UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

TRABAJO FIN DE MÁSTER

Computational analysis of a plant receptor interaction network

Máster Universitario en Bioinformática y Biología Computacional

Autor: Manosalva Pérez, Nicolás

Tutor: Youssef Belkhadir
Ponente: Carlos Aguirre Maeso
Departamento de Ingeniería Informática

Febrero 2020

Table of contents

Index of Figures

Index of Tables

Key words

LRR-RKs, Network Biology, Arabidopsis, Protein-protein interaction network, RNA-Seq.

List of abbreviations

CSI^{LRR}: leucine rich repeat-based sell surface interaction
ECD: extracellular domain
FPKM: fragment per kilobase million
GEO: Gene Expression Omnibus
GNR: gene regulatory network
KS: Kolmogorov-Smirnov
LCC: largest connected component
LRR: leucine rich repeat
LRR-RK: leucine-rich repeat receptor kinase
mRNA: messenger ribonucleic acid
NCBI: National Center for Biotechnology Information
PCA: principal component analysis
PPI: protein-protein interaction
RBE: rabbit ears
RK: receptor kinase
RNA-Seq: ribonucleic acid sequencing
TF: transcription factor
ZFP: zinc finger protein

Abstract

In all organisms, complex protein-protein interactions (PPI) networks control major biological functions yet studying their structural features presents a major analytical challenge. In plants, leucine-rich-repeat receptor kinases (LRR-RKs) are key in sensing and transmitting non-self as well as self-signals from the cell surface. As such, LRR-RKs have both developmental and immune functions that allow plants to make the most of their environments. In the model organism in plant molecular biology, *Arabidopsis thaliana*, most LRR-RKs are still represented by biochemically and genetically uncharacterized receptors. To fix this an LRR-based Cell Surface Interaction (CSI^{LRR}) network was obtained in 2018, a protein-protein interaction network of the extracellular domain of 170 LRR-RKs that contains 567 bidirectional interactions. Several network analyses have been performed with CSI^{LRR}. However, these analyses have so far not considered the spatial and temporal expression of its proteins. Neither has it been characterized in detail the role of the extracellular domain (ECD) size in the network structure. Because of that, the objective of the present work is to continue with more in depth analyses with the CSI^{LRR} network. This would provide important insights that will facilitate LRR-RKs function characterization.

The first aim of this work is to test out the fit of the CSI^{LRR} network to a scale-free topology. To accomplish that, the degree distribution of the CSI^{LRR} network was compared with the degree distribution of the known network models of scale-free and random. Additionally, three network attack algorithms were implemented and applied to these two network models and the CSI^{LRR} network to compare their behavior. However, since the CSI^{LRR} interaction data comes from an in vitro screening, there is no direct evidence whether its protein-protein interactions occur inside the plant cells. To gain insight on how the network composition changes depending on the transcriptional regulation, the interaction data of the CSI^{LRR} was integrated with 4 different RNA-Seq datasets related with the network biological functions. To automatize this task a Python script was written. Furthermore, it was evaluated the role of the LRR-RKs in the network structure depending on the size of their extracellular domain (large or small). For that, centrality parameters were measured, and size-targeted attacks performed. Finally, gene regulatory information was integrated into the CSI^{LRR} to classify the different network proteins according to the function of the transcription factors that regulate its expression.

The results were that CSI^{LRR} fits a power law degree distribution and approximates a scale-free topology. Moreover, CSI^{LRR} displays high resistance to random attacks and reduced resistance to hub/bottleneck-directed attacks, similarly to scale-free network model. Also, the integration of CSI^{LRR} interaction data and RNA-Seq data suggests that the transcriptional regulation of the network is more relevant for developmental programs than for defense responses. Another result was that the LRR-RKs with a small ECD size have a major role in the maintenance of the CSI^{LRR} integrity. Lastly, it was hypothesized that the integration of CSI^{LRR} interaction data with predicted gene regulatory networks could shed light upon the functioning of growth-immunity signaling crosstalk.

Abstract in Spanish

En todos los organismos, complejas redes de interacción proteína-proteína (PPI) controlan funciones biológicas de gran relevancia. Sin embargo, estudiar sus características estructurales aún presenta un gran desafío analítico. En las plantas, los receptores quinasas con repeticiones ricas en leucina (LRR-RK) son clave en la detección y transmisión de señales endógenas y exógenas en la superficie celular. Como tal, los LRR-RK tienen funciones de desarrollo e inmunes que permiten a las plantas aprovechar al máximo sus entornos. En el organismo modelo en biología molecular de plantas, *Arabidopsis thaliana*, la mayoría de los LRR-RK todavía están representados por receptores bioquímica y genéticamente no caracterizados. Para solucionar esto, se obtuvo una red de interacción de superficie celular basada en LRR (CSI^{LRR}) en 2018, una red de interacción proteína-proteína del dominio extracelular de 170 LRR-RK, que contiene 567 interacciones bidireccionales. Se han realizado varios análisis de red con CSI^{LRR}. Sin embargo, estos análisis hasta ahora no han considerado la expresión espacial y temporal de sus proteínas. Tampoco se ha caracterizado en detalle el papel del tamaño del dominio extracelular (ECD) en la estructura de la red. Por eso, el objetivo del presente trabajo es continuar con análisis más profundos de la red CSI^{LRR}. Esto proporcionaría información importante que facilitará la caracterización de la función de los LRR-RK.

El primer objetivo de este trabajo es comprobar el ajuste de la red CSI^{LRR} a una topología de libre escala. Para lograr eso, la distribución de grado de la red CSI^{LRR} se comparó con la distribución de grado de los modelos de red ya conocidos de libre escala y aleatorio. Además, se implementaron tres algoritmos de ataque de red y se aplicaron a estos dos modelos de red y a la red CSI^{LRR} para comparar su comportamiento. Sin embargo, dado que los datos de interacción CSI^{LRR} provienen de un cribado in vitro, no hay evidencia directa de si sus interacciones proteína-proteína ocurren dentro de las células de la planta. Para obtener información sobre cómo cambia la composición de la red en función de la regulación transcripcional, los datos de interacción del CSI^{LRR} se integraron con 4 conjuntos de datos de RNA-Seq diferentes relacionados con las funciones biológicas de la red. Para automatizar esta tarea, se escribió un script de Python. Además, se evaluó el papel de los LRR-RK en la estructura de la red dependiendo del tamaño de su dominio extracelular (grande o pequeño). Para ello, se midieron diferentes parámetros de centralidad y se realizaron ataques de tamaño específico. Finalmente, información sobre la regulación de genes se integró en el CSI^{LRR} para clasificar las diferentes proteínas de la red de acuerdo con la función de los factores de transcripción que regulan su expresión.

Los resultados fueron que la distribución de grado de CSI^{LRR} se ajusta a una ley potencial y se aproxima a una topología de libre escala. Además, CSI^{LRR} muestra una alta resistencia a ataques aleatorios y una resistencia reducida a ataques dirigidos a nodos concentradores / cuellos de botella, similar al modelo de red de libre escala. Asimismo, la integración de datos de interacción de CSI^{LRR} y datos de RNA-Seq sugiere que la regulación transcripcional de la red es más relevante para programas de desarrollo que para respuestas de defensa. Otro resultado fue que los LRR-RK con un tamaño ECD pequeño tienen un papel importante en el mantenimiento de la integridad de CSI^{LRR}. Por último, se planteó la hipótesis de que la integración de los datos de interacción de CSI^{LRR} con redes reguladoras de genes predichas podría arrojar luz sobre el funcionamiento de la diafonía entre la señalización por inmunidad y el crecimiento.

1 Introduction

1.1 Network Biology: approaches and applications

Networks have been used to describe interactions between entities in a wide array of different research areas, without biology being an exception[1]. Despite the success of research of individual cellular components and their function, only in rare cases a discrete biological function can be attributed to an individual molecule. Instead, most biological characteristics arise from complex interactions between the extensive cell constituents, such as proteins, nucleic acids and small molecules. This is where the Network Biology comes into play as a Systems Biology integrative approach to help understand the cell's internal organization and evolution[2].

The development of high-throughput data-collection techniques (genomics, transcriptomics, proteomics and semi-automated screens) has allowed to get snapshots of the status of the cell's components at any given time and how the interact with each other. From them, enough information can be extracted to create experimental interaction webs of different nature, like protein-protein interaction (PPI), metabolic, signaling and transcription regulatory networks. These four are the main network types in molecular biology, but for the purpose of this work we will only discuss in further detail the protein-protein interaction networks[2,3].

In PPI networks, the nodes are proteins, and two nodes are connected by a undirected edge if the two proteins bind[3]. Because of the nature of this relationship between the nodes (proteins), the links do not have an assigned direction, but a mutual binding relationship: if protein A binds to protein B, then protein B also binds to protein A[2]. The extraction of structural and topological features from a PPI could potentially provide information on individual nodes and edges, distinct modules, and the entire network. Some of these features include degree, the number of connections of a node; betweenness, the fraction of the shortest paths that pass through a node; and eigenvector, a measure of the influence of a node in a network[4]. However, to get a complete overview of the network organization and in which manner their elements are arranged, the different networks are classified according to their fit into certain mathematical models, such as random networks or scale-free networks[2] (figure 1A).

An important finding in the field of Network Biology was that most networks within the cell approximate a scale-free topology[5]. A network is considered scale-free if the degree distribution follows a power-law. This means that the degree of its nodes is highly non-uniform: most of its nodes have only a few connections, and only a few nodes have many connections (also called hubs). They are called scale-free because of the absence of a typical node in the network that could be used to characterize the rest. In scale-free networks, the probability that a node has k neighbors follows $P(k) \sim k^{-\gamma}$, where γ is the degree exponent. This degree exponent ranges between 2 and 3[2,3]. Scale-free networks are frequently discussed regarding network assembly mechanisms, particularly in the context of preferential attachment, in which the probability that a node gains a connection is proportional to its current degree k. Although preferential attachment is the most known

mechanism that produces scale-free networks, there are other mechanisms that can also produce them[6].

Another network model relevant in Network Biology is the random networks. The degree distribution of this type follows a Poisson distribution. In this case, there will be many nodes with a mean degree value, while nodes with extreme degree values (very high and very low) will be less represented[2].

