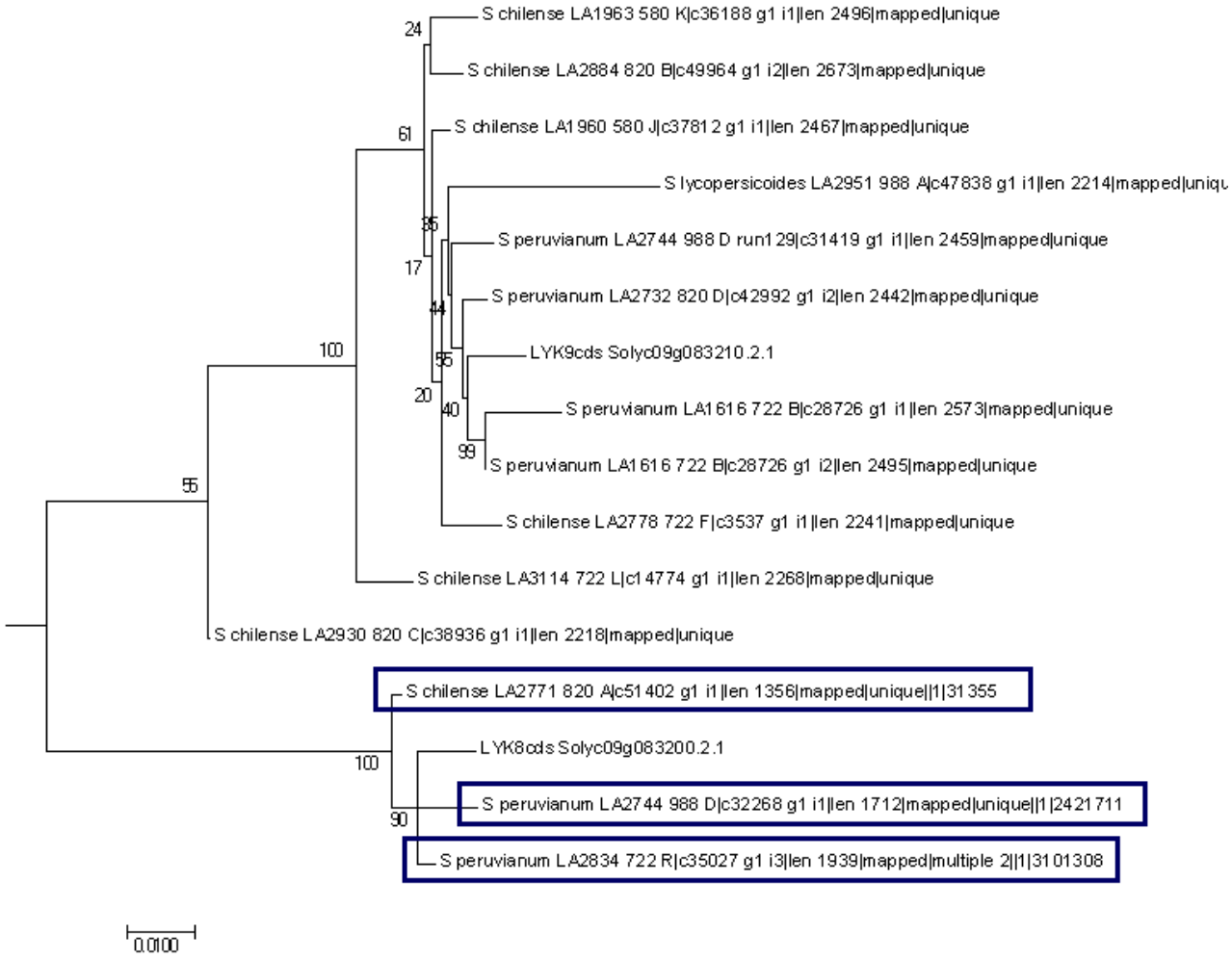
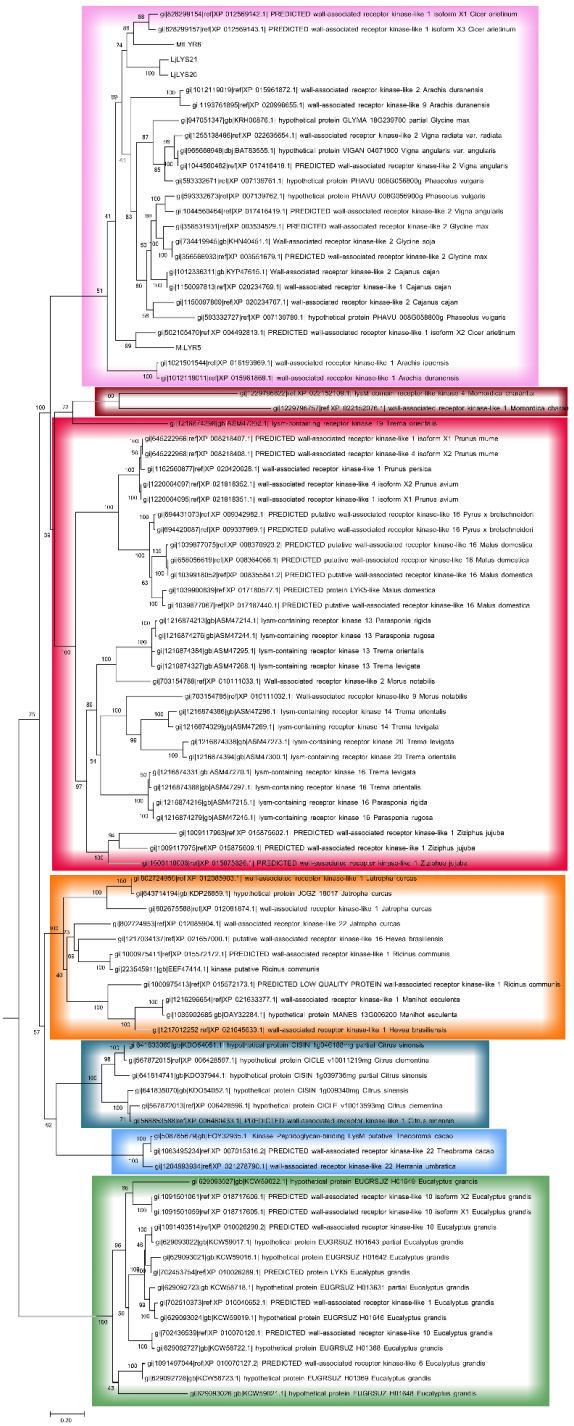


Fig. 5 Phylogeny of wild tomato *Sl*LYK8 and *Sl*LYK9 orthologs with intact kinases

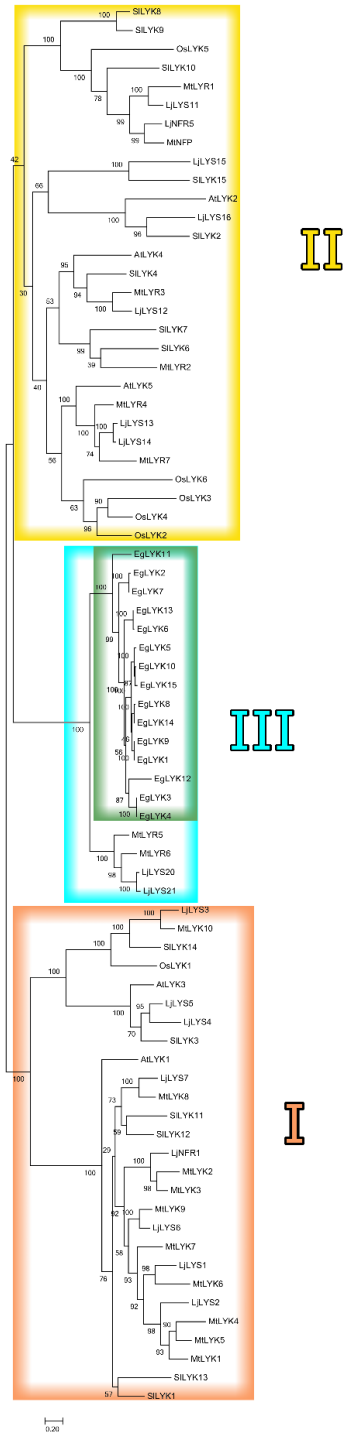
This phylogeny includes *Sl*LYK8 and *Sl*LYK9 BLAST hits which extend past the point of *Sl*LYK8 truncation. Three sequences from *Solanum peruvianum* and *Solanum chilense* with intact kinase domains more closely match *Solanum lycopersicum* LysM-RLK *Sl*LYK8 than *Sl*LYK9 in terms of both phylogeny and percent identity.