1.2 Plant Leucine-Rich Repeat Receptor Kinases (LRR-RKs)

Plant growth and development are regulated by endogenous growth regulators, as well as by both beneficial and detrimental environmental cues. Hormonal, environmental, or pathogenic signals are mostly perceived by membrane-localized receptors that transduce those signals inside plant cells to activate genetic programs that regulate growth, development, and defense responses[7]. Among all of them, the receptor kinases (RK) constitute one of the largest gene families in the plant kingdom[8]. The whole family of RK consists of over 600 members and represents nearly 2.5% of protein coding sequences in the *Arabidopsis thaliana* (hereafter Arabidopsis) genome[9]. Upon ligand binding there is a dimerization or oligomerization of the RKs with either themselves or with a co-receptor. This leads to the activation of intracellular kinase domains (KD) which initiate downstream signaling transduction.[10] Protein kinases are a group of enzymes that move a phosphate group onto proteins, in a process called phosphorylation. This functions as an on/off switch for many cellular processes[11].

Based on the extracellular domain structure, plant RKs can be categorized into 14 subfamilies.[12] The leucine-rich repeat receptor kinase (LRR-RK) family of membrane integral receptors contains more than 200 members in Arabidopsis and is considered to be the largest family of plant receptor kinases.[13] The LRR-RKs are composed of three domains: an extracellular domain (ECD) containing tandem repetitions of a consensus sequence (enriched in residues of the hydrophobic amino acid leucine), a single membrane-spanning domain and an intracellular kinase domain (figure 1B). This kinase domain phosphorylates or self-phosphorylates the hydroxyl groups of serine or threonine on the proteins that start the signaling process.[10,14]

Regarding the extracellular domain of the LRR-RKs, many have been structurally characterized. This showed that LRR ectodomains or extracellular domains mostly function as a platform for ligand perception or either as co-receptor association. Based on the length of these ectodomains structure, plant LRR-RKs can be classified in two groups: the LRR-RKs with a large ectodomain and LRR-RKs with a small ectodomain[10]. Furthermore, it has been proposed in numerous studies that the ECD size of an LRR-RK could be used to predict whether they function as receptors or co-receptors[15].

In 2019 Xi *et al.* used this concept to classify the 225 members of the LRR-RKs. It was observed a bimodal distribution of the ECDs lengths with one maximum at 250 amino acids and another maximum at 550 amino acids. Based on this distribution, they defined the LRR-RKs with ECD length up to 400 amino acids as small, being putative co-receptors. ECDs with more than 400 amino acids were classified as large, which means putative ligand-recognition receptors[15].

LRR-RKs can mainly be grouped into either regulating plant growth and development or being involved in plant immunity and defense. Therefore, they are key proteins that regulate the interaction of the plant with the environment to secure its adaptability. However, most of them are still uncharacterized. In Arabidopsis only less than 20 RKs have a biochemically defined ligand and only less than 50 RKs have a genetically defined function[15,16]. Due the size of the LRR-RKs gene family (more than 200), a reductionist approach of trying to characterize the receptors individually is not practical. Consequently, the integrative approach that network biology offers is needed to understand more in depth this important but largely unknown group of genes.

1.3 LRR-based Cell Surface Interaction (CSI^{LRR}) network

In 2018 Smakowska *et al.*[17] published an extracellular network of Arabidopsis LRR-RKs termed LRR-based Cell Surface Interaction (CSI^{LRR}) network. It was implemented an extracellular protein-protein interaction assay previously stablished in an all-by-all screen of 200 formerly cloned LRR-RK ECDs. Given that the Arabidopsis genome encodes 225 LRR-RKs, it was tested the extracellular LRRs interaction space to a totality of 79%. The result was a PPI network that contained 170 proteins and 567 bidirectional interactions (figure 1C). It is important to note that only a 26.4% of the interactions tested passed the extremely stringent statistical cut-off for the network construction. This is the reason not all the 200 LRRs tested are included in the network[17].

In 2018 also Ahmed *et al.*[4] used this same network to discover pathogen contact points in host protein-protein interactomes. By computing network centrality parameters and integrating it with phenotypic data, it was possible to predict preferential targets of pathogen effectors. Effectors are pathogen proteins that are translocated inside the plant cells during infection to alter the cellular machinery in favor of the pathogen[4]. These predictions were also confirmed experimentally. All of this allowed to conclude that nodes with increased connectivity that are located closer to the network core are the preferred targets of pathogen attack. This result contrast with the centrality-lethality rule that states that high degree (hubs) and high betweenness (bottlenecks) nodes in a biological network are likely to be encoded by essential genes[4].

1.4 Motivation and Objectives

Since the CSI^{LRR} network was obtained through experimental *in vitro* assays, there is no direct evidence of whether these interactions occur inside the plant cells. There are many different layers of regulation inside the cell that condition two proteins being able to interact. The first one is the possibility of physical interaction, which has already been elucidated. The other ones are determined by transcriptional, translational, post-translational regulation and subcellular localization. The latter determines if the two proteins that are to interact find themselves in the same cellular organelle. However, given that all the proteins in the network are assumed to be extracellular membrane proteins, the subcellular localization should not be considered.

Nevertheless, there are other layers of regulation that influence the composition of the network inside the cell. This regulation layers affect the network spatially and temporally. This means that the network composition changes between the tissues of the plant, even

between cell types, but it also changes according to the different developmental stages of the plant. Another factor that can influence the network are the environmental conditions. Two plants with identical genetic composition could be regulating the network differently if they are in despair habitats.

In this work, the regulation layer to unveil in the transcription regulation. One of the main and most used techniques in profiling transcriptomes in biology is the RNA sequencing (RNA-Seq; ribonucleic acid sequencing). RNA sequencing is the application of any next-generation sequencing technique to profile all the present RNA in a biological sample at any moment[18]. One of the most studied RNA types is the messenger RNA (mRNA). It is the intermediary molecule between a gene and a protein. This way, profiling the mRNA in a sample gives direct evidence of the genes that are being expressed.

Under these assumptions, one of the main aims of the present work is to unveil how the transcriptional regulation layer affects the composition and structure of the CSI^{LRR} network in different plant tissues and experimental conditions using publicly available transcriptomics datasets. For this, a computational python-based tool has been developed in order to automatize the integration of PPI data and RNA-Seq[18] data. Other objectives of the work are the comparison of the CSI^{LRR} network with other known network models relevant in Network Biology to determine its fit into a scale-free network model, the analysis of the differential function of the LRR-RKs ECDs size inside the network, and an integration of gene regulation data with the PPI data already available.

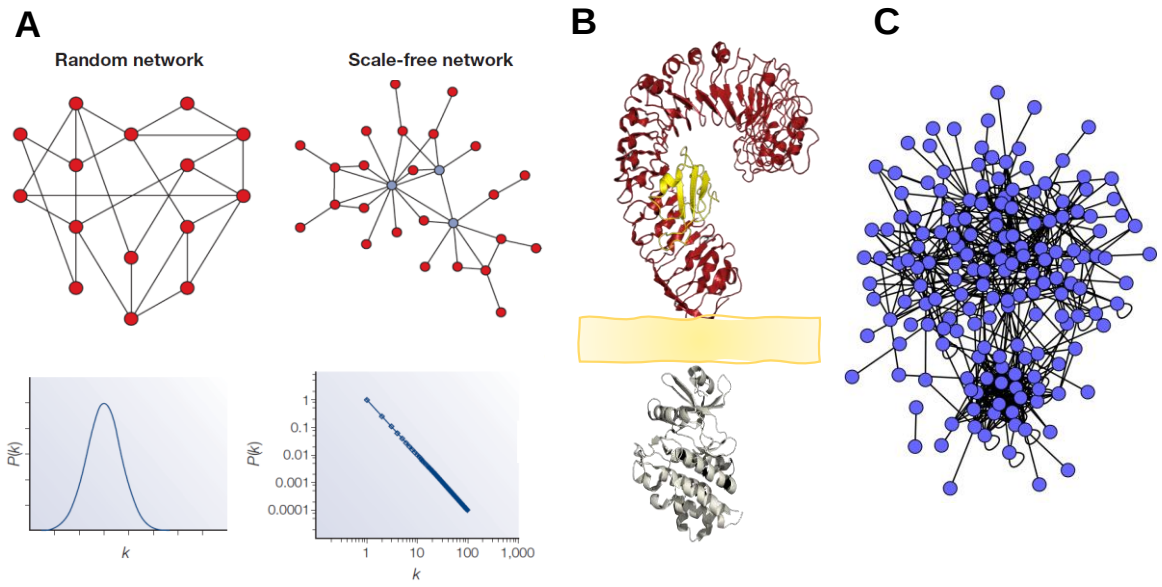

Figure 1. Network models, CSI^{LRR} and LRR-RKs structure. (A) Examples of a random network model, a scale-free network model and their respective typical degree distribution shapes, adapted from Barabási and Oltvai 2004[2] **(B)** Representation of the tridimensional structure of an LRR receptor kinase. In red there is a large extracellular domain, in yellow a small extracellular domain and in gray the intracellular kinase domain. **(C)** Representation of the CSI^{LRR} network with Cytoscape[19].

2 Methods

2.1 Datasets and databases

The protein-protein interaction data was obtained from Smakowska *et al.* 2018[17] supplementary data table 2 as an edge list. The RNA-Seq datasets were obtained from the Gene Expression Omnibus (GEO) of the National Center for Biotechnology Information (NCBI) with the following accession numbers: GSE90075 (*P. syringae* dataset), GSE63603 (flg22 dataset), GSE79709 (root dataset) and GSE38612 (organs dataset)[20–23] (figure 2). The expression values in the four datasets were in Fragments Per Kilobase Million (FPKM) and were not modified for the purpose of this work. FPKM is a variation from RPKM (Reads Per Kilobase Million)[24]. FPKM is a normalized estimation of the gene expression based on RNA-Seq data. They are computed from the number of fragments/reads that mapped a specific gene, considering the gene length and the sequencing depth.