Additional Files



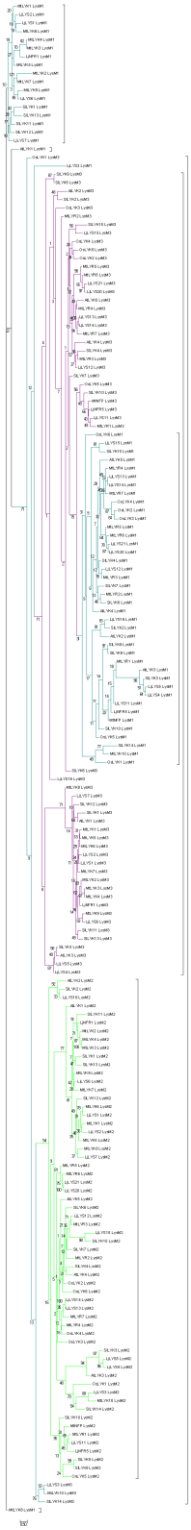
Additional File 1 Figure S1 Full tree of Group III BLAST hits

This phylogeny is identical to the Group III BLAST hit phylogeny in Fig. 2, but the sequence names have not been condensed. Names of the genes' species of origin have been appended to the original names from NCBI, and special characters (spaces and colons) have been removed. (PNG)



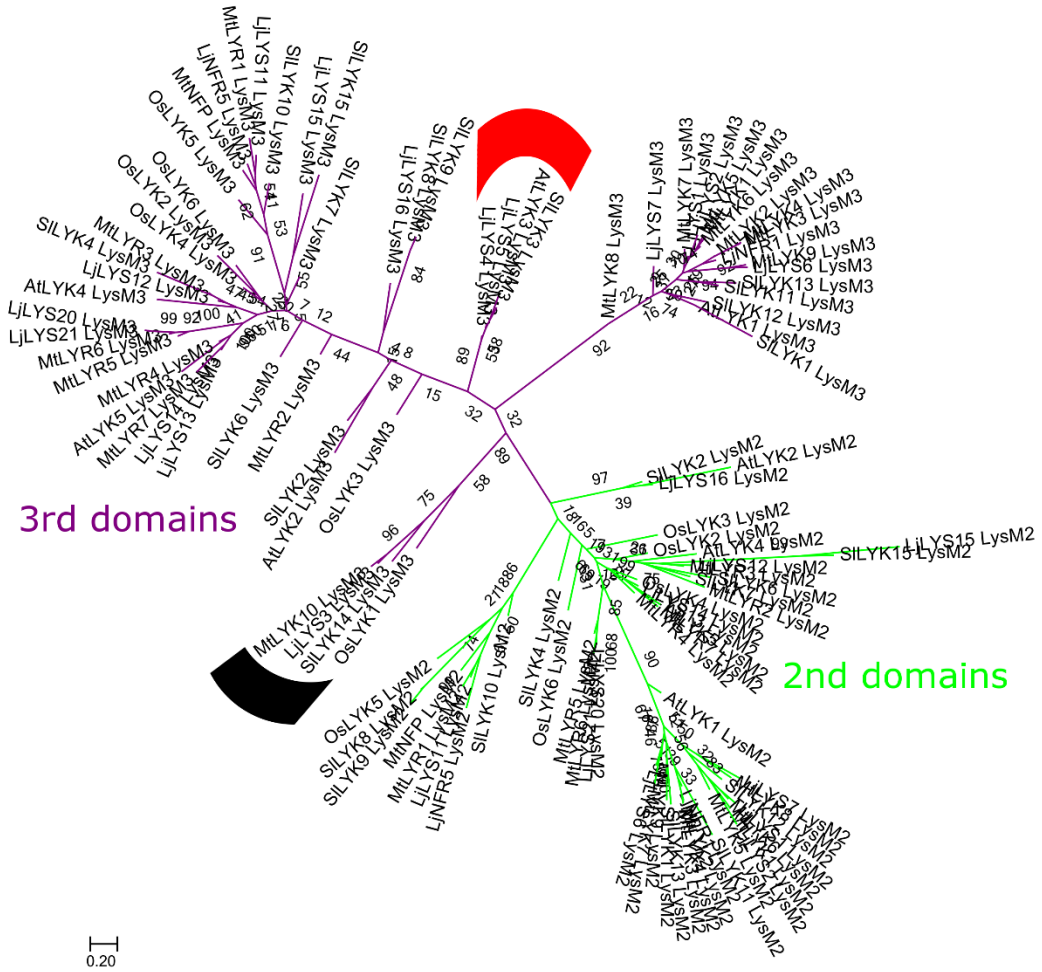
Additional File 2 Figure S2 Phylogeny of known LysM-RLKs and the Group III *Eucalyptus grandis* sequences

Phylogeny of *Eucalyptus grandis* (*Eg*) LysM-RLK amino acid sequences and known LysM-RLKs from *Solanum lycopersicum* (*Sl*), *Arabidopsis thaliana* (*At*), *Lotus japonicus* (*Lj*), *Oryza sativa* (*Os*), and *Medicago truncatula* (*Mt*). The *Eucalyptus grandis* LysM-RLKs discovered in BLAST searches with Group III LysM-RLK queries are most closely related to the Group III LysM-RLKs. (PNG)



Additional File 3 Figure S3 Phylogeny of LysM-RLK domains with all names visible

This phylogeny is identical to the phylogeny in Fig. 3, but all names are clearly visible. The phylogeny is unrooted but is shown in rectangular format. (PNG)



Additional File 4 Figure S4 Phylogeny of reliably aligned individual LysM-RLK domains

Phylogeny of amino acid sequences of the LysM-RLK LysM domains which scored 0.80 or higher when evaluated with GUIDANCE. Each of the first domains scored poorly, and all were omitted. All third domains were included. Second domains of genes highlighted in red and black were omitted. The second and third domains of the sequences included form distinct clades. (PNG)

[included in the digital version of this thesis only]

Additional File 5 *Sl*LYK8 ortholog sequences generated for this study

This is a fasta-formatted text file containing the sequences generated from reads from several wild tomato species (Solanum peruvianum: peru, Solanum chilense: chil, Solanum lycopersicoides:lyco, and Solanum ochranthum: ochr) which were mapped to the region corresponding to *Sl*LYK8 in Solanum lycopersicum. Unlike the rest of the LysM-RLK orthologs obtained from [30], these sequences were assessed with a minimum read depth of five sequences. The rest of the mapping procedure was the same as that used for the other sequences.

9 Acknowledgements

I don't know another PhD student who could say their last two years were mostly relaxed and enjoyable. Laura, your optimism and seemingly inexhaustible willingness to help everyone around you have been huge sources of relief for me when I needed it most. You are the main reason I was able to learn a new discipline and subject, publish, and finish what I started here. I am astounded and incredibly grateful.

Much of the relaxed nature of this last stretch was due to my source of funding. Thank you to Axel Görlitz, for organizing the Vorkurs program for natural science students, and Achim Schädle for recommending me to teach the mathematics course. It has given me a rewarding job, independence, and no time limit.

Thank you Oliver Ebenhöh for being my second reviewer for this thesis when I finally did get into the crunch time to have my "dream team" of committee members, including the fun and always supportive Shin-han Shiu, who supervised my work at MSU and made sure I ate well, and Markus Kollmann, who regularly provided his expertise throughout my work here though he wasn't obligated.

I wouldn't be here at all if it weren't for Rüdiger Simon, who convinced me to come back after a 3-month internship in his lab. Thank you Rüdiger for your simple explanations and eagerness to share your knowledge and contacts, as well as your support while I was writing my first publication.

Thank you to Fred for your friendship, support, and understanding. You made my life bearable during the worst times. Andrea, thank you for sharing your wisdom and encouragement, and for warning me when I was being too American. Thanks to Barbara Berckmans and the rest of the Simon lab for your friendliness whenever I visited the lab, and an extra thanks to those of you who proofread my first paper for me and had fun with me outside the lab (and very generously gave me wedding gifts). Thanks to the Rose lab, especially Sophie, Janina, and Anika for stimulating discussions, outings, and Game of Thrones night. Anastasia and my friends in various other labs, thank you for all the fun, support, and inspiration.