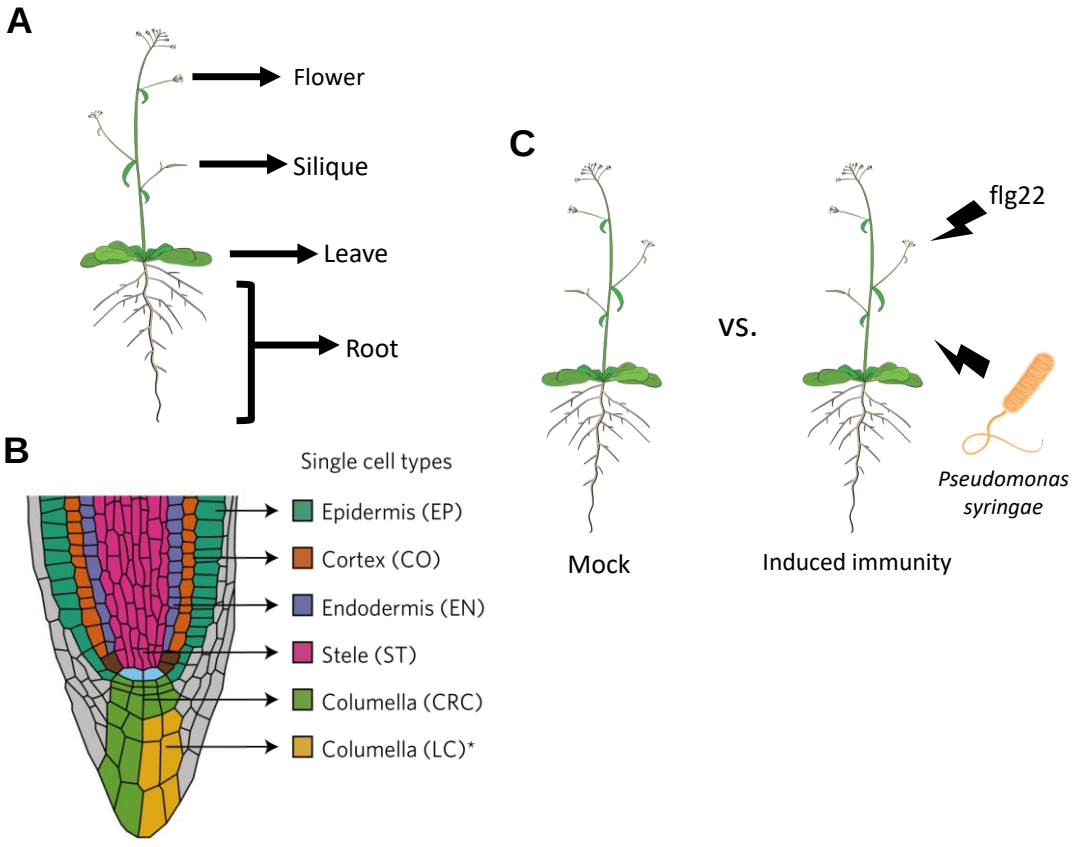

Figure 2. Datasets used in this work. (A) Schematic representation of the different *Arabidopsis thaliana* organs of which its transcriptome profiling composes the organs data set. **(B)** Schematic representation of the six root cell types used for the root datasets, adapted from Kawakatsu 2016[22]. **(C)** Scheme representing the two immunity-related datasets, in which mock plants and plants treated with the bacterial elicitor flg22 and with the plant pathogen *Pseudomonas syringae* to activate an effector triggered immunity (ETI) response.

2.2 Calculation of network parameters and network visualization

The parameter computation of all the networks in the work was done using the Python package NetworkX 2.5[25]. For the heatmaps and the Principal Component Analysis (PCA)[26]

plots a set of 18 network parameters were computed. In the appendix 1 table there are all the parameters, the functions used for its computation and a brief explanation of its meaning. The software Cytoscape 3.7.2 was used for the visualization of the networks[19].

2.3 Statistical analyses

For the statistical comparison of the resistance-damage value in the network attacks and the small/large centrality parameters it was used a one-sided non-parametric Mann-Whitney U test[27] given the distribution of the samples. For the small/large targeted attacks it was used a Welch T-test[28]. To test the goodness of the fit of CSI^{LRR} to a power-law distribution it was used a two-sided Kolmogorov–Smirnov (KS) test[29]. All these test were done using the Python 3.7 libraries Scipy and powerlaw[30,31].

2.4 Generation of scale-free and random networks

The degree distribution of one hundred scale-free networks and one hundred random networks was computed using the NetworkX functions networkx.barabasi_albert_graph(169, 3) and network.gnm_random_graph(169, 567). Then these one hundred-degree distributions were averaged and plotted along with the degree distribution of the CSI^{LRR} network.

2.5 Attack algorithms implementation

The three attack algorithms were implemented in Python from the pseudocode published in Aguirre *et al.* 2002[32]. In table 1 there is a summary of how the three attack algorithms work. For the random attacks to all networks and the degree/minimum path attacks to the scale-free and random networks, the functions were iterated 100 times to reduce the variance in the results. It is important to note that in the case of the small/large targeted attacks the algorithms were slightly modified. Instead of receiving as an input all the network nodes as potential objects to be disabled, they are first filtered according to their size, so only the nodes of one type can be deleted by the algorithm. In the appendix 2 can be found the python attack functions implemented for this work.

Table 1. Network attack strategies used in this work.

Random attack algorithm	Degree attack algorithm	Minimum path attack algorithm
Nodes are targeted and deleted from the network in a random manner.	Nodes are targeted and deleted from the network according to their number of connections.	Nodes are targeted and deleted from the network according to the number of shortest paths in which they are.

2.6 Automatization of the integration of PPI data with RNA-Seq data

For the automatization of the integration of PPI data and RNA-Seq data it was implemented a Python 3.7 script. The script file and a tutorial can be found in https://github.com/nicomaper/intransnet . For information on how to use the program see the tutorial in the previous link. For further information on the characteristics of the program see chapter Results 3.3 and figure 5.

2.7 Cut-off expression value computation

Technical noise is unavoidable in RNA-Seq experiments and it must be quantified in order to avoid mistaking it for genuine differences in biological expression levels[33]. Thus, determining a cut-off expression value to discriminate genes that have an FPKM higher than 0 due to technical noise is key for the integration of RNA-Seq data with PPI data. To implement this concept into the tool developed during this work, there is a cut-off FPKM value that can be given in the form of a numerical value or automatically computed. For the automatic computation the whole RNA-Seq dataset must be provided.

Considering that in an RNA-Seq experiment data there is technical noise, a bimodal distribution of the logarithmic FPKM expression values should be expected[34]. If this is the case, the cut-off value that would discriminate the technical noise from the actual expressed genes would be in between of the two normal distributions. To find such value from a whole RNA-Seq data it was implemented in Python a function that uses Gaussian Mixture Model clustering algorithm from the scikitlearn module to differentiate the two normal distributions[35]. After assigning each expression value to a cluster, the function takes the maximal and the minimum number of the first and second cluster respectively and computes the exponential of the mean. This number is used as cut-off value to categorize which genes are considered expressed, and therefore, kept in the network, and which ones are noise and then eliminated from the network. In the appendix 3 there is the Python implemented function for the cut-off computation from an RNA-Seq whole dataset.

2.8 PPI integration with GRN

The TF2Network software was used to predict the transcription factor (TF) regulators of CSI^{LRR} network proteins. This tool exploits the large volume of TF binding information and allows the prediction of gene regulatory networks (GNR) by identifying potential regulators for a set of functionally related genes[36]. In this case that list of genes were the LRR-RKs that shape the CSI^{LRR} network.

Once TF2Network was run with the genes that encode the network proteins, the results returned were filtered by their q-value and the functional annotation of the TF. First, the three TF with the lowest q-value that had been functionally annotated as involved in plant development. Then, the three TF with the lowest q-value and immunity-related annotations. This gene regulatory information was integrated into the CSI^{LRR} network. This way, the nodes or proteins in the network were classified the three different groups:

proteins regulated only by development-related TFs, proteins regulated only by immunity-related TFs, and proteins regulated by both types of TFs.

3 Results

3.1 CSI^{LRR} fits a power-law degree distribution but is not scale-free

To gain more insight into the characteristics of the CSI^{LRR} network, it was made a comparison with other known network models relevant in Network Biology: random networks and scale-free networks. Since this network topologies are mainly defined by their degree distribution, the averaged degree distribution of 100 random networks and 100 scale-free random networks was compared with the degree distribution of CSI^{LRR}. As expected, the random network has a normal distribution while the scale-free network has a power-law degree distribution. As observed in figure 3, the CSI^{LRR} network has a heavy-tailed distribution which could approximate a power-law.

Then a linear regression was fit to the obtained degree distributions in a logarithmic scale. The scale-free network model has an R-squared of -0.93 with an associated p-value less than 0.001. In the case of the CSI^{LRR} network the R-squared is -0.92 and the associated p-value is as well less than 0.001. This proves that, similarly to scale-free networks, CSI^{LRR} possesses many nodes with low degree and few nodes with high degree. However, to verify whether it is a scale-free network, first it needs to be determined whether a power-law distribution fits the CSI^{LRR} degree distribution and to determine the value of the γ parameter in such model, which should be between 2 and 3. With a Kolmogorov-Smirnov test the CSI^{LRR} degree distribution was compared with a fitted power-law distribution. The test yielded a p-value of less than 0.001. This indicates that the CSI^{LRR} distribution fits well a power-law distribution. However, the computed value of γ for such model is 3.32. This means that the γ value, one strong requirement to consider a network scale-free[6], fails to be fulfilled.

3.2 CSI^{LRR} displays high tolerance to random attacks and reduced tolerance to hub/bottleneck-directed attacks, similarly to random scale-free network models.

To study and compare further the behavior of the CSI^{LRR} network and the other two models (scale-free and random), three different algorithms were implemented to perform attacks. A network attack is a set of network objects (in this case, nodes, but could also be edges) which are disabled or deleted from the network. Its purpose is to produce damage in terms of connectivity[32]. For the attacks performed in this work, the damage has been defined as the order (number of nodes) in the largest connected component (LCC) after the attack[32]. Therefore, the resistance of a network to an attack is the proportion of initial nodes that the LCC has after the attack. The cost of an attack has been defined as the number of objects (nodes) that the algorithm disables, the proportion of total nodes in the network that the algorithm deletes. In table 1 there are summarized the three different algorithms that were implemented. Each of them targets nodes in a different manner, either randomly, according to their degree or to the number of minimum paths that cross them.

In response to the random attacks, the three networks behave similarly. In all the cases, the resistance to the network attack seems to be proportionate with the cost of the attack. Also, there were not significant differences when comparing the product resistance-cost of the random attacks to the three different networks.

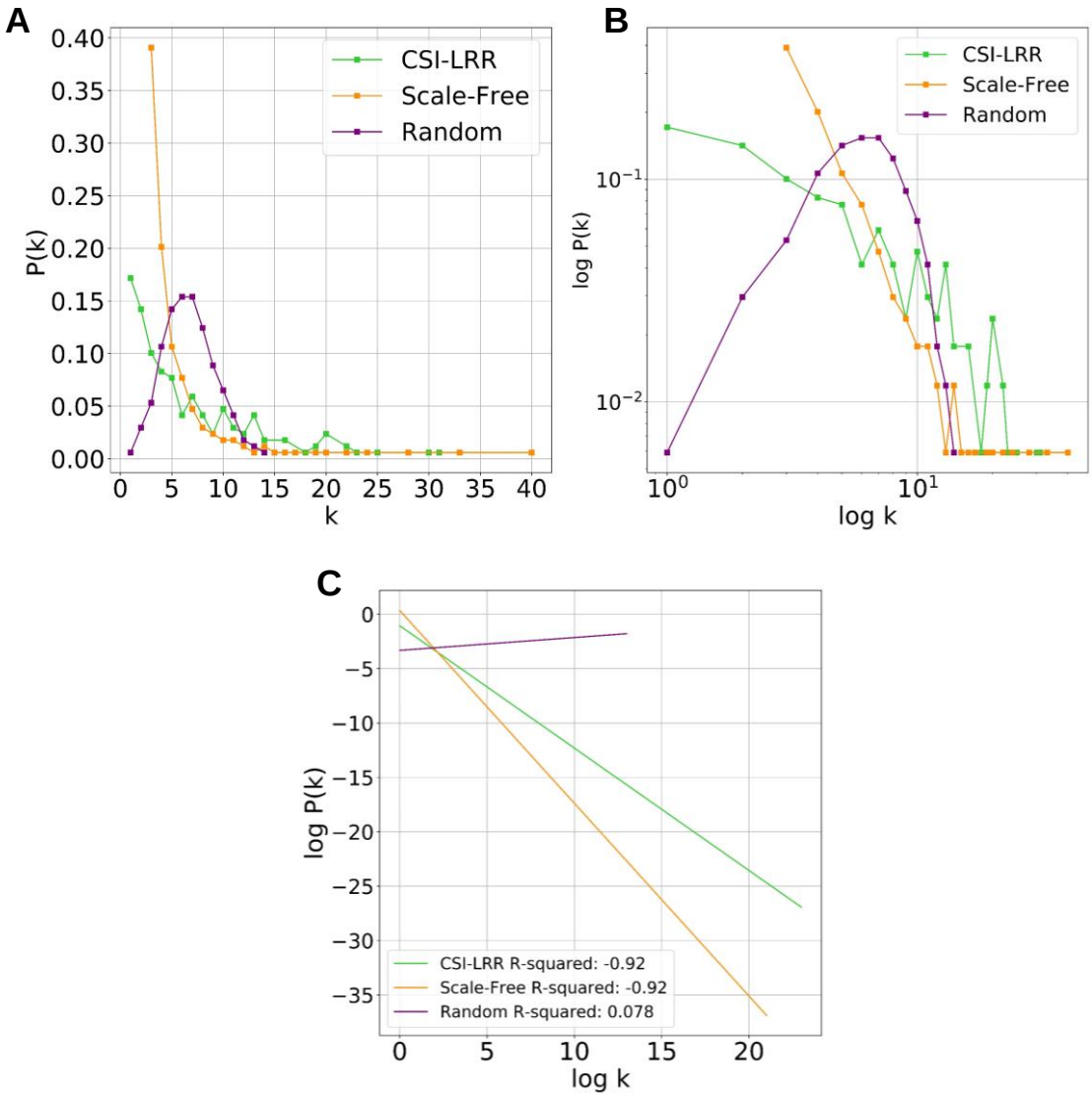

Figure 3. Degree distribution and linear regression of CSI^{LRR}, scale-free and random network models. (A) Degree distribution of the three networks used in this work. For the random and scale-free models, the degree distribution is the average of 100 random versions of each. **(B)** Degree distribution in a logarithmic scale. **(C)** Linear regression of the logarithmic degree distribution of the three networks. The associated p-values are $7.012*10^{-11}$ for CSI^{LRR}, $5.4*10^{-10}$ for scale-free and 0.79 for the random network.

However, when the hubs (i.e. the nodes with higher degree) of the network are directly targeted, there is an important change in the behavior of all the networks, especially in the case of the scale-free network model and the CSI^{LRR} network. For the CSI^{LRR} network, when deleted approximately a 35% of the nodes with most degree, the network has been completely disconnected and disabled. In this case, the scale-free network model and the CSI^{LRR} network display a higher sensitivity to the degree targeted attacks than the random network model, which seems to be more resilient. For this case there were significant differences between the product damage-cost of the CSI^{LRR} network and scale-free network with the random network.

The results are very similar when the bottlenecks of the networks are targeted. When the algorithm targets the nodes that are in the higher number of minimum paths, the damage to the connectiveness of the network is higher, especially for the CSI^{LRR} network. The results for the statistical differences are the same in this case as for the degree attacks.

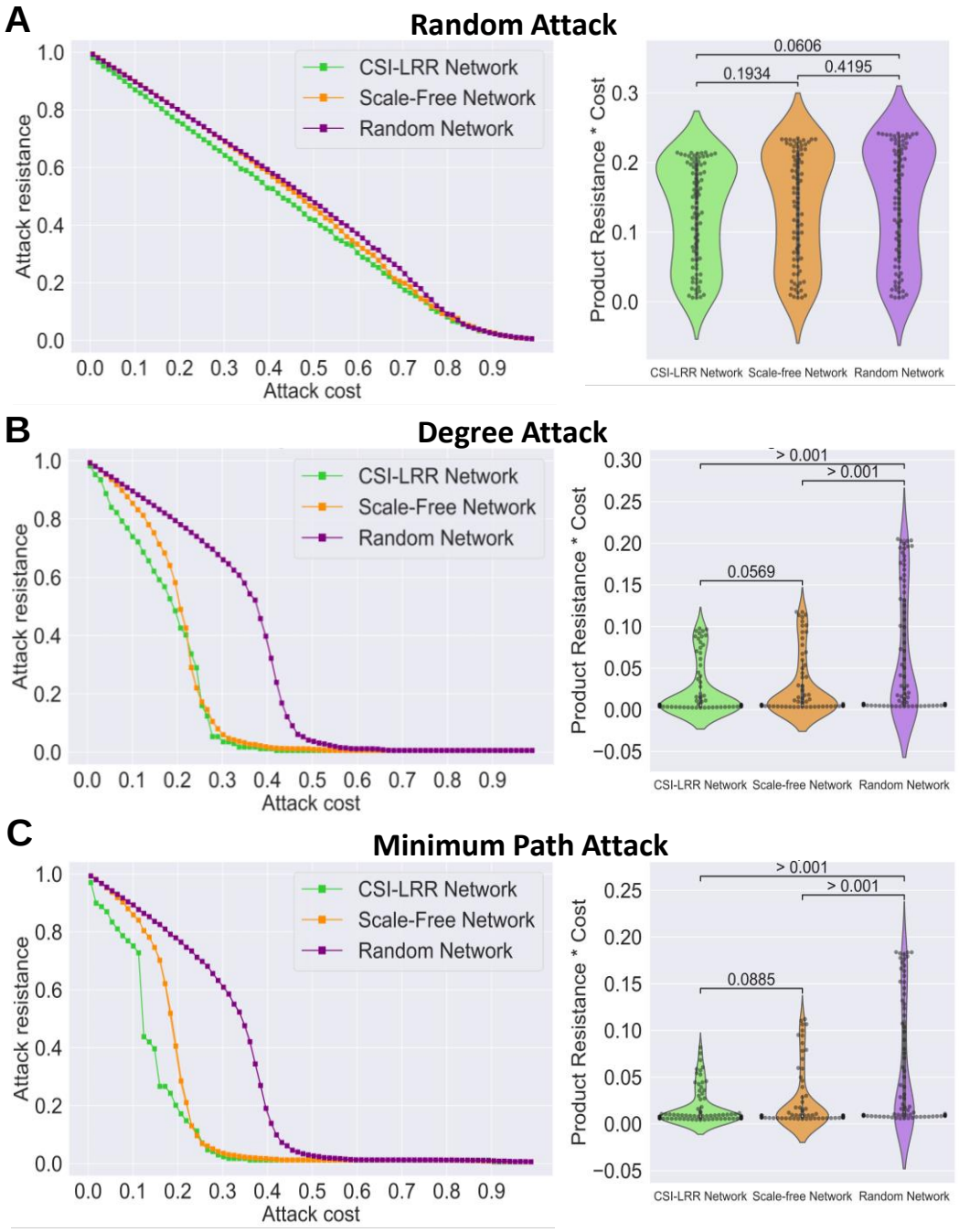

Figure 4. Network attacks. **(A)** Random attack on CSI^{LRR} network and on scale-free and random network models. The results presented correspond to the average of 100 attacks. The three networks present the same resilient behavior to random attacks. There is a direct relationship between the cost of the attack and the decrease in the resistance. **(B)** Degree attacks on the same previously mentioned networks. In this case the decrease in the resistance is larger since the nodes with a higher degree are being targeted. The decrease in the resistance is more dramatic for the CSI^{LRR} and the scale-free network than for the random network due to its lack of a structural bias. **(C)** Minimum path attack strategy, in which the nodes that are in largest number of shortest paths are targeted. The decrease is the resistance is similar for the CSI^{LRR} and the scale-free network, which denotes similarity in their topologies.

3.3 Development of a computational tool to integrate PPI data with RNA-Seq data.

To automatize the integration of PPI data and RNA-Seq data, a command line-based Python script was written. In the figure 5 there is a scheme of the functioning of the program, as well as their input arguments and its outputs. The objective of this tool is to simplify the network deleting the nodes that have expression values below a cut-off and are considered not expressed under certain experimental conditions or tissues/cell types. This way, for each RNA-Seq dataset new networks will be generated that lack the nodes of the proteins which genes are not being expressed.

The first main input argument is a PPI network in the form of an edge list. The second is a data frame with the expression values of the network protein's genes for a set of experimental conditions/tissues/cell types from an RNA-Seq experiment. The third input argument is a cut-off expression value either given in the form of a single number, or that can be computed automatically by providing the expression values of the whole RNA-Seq experiment. Both ways of providing the cut-off value are mutually exclusive but at least one of them mandatory. The purpose of the cut-off expression value is to determine what genes are really expressed and which ones have an expression value due to the technical noise. For further information on the cut-off value computation see chapter Methods 2.7.

Once provided the three mandatory input arguments, the program computes a network for each sample. Then a set of 18 network topology parameters are computed that are later used for statistical comparison, for principal component analysis[26] and for data visualization.