I also owe many thanks to many members of the iGRAD-Plant team. Sigrun Wegener-Feldbrügge is chief among them for being amazingly organized, professional, and clever and just making everything work. I got help and resources from most of the people involved in iGRAD-Plant at one time or another, and I appreciate you all taking the time to give me guidance. Thank you to Sophie, Jan, Fred, Anastasia, Woogie, Janina, Alex, Anika, Barbara, Jenia, Alisandra, Nina, Thorsten, and Melanie for volunteering to give me comments on this thesis. Thanks also goes to Mayo and Sven for introducing me to GUIDANCE.

Mom and grandma, your "Can she swim? I don't know. They'll toss her in the deep end, and we'll find out" approach has granted me a great deal of freedom, and it is because of you that I had the courage to do this. Thank you for all of your support and for braving the plane to come see me. Renate, thank you for adopting me and treating me like a daughter, and thank you Patricia for being my sister. I was so lucky to have a family here while mine was an ocean away.

Lastly, to Chris, you witnessed all the madness of this and not only tolerated it, but actively took care of me when I couldn't do it myself. Everyone owes their spouse a thank-you when they graduate, but few of them know for sure that the reason they are graduating is because their spouse insisted they go to the Dean's office. I could not have done this without you.

10 References

1. Graur D, Li WH. Fundamentals of molecular evolution, 2nd edition. Sunderland, Massachusetts: Sinauer; 2000.
2. Karki R, Pandya D, Elston RC, Ferlini C. Defining "mutation" and "polymorphism" in the era of personal genomics. BMC medical genomics. 2015 Dec;8(1):37.
3. Gillespie JH. Population Genetics: A Concise Guide, 2nd edition. JHU Press; 2004 Jun 28.
4. Curtis H, Barnes NS. Biology, 5th edition. Worth Publishers; 1989.
5. Nei M, Kumar S. Molecular evolution and phylogenetics. Oxford university press; 2000 Jul 27.
6. McDonald JH, Kreitman M. Adaptive protein evolution at the Adh locus in Drosophila. Nature. 1991;351.6328:652.
7. Allison AC. Protection afforded by sickle-cell trait against subtertian malarial infection. British medical journal. 1954 Feb 6;1(4857):290.
8. Matsumoto T, John A, Baeza-Centurion P, Li B, Akashi H. Codon usage selection can bias estimation of the fraction of adaptive amino acid fixations. Molecular biology and evolution. 2016 Feb 12;33(6):1580-9.
9. Huelsenbeck JP, Crandall KA. Phylogeny estimation and hypothesis testing using maximum likelihood. Annual Review of Ecology and Systematics. 1997 Nov;28(1):437-66.
10. Polynikis A, Hogan SJ, di Bernardo M. Comparing different ODE modelling approaches for gene regulatory networks. Journal of theoretical biology. 2009 Dec 21;261(4):511-30.
11. Alon U. An introduction to systems biology: design principles of biological circuits. CRC press; 2006 Jul 7.
12. Karlebach G, Shamir R. Modelling and analysis of gene regulatory networks. Nature Reviews Molecular Cell Biology. 2008 Oct;9(10):770.
13. Gutenkunst RN, Waterfall JJ, Casey FP, Brown KS, Myers CR, Sethna JP. Universally sloppy parameter sensitivities in systems biology models. PLoS computational biology. 2007 Oct 5;3(10):e189.
14. Marino S, Hogue IB, Ray CJ, Kirschner DE. A methodology for performing global uncertainty and sensitivity analysis in systems biology. Journal of theoretical biology. 2008 Sep 7;254(1):178-96.
15. Hasty J, McMillen D, Isaacs F, Collins JJ. Computational studies of gene regulatory networks: in numero molecular biology. Nature Reviews Genetics. 2001 Apr;2(4):268.
16. Oldroyd GED, Robatzek S. The broad spectrum of plant associations with other organisms. Current Opinion in Plant Biology. 2011;14.4:347-350.
17. Glazebrook J. Contrasting mechanisms of defense against biotrophic and necrotrophic pathogens. Annual review of phytopathology. 2005 Jul 28;43:205-27.
18. Voigt CA. Callose-mediated resistance to pathogenic intruders in plant defense-related papillae. Frontiers in plant science. 2014 Apr 28;5:168.
19. Brunkard JO, Zambryski PC. Plasmodesmata enable multicellularity: new insights into their evolution, biogenesis, and functions in development and immunity. Current Opinion in Plant Biology. 2017 Feb 1;35:76-83.
20. De Coninck B, Timmermans P, Vos C, Cammue BP, Kazan K. What lies beneath: belowground defense strategies in plants. Trends in plant science. 2015 Feb 1;20(2):91-101.
21. Zipfel C, Oldroyd GE. Plant signalling in symbiosis and immunity. Nature. 2017 Mar;543(7645):328.
22. Hein I, Gilroy EM, Armstrong MR, Birch PR. The zig-zag-zig in oomycete–plant interactions. Molecular Plant Pathology. 2009 Jul 1;10(4):547-62.