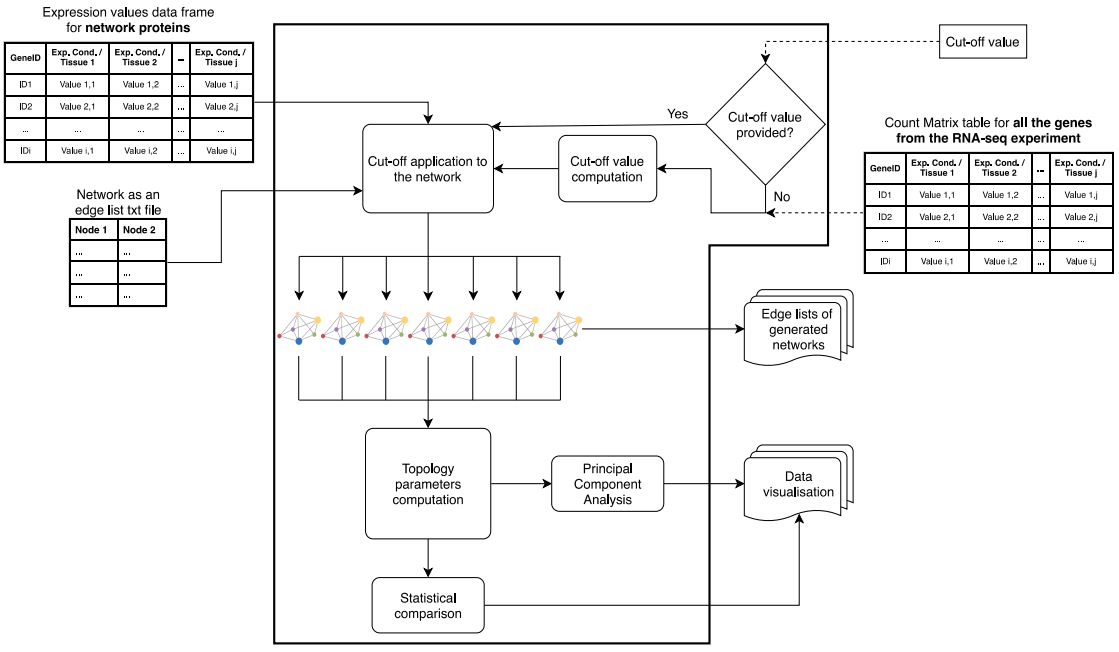

Figure 5. Scheme of the functioning of the script to integrate PPI and RNA-Seq data. The Python script implemented for the integration of PPI data with RNA-Seq data takes as input arguments: the network as an edge list, a data frame with the mRNA expression values in FPKM of the genes in the network and a cut-off expression value in the form of a number. The cut-off value can also be computed if the whole RNA-Seq data frame is provided. The main output of the program are the edge lists of the new networks generated by eliminating the nodes in the original network that do not pass the cut-off. Also, some plots are generated for visualization of the data.

3.4 Integration of CSI^{LRR} interaction data and RNA-Seq data suggests that the transcriptional regulation of the network is more relevant for developmental programs than for defense responses.

LRR-RKs can mainly be grouped into either regulating plant growth and development or being involved in plant immunity and defense[15]. Because of this, the Python script was tested with two development-related and two immunity-related RNA-Seq datasets. One of the developmental datasets makes a transcriptome profile of four Arabidopsis organs: root, leaf, flower and silique[23], while the other is from different Arabidopsis root cell types[22]. The immunity datasets are the transcriptome profiling of the Arabidopsis ecotype Columbia-0 that has been treated either with the plant pathogen *Pseudomonas syringae*[20] or with the bacterial peptide flg22[21]. In both treatments it is expected that the plant responds triggering an immune response[37].

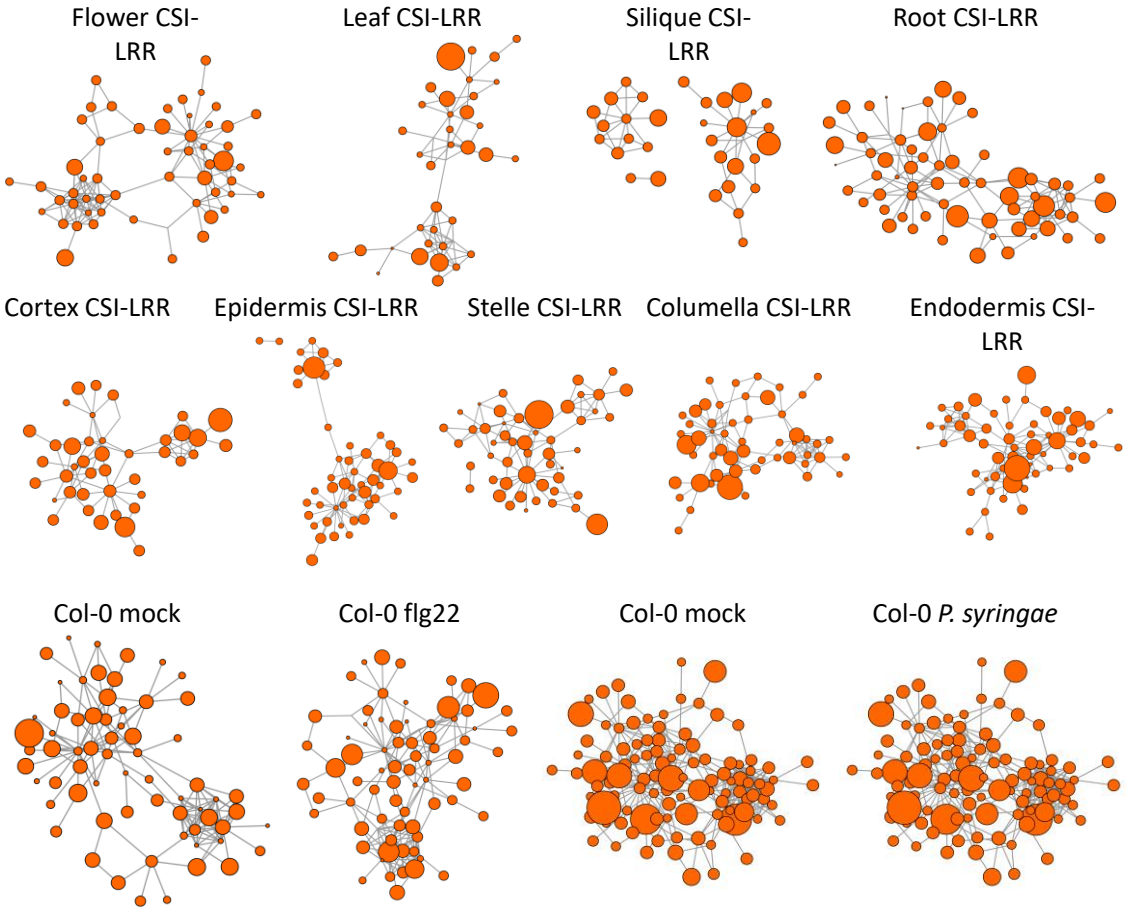

Figure 6. Networks generated during this work. Networks generated from the integration of the PPI data from CSI^{LRR} and the RNA-Seq data sets mentioned in the Methods section. The node size is relative to the expression of the gene that encodes the protein in each data set. It is noticeable just from the network representation that the networks integrated the immunity datasets are more complex than the ones that come from developmental datasets. The visualization software was Cytoscape[19].

In figure 6 there are the generated networks for each dataset. As mentioned previously, from the generated networks a set of 18 network parameters were computed. To this data frame of new networks and their as sociated parameter values was applied an algorithm of

dimensionality reduction of principal component analysis[26]. The purpose of this was to find the parameters that contained the largest amount of variability and make the results visualizable in two dimensions. The results of this are in figure 7.

For the first developmental dataset of Arabidopsis organs, the networks have been surprisingly reduced in order and size (appendix 4) compared to the original. This suggest a tight transcriptional control of the CSI^{LRR} network in these tissues. Especially in silique, the network with the least order and that has been disconnected (figure 6). Furthermore, when looking at the percentage of expressed genes for this dataset, despite that the values are about homogeneous, the silique network has a reduced number of nodes (appendix 5). It is also noticeable that in the leaf network there is an articulation point (and edge that being removed would disconnect the network), which affects parameters related with network distance (appendix 4). From the PCA plot it can be deduced that while the flower and the root are the organs with the most similar networks, the silique is the most divergent from all the others.

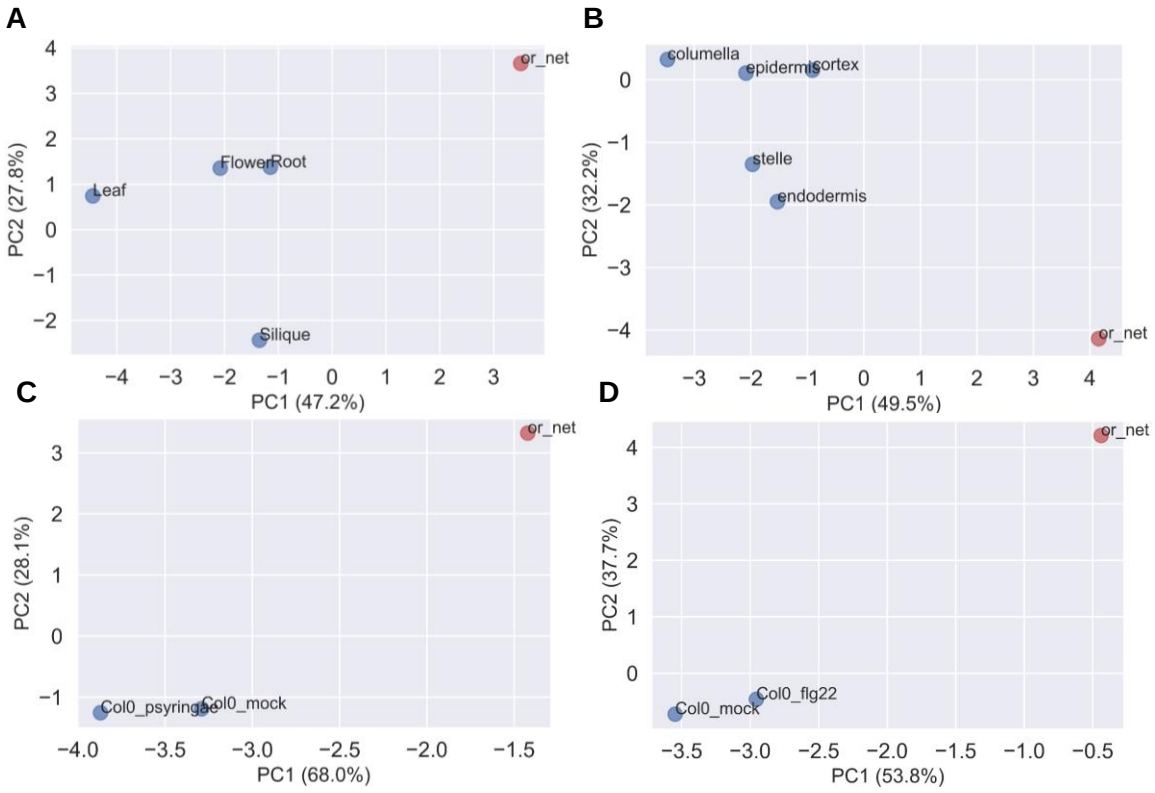

Figure 7. Principal component analysis plots. 18 network parameters for each network were computed, from which a PCA was performed. **(A)** PCA for the organs data set. While the Flower and Root network are the most similar ones among them and with the original network, the Silique network seems to be the most different one. **(B)** PCA for the root data set. In this case the networks seem to be more similar between them, probably because all the networks are from cell types from the same tissue. **(C)** PCA for the *P. syringae* data set. **(D)** PCA for the flg22 data set. In both (C) and (D) the networks are highly similar between them and have are more similar with the original network than in the developmental PCAs.

For the second developmental dataset of RNA-Seq from root cell-types, it can be seen in the PCA plot a higher similarity between the networks, probably because they all belong to the same organ. The most similar networks seem to be endodermis and stelle, two neighbor tissues, while the epidermis and the cortex, also neighbor tissues, seem to be more related. According to the heatmap in the appendix 4 and figure 6B, the most dissimilar network is

the one from columella, the most distinct tissue from all the others due to its especial functions in gravity perception[38].

For the immunity datasets, in both cases there is not much difference between the mock and the immunity-induced networks. The amount of variance in the principal components is 96.1% for the *P. syringae* dataset and 91.5% for the flg22 dataset, and yet they seem to be remarkably close to each other. From the heatmaps in the appendix 4 it is observed a high similarity between the two generated networks parameters.

These results suggest, on one hand, that the transcriptional regulation of the network genes under conditions where the immunity is triggered is not very relevant for the network composition. This makes sense given the rapidity and urgency of the innate immunity, which would require faster and finer mechanisms such as post translational modifications, rather than transcriptional regulation. On the other hand, when it comes to transcriptional regulation of the network in developmental programs, the networks show significant differences, suggesting than in this case, it would play a key role for organ differentiation.

3.5 ECDs with a small size have a major role in the maintenance of the CSI^{LRR} integrity.

For the characterization of the functionality of the proteins in the network according to their size, the first approach was to compute different centrality parameters according to the size label of each node. This showed that the proteins in the network that have a small ECD, despite of being the minority, are hubs and bottlenecks in a higher proportion than proteins that have a large ECD (102 large nodes vs. 70 small nodes). In the all five centrality parameters computed (degree, betweenness, clustering coefficient, eigenvector centrality and load centrality) there were significant differences between the values of the small and the large nodes. The nodes with a small ECD displayed a higher average value in the parameters computed.

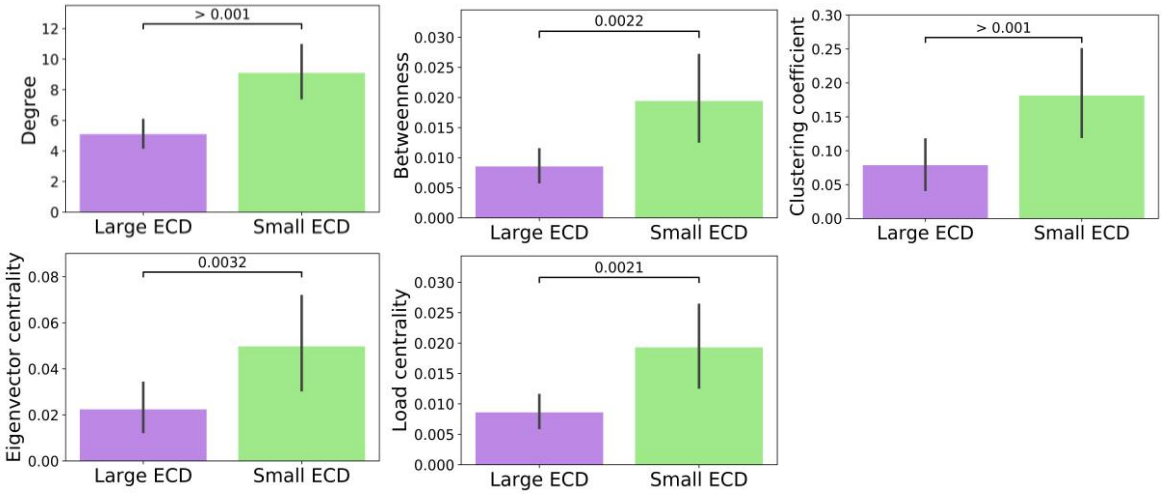

Figure 8. Centrality parameters for small and large nodes in CSI^{LRR}. Mean values of degree, betweenness, clustering coefficient, eigenvector centrality and load centrality for the small and large nodes in the CSI^{LRR} network. The statistical test applied was a one-sided Whitney-Mann U test. In all cases it was considered a statistical significance given that all the p-values are below 0.001.

To study these differences further, the next step was to perform size-targeted attacks on the CSI^{LRR} network. This means, that the three implemented attack algorithms mentioned previously were used in the network, but they were modified in order to target only nodes of one type: either small or large. The result from this was that with the three attack algorithms, the resistance of the network to the attacks was lower when the small nodes were targeted. Also, in the three attack types, there were significant differences from the product of the resistance and the cost when the small nodes are targeted than when are the large. These results strongly suggest that the small nodes play a differential and key role inside the network when it comes to maintenance and flow of information.

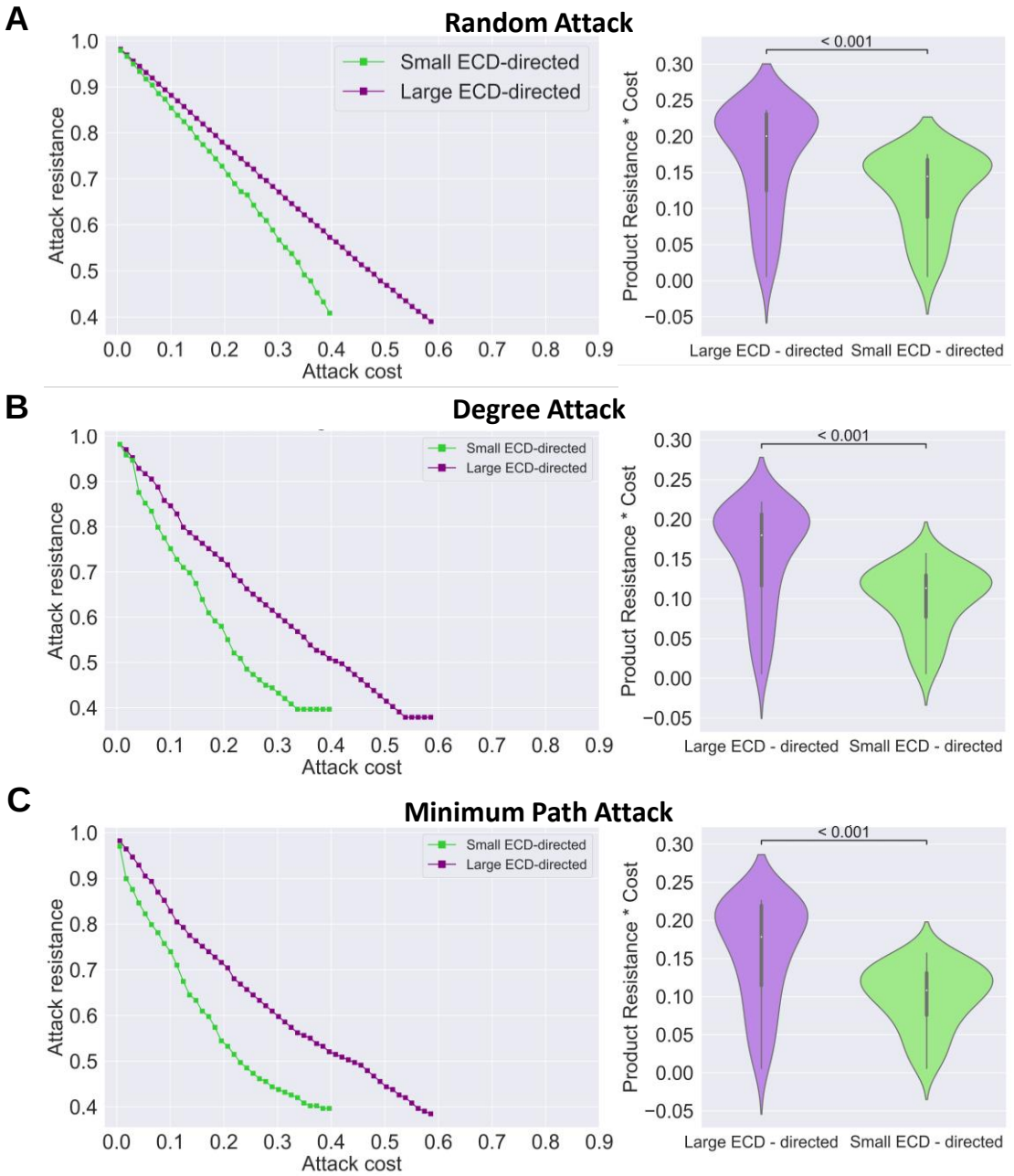

Figure 9. Small and large targeted attacks on the CSI^{LRR} network. (A) Random targeted attack to either small or large nodes. This corresponds to the mean of 100 attacks. The network has a less resistance when the small nodes are targeted randomly, which shows their importance in the connectivity of CSI^{LRR}. **(B)** and **(C)** are the degree attack and minimum path attack targeted attacks. Again, the resistance of the network is more affected when only the small nodes are targeted.

The k-core of a network is a subnetwork that contains the nodes with a degree k or more that are connected. The CSI^{LRR} max k-core is the 10-core and it contains an observed frequency of 0.65 small nodes, and a 0.35 frequency of large nodes. This data contrasts with the expected frequency of each node type, which is 0.4 for the small nodes and 0.6 for the large nodes. This provides one final piece of evidence that the proteins that have a small ECD in the network have a more important role for the connectivity of the network.