23. Pritchard L, Birch PR. The zigzag model of plant–microbe interactions: is it time to move on?. Molecular plant pathology. 2014 Dec 1;15(9):865-70.
24. Kiers ET, Denison RF. Sanctions, cooperation, and the stability of plant-rhizosphere mutualisms. Annual Review of Ecology, Evolution, and Systematics. 2008 Dec 1;39:215-36.
25. Gust AA, Willmann R, Desaki Y, Grabherr HM, Nürnberger T. Plant LysM proteins: modules mediating symbiosis and immunity. Trends in Plant Science. 2012;17.8:495-502.
26. Arrighi JF, Barre A, Amor BB, Bersoult A, Soriano LC, Mirabella R, et al. The Medicago truncatula Lysine Motif-Receptor-Like Kinase Gene Family Includes NFP and New Nodule-Expressed Genes. Plant Physiology. 2006;142.1:265-279.
27. Zhang X, Dong W, Sun J, Feng F, Deng Y, He Z et al. The receptor kinase CERK1 has dual functions in symbiosis and immunity signaling. The Plant Journal. 2014;81.2:258-267.
28. Buendia L, Wang T, Girardin A, Lefebvre B. The LysM receptor-like kinase SlLYK10 regulates the arbuscular mycorrhizal symbiosis in tomato. New Phytologist. 2016;210.1:184-195.
29. Lohmann GV, Shimoda Y, Nielsen MW, Jørgensen FG, Grossmann C, Sandal N et al. Evolution and Regulation of the Lotus japonicus LysM Receptor Gene Family. Molecular Plant-Microbe Interactions Journal. 2010;23.4:510-521.
30. Doyle JJ, Luckow MA. The rest of the iceberg. Legume diversity and evolution in a phylogenetic context. Plant Physiology. 2003 Mar 1;131(3):900-10.
31. Bai Y, Lindhout P. Domestication and breeding of tomatoes: what have we gained and what can we gain in the future?. Annals of Botany. 2007 Aug 23;100(5):1085-94.
32. Beddows I, Reddy A, Kloesges T, Rose LE. Population Genomics in Wild Tomatoes—The Interplay of Divergence and Admixture. Genome Biology and Evolution. 2017 Oct 24;9(11):3023-38.
33. Gheysen G, Mitchum MG. How nematodes manipulate plant development pathways for infection. Current Opinion in Plant Biology. 2011 Aug 1;14(4):415-21.
34. Stahl Y, Wink RH, Ingram GC, Simon R. A signaling module controlling the stem cell niche in Arabidopsis root meristems. Current Biology. 2009 Jun 9;19(11):909-14.
35. Hohm T, Zitzler E, Simon R. A dynamic model for stem cell homeostasis and patterning in Arabidopsis meristems. PLoS One. 2010 Feb 12;5(2):e9189.
36. Band LR, Fozard JA, Godin C, Jensen OE, Pridmore T, Bennett MJ, King JR. Multiscale systems analysis of root growth and development: modeling beyond the network and cellular scales. The Plant Cell. 2012 Oct 1;24(10):3892-906.
37. Band LR, Wells DM, Larrieu A, Sun J, Middleton AM, French AP, Brunoud G, Sato EM, Wilson MH, Péret B, Oliva M. Root gravitropism is regulated by a transient lateral auxin gradient controlled by a tipping-point mechanism. Proceedings of the National Academy of Sciences. 2012 Mar 20;109(12):4668-73.