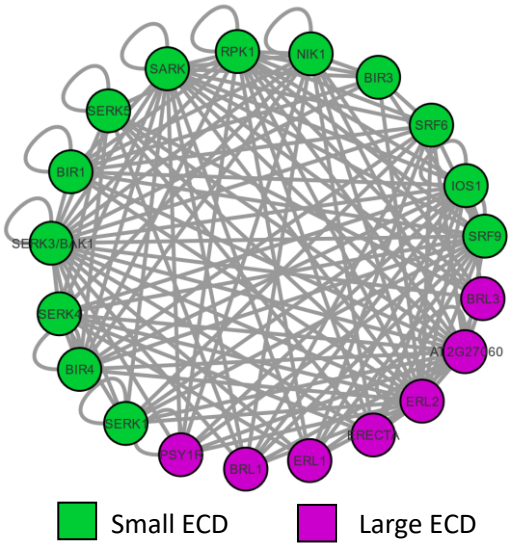

Figure 10. 10-core of CSI^{LRR} and small/large node composition.

Table 2. Expected and observed frequencies of the small and large ECD nodes in CSI^{LRR}

	Small ECD	Large ECD
Expected frequency in max k-core	0.4	0.6
Observed frequency in max k-core	0.65	0.35

3.6 Integration of CSI^{LRR} data with predicted GRN could shed light upon the functioning of growth-immunity signaling crosstalk.

To understand in more detail how the CSI^{LRR} network works and what the function of its constituents is, the final part of the present work is to integrate gene regulatory data into the CSI^{LRR} network. For this, the known interaction data was combined with predicted data of the transcription factors that regulate protein network genes. The predicted data for this gene regulatory network was obtained using the tool TF2Network. This program takes experimental transcription factors binding site and differential expression data and predicts what are the main transcription factors that regulate a set of genes[36]. That is, in this case, all the protein network genes.

Since the LRR-RK can be mainly grouped into either regulating plant growth and development or being involved in plant immunity and defense[15], the list of transcription

factors returned by TF2Network were categorized on being immunity or development related. This tool provides the functional annotation of the input and the predicted TFs, which facilitated the categorization. For simplification purposes, only the three transcription factors with the lowest q-value for each group were selected. The selected immunity-related transcription factors were WRKY33, WRKY50 and WRKY51. The WRKY transcription factor family has been well characterized as being involved in defense response in plants[39]. The selected development related transcription factors were ZFP5, ZFP8 and RBE[40,41].

The result of the integration of PPI data and GRN data allowed to determine three different groups inside the CSI^{LRR} network. On one hand, there is a group of nodes or proteins that are regulated only by the development-related transcription factors and other group only regulated by the immunity-related transcription factors. On the other hand, there is a numerous group of nodes inside the network that are regulated, at least, by one transcription factor of each group. From the biological point of view, this is a very interesting group inside the network. In this group there could be candidates responsible for the integration of environmental signals that determine the plants decision of whether to defend against a potential pathogen or to continue with the growth and development[42].

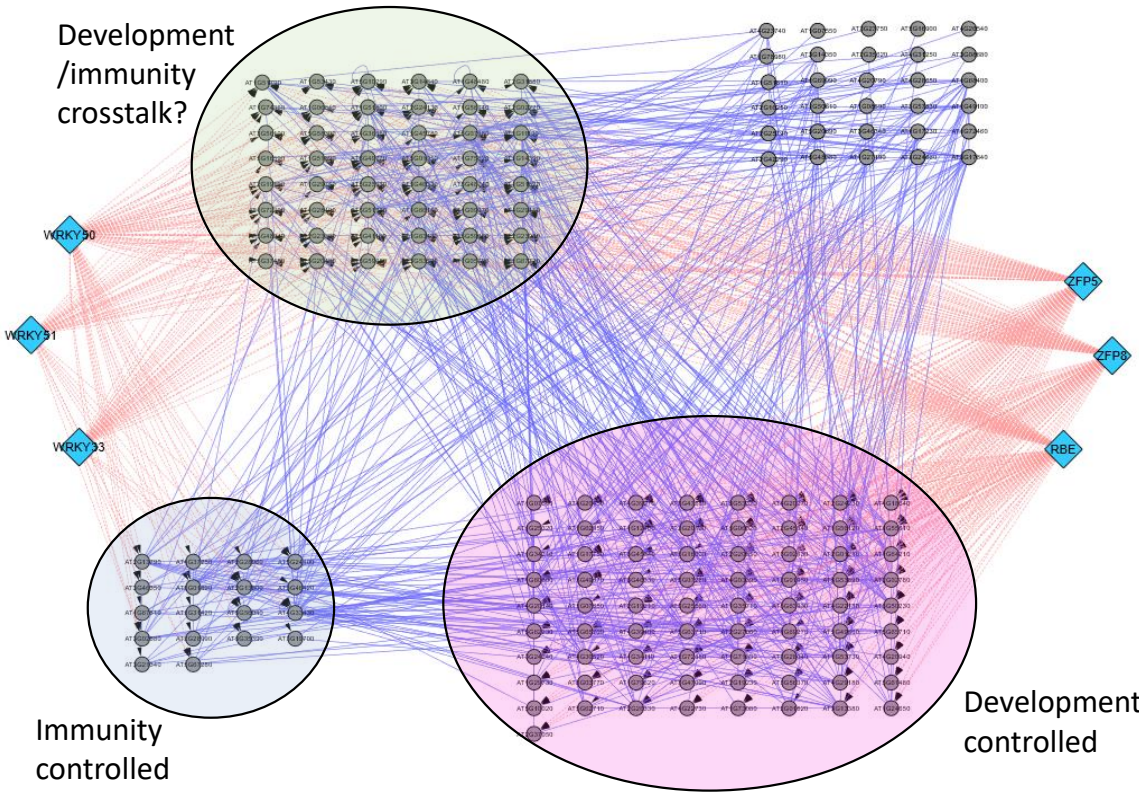

Figure 11. Integration of CSI^{LRR} interaction data with gene regulatory network data from TF2Network predictions. The transcription factors returned by TF2Network as regulatory of the CSI^{LRR} network proteins were categorized according to their functional annotation in two groups: immunity-related and development-related. Thus, the proteins they regulated inside the network were clustered according to the function of the TF that regulates them. This gave three cluster: proteins immunity controlled, proteins developmentally controlled, and proteins controlled by both processes. It was speculated that these proteins could be involved somehow in the crosstalk growth-defense.

4 Discussion and conclusions

The massive progress that the high-throughput technologies are experimenting along with the improvements in robust statistical and analytical tools have allowed a huge development in the field of Systems Biology in the recent years[43]. The traditional reductionist approach in molecular biology has generated a vast amount of knowledge but it fails to understand the large-scale interactions of the individual components within cellular and environmental contexts[5].

Network Biology seeks to lessen those limitations of the reductionist approach by implementing an integrative or holistic approach view of the genes, proteins and metabolites. The graph framework of considering cell constituents as interacting elements can be exploited to comprehend more deeply how cellular systems work and to make predictions for experimental validation. An example of this is the use of Network Biology to predict pathogen contact points in plant PPI networks that were experimentally validated[4]. Other cases are when the centrality of a host-pathogen interactome was associated with pathogen fitness during infection or when the structural robustness of mammalian TFs networks revealed plasticity across development[43,44]. In this occasion the CSI^{LRR} has been analyzed more in depth. The aim was to understand better the functioning of the large family of plant receptors LRR-RKs. They are key in the integration of environmental signals to help the plants adapt better to the habitats they find themselves in[17].

In the present work the CSI^{LRR} network is further analyzed and compared with other known network models in biology like the scale-free network and the random network. From this it was determined that the CSI^{LRR} degree distribution fits a power-law distribution, but it cannot be considered strictly a scale-free network, since the value of its γ parameter is not between 2 and 3. However, it could be said that the CSI^{LRR} approximates an scale-free network model, as it is common in many biological networks[3]. However, it is yet to determine whether this power-law degree distribution of nodes is due to a bias in the technologies currently used to determine *in vitro* protein-protein interactions, like the yeast two-hybrid or the affinity purification along with the mass spectrometry[4]. However, despite the experimental procedure to obtain a biological network, many large-scale interactomes and PPI networks display scale-free properties. The evolutionary origin of the scale-free networks in biology is still unknown, but it has been hypothesized that the growth and preferential attachment is probably rooted in gene duplication. Genes that have been duplicated produce the same proteins and therefore share interacting partners[2].

Regarding the different network attacks performed on CSI^{LRR} and on scale-free and random models, the results show that the three of them are resistant to random attacks. In the case of the random networks they tend to be very resilient to different attacks strategies. This is because they do not have any structural bias that the attack algorithm can take advantage of to dismantle the network. For the three attack algorithms the CSI^{LRR} network displayed a behavior very similar as the scale-free network. This is because of the power-law distribution of their degree. Only a minor proportion of its nodes have a large degree, therefore a large importance. The chances that a random attack disables the

important nodes are lower than for nodes with a low degree. This provides more evidence that the CSI^{LRR} approximates a scale-free network model.

The integration of PPI data with RNA-Seq expression data is expected to provide more information about the dynamical behavior of a network *in vivo*. In this work a Python script was implemented to tackle such task with the CSI^{LRR} network. Two types of datasets were used to test the script related with the functions associated with the network: development and immunity[10]. The results showed that for developmental datasets the changes in the network were more dramatic than when immunity-related datasets were used. This suggests that the layer of transcriptional regulation that controls the network composition is more relevant for developmental programs than for immunity responses. From a biological perspective, on one hand, makes sense that the transcriptional control in differentiated tissues and organs is tighter than for transient and rapid responses, like the immune response in plants triggered by effectors. It is important to notice that the disease in plants is an exception rather than the rule. On the other hand, severe expression changes of the network genes have been reported in Arabidopsis after treatment with flg22, the bacterial immunity elicitor[21].

It is also relevant to consider that the choice of a cut-off expression value is a difficult task to handle, especially in cases where the genes under study have low expression values, like the Arabidopsis LRR-RK. Finally, in this work it has only been studied how the transcriptional regulation affects the network, while there are many other types of regulation that also affect the network composition and dynamics, like the post-translational modifications. Especially phosphorylation is relevant in modulating the protein activity and protein-protein network interactions, hence the network composition[45].

LRR-RKs have previously been classified according the number of LRR in their extracellular domain as either small or large. Despite most of the small LRR-RKs function as co-receptors and the large LRR-RKs as ligand receptors, there are cases in which this rule does not hold[15]. Results in this work showed that small and large proteins in the network have statistical differences in key centrality parameters like degree, betweenness, clustering, eigenvector centrality and load centrality. Furthermore, their elimination from the network cause differential effects in terms of connectivity. This strongly suggests that they play different functional roles inside the network. Whether this differential role is related somehow to their function as ligand receptors or co-receptors would yet to be clarified. However, it would make sense that proteins that act as co-receptors have a wider number of interactions and are more flexible than proteins than bind a single ligand and are related to a single signaling pathway.

Networks in biology often include only one type of nodes: genes, proteins or metabolites. However, the integration of different kind of networks into only one could provide more information about both separately and together. In this work information provided by the software TF2Network was integrated with the PPI data from CSI^{LRR}. This allowed to classify some of the network nodes into being regulated by immunity-related, development-related transcription factors, or both. In Arabidopsis and other plant species it has been well characterized the negative feedback that exists between growth and defense. The immune response in plants, and in all organisms in general, are very expensive in terms of energy and resources. Because of that, when a plant is being infected by a pathogen, it stops growing and developing. In contrast, when the plant is inverting energy and resources in development, the immune response gets weakened[42]. However, the

molecular mechanism of this decision has not been completely elucidated. Using this kind of approach potential receptors involved in the immunity/development crosstalk could be identified. This would provide a deeper understanding about the molecular mechanism responsible for this biological phenomenon.

In summary, the results presented in this work are another example of the applications of a systems biology integrative approach to biological data. It has been determined that despite fitting a power-law degree distribution, the CSI^{LRR} network does not strictly fit a scale-free network model. However, this network displays high tolerance to random attacks and reduced tolerance to hub/bottleneck-directed attacks, similarly to random scale-free network models. After studying its basic characteristics, the integration of interaction data with RNA-Seq data revealed that the transcriptional regulation of the network could be more relevant for developmental programs than for defense responses. Additionally, it has been found that LRR-RKs ECDs with a small size have a major role in the maintenance of the CSI^{LRR} network integrity. Finally, it was hypothesized that the integration of CSI^{LRR} data with predicted GRN could shed light upon the functioning of growth-immunity signaling crosstalk.

5 Bibliography

1. Yu D, Kim M, Xiao G, Hwang TH. Review of Biological Network Data and Its Applications. *Genomics Inform*. 2013;11(4):200-210. doi:10.5808/gi.2013.11.4.200
2. Barabási AL, Oltvai ZN. Network biology: Understanding the cell's functional organization. *Nat Rev Genet*. 2004;5(2):101-113. doi:10.1038/nrg1272
3. Albert R. Scale-free networks in cell biology. *J Cell Sci*. 2005;118(21):4947-4957. doi:10.1242/jcs.02714
4. Ahmed H, Howton TC, Sun Y, Weinberger N, Belkhadir Y, Mukhtar MS. Network biology discovers pathogen contact points in host protein-protein interactomes. *Nat Commun*. 2018;9(1):1-13. doi:10.1038/s41467-018-04632-8
5. McCormack ME, Lopez JA, Crocker TH, Mukhtar MS. Making the right connections: Network biology and plant immune system dynamics. *Curr Plant Biol*. 2016;5:2-12. doi:10.1016/j.cpb.2015.10.002
6. Broido AD, Clauset A. Scale-free networks are rare. *Nat Commun*. 2019;10(1):1-10. doi:10.1038/s41467-019-08746-5
7. Yin Y, Wu D, Chory J. Plant receptor kinases: Systemin receptor identified. *Proc Natl Acad Sci U S A*. 2002;99(14):9090-9092. doi:10.1073/pnas.152330799
8. Dufayard JF, Bettembourg M, Fischer I, *et al.* New insights on Leucine-Rich repeats receptor-like kinase orthologous relationships in angiosperms. *Front Plant Sci*. 2017;8(381):1-18. doi:10.3389/fpls.2017.00381
9. Zulawski M, Schulze G, Braginets R, Hartmann S, Schulze WX. The Arabidopsis Kinome: Phylogeny and evolutionary insights into functional diversification. *BMC Genomics*. 2014;15(548):1-14. doi:10.1186/1471-2164-15-548
10. Chakraborty S, Nguyen B, Danyal Wasti S, Xu G. Plant Leucine-Rich Repeat Receptor Kinase (LRR-RK): Structure, Ligand Perception, and Activation Mechanism. *Molecules*. 2019;24(3081):1-37. doi:10.3390/molecules24173081
11. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. The protein kinase complement of the human genome. *Science*. 2002;298(5600):1912-1934. doi:10.1126/science.1075762
12. Shiu SH, Bleecker AB. Receptor-like kinases from Arabidopsis form a monophyletic gene family related to animal receptor kinases. *Proc Natl Acad Sci U S A*. 2001;98(19):10763-10768. doi:10.1073/pnas.181141598
13. Gou X, He K, Yang H, *et al.* Genome-wide cloning and sequence analysis of leucine-rich repeat receptor-like protein kinase genes in Arabidopsis thaliana. *BMC Genomics*. 2010;11(19):1-15. doi:10.1186/1471-2164-11-19
14. Alberts B, Johnson A, Lewis J, *et al.* Chapter 3. Proteins. In: *Molecular Biology of the Cell*. Sixth. ; 2015:819.
15. Xi L, Wu XN, Gilbert M, Schulze WX. Classification and interactions of LRR receptors and co-receptors within the arabidopsis plasma membrane – An overview. *Front Plant Sci*. 2019;10(472):1-8. doi:10.3389/fpls.2019.00472
16. Torii KU. Leucine-Rich Repeat Receptor Kinases in Plants: Structure, Function, and Signal Transduction Pathways. *Int Rev Cytol*. 2004;234:1-46. doi:10.1016/S0074-7696(04)34001-5
17. Smakowska-Luzan E, Mott GA, Parys K, *et al.* An extracellular network of Arabidopsis leucine-rich repeat receptor kinases. *Nature*. 2018;553(7688):342-346. doi:10.1038/nature25184
18. Chu Y, Corey DR. RNA Sequencing: Platform Selection, Experimental Design, and

Data Interpretation. *Nucleic Acid Ther*. 2012;22(4):10-13. doi:10.1089/nat.2012.0367
19. Shannon P, Markiel A, Ozier O, *et al.* Cytoscape: A Software Environment for Integrated Models. *Genome Res*. 1971;13(22):2498–2504. doi:10.1101/gr.1239303.metabolite
20. Yang L, Biswas S, Finkel OM, *et al.* Pseudomonas syringae type III effector HopBB1 promotes host transcriptional repressor degradation to regulate phytohormone responses and virulence. 2018;21(2):156-168. doi:10.1016/j.chom.2017.01.003.Pseudomonas
21. Li B, Jiang S, Yu X, *et al.* Phosphorylation of trihelix transcriptional repressor ASR3 by MAP KINASE4 negatively regulates arabidopsis immunity. *Plant Cell*. 2015;27(3):839-856. doi:10.1105/tpc.114.134809
22. Kawakatsu T, Stuart T, Valdes M, *et al.* Unique cell-type-specific patterns of DNA methylation in the root meristem. *Nat Plants*. 2016;2(5):1-8. doi:10.1038/NPLANTS.2016.58
23. Liu J, Jung C, Xu J, *et al.* Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in arabidopsisC W. *Plant Cell*. 2012;24(11):4333-4345. doi:10.1105/tpc.112.102855
24. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*. 2008;5(7):621-628. doi:10.1038/nmeth.1226
25. Hagberg AA, Schult DA, Swart PJ. Exploring network structure, dynamics, and function using NetworkX. In: *7th Python in Science Conference (SciPy 2008)*. ; 2008:11-16.
26. Hotelling H. Analysis of a complex of statistical variables into Principal Components. *J Educ Psychol*. 1933;24:417-441.
27. Mann HB, Whitney DR. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Ann Math Stat*. 1947;18(1):50-60. doi:10.1214/aoms/1177730491
28. Welch BL. The generalisation of student's problems when several different population variances are involved. *Biometrika*. 1947;34(1-2):28-35. doi:10.1093/biomet/34.1-2.28
29. Massey FJ. The Kolmogorov-Smirnov Test for Goodness of Fit. *J Am Stat Assoc*. 1951;46(253):68-78.
30. Virtanen P, Gommers R, Oliphant TE, *et al.* SciPy 1.0--Fundamental Algorithms for Scientific Computing in Python. *Prepr arXiv190710121*. 2019:1-22. http://arxiv.org/abs/1907.10121.
31. Alstott J, Bullmore E, Plenz D. Powerlaw: A python package for analysis of heavy-tailed distributions. *PLoS One*. 2014;9(1):e85777. doi:10.1371/journal.pone.0085777
32. Maeso CA. Estudio de propiedades estáticas y dinámicas para modelos aplicados a redes. 2002.
33. Brennecke P, Anders S, Kim JK, *et al.* Accounting for technical noise in single-cell RNA-seq experiments. *Nat Methods*. 2013;10(11):1093-1098. doi:10.1038/nmeth.2645
34. Mokry M, Hatzis P, Schuijers J, *et al.* Integrated genome-wide analysis of transcription factor occupancy, RNA polymerase II binding and steady-state RNA levels identify differentially regulated functional gene classes. *Nucleic Acids Res*. 2012;40(1):148-158. doi:10.1093/nar/gkr720
35. Varoquaux G, Buitinck L, Louppe G, Grisel O, Pedregosa F, Mueller A. Scikit-

learn: Machine Learning in Python. *J Mach Learn Res*. 2011;12(1):2825-2830. doi:10.1145/2786984.2786995

36. Kulkarni SR, Vaneechoutte D, Van de Velde J, Vandepoele K. TF2Network: predicting transcription factor regulators and gene regulatory networks in Arabidopsis using publicly available binding site information. *Nucleic Acids Res*. 2018;46(6):e31. doi:10.1093/nar/gkx1279
37. Dodds PN, Rathjen JP. Plant immunity: Towards an integrated view of plantĝ€" pathogen interactions. *Nat Rev Genet*. 2010;11(8):539-548. doi:10.1038/nrg2812
38. Kuya N, Sato S. The relationship between profiles of plagiogravitropism and morphometry of columella cells during the development of lateral roots of Vigna angularis. *Adv Sp Res*. 2011;47(3):553-562. doi:10.1016/j.asr.2010.09.009
39. Phukan UJ, Jeena GS, Shukla RK. WRKY transcription factors: Molecular regulation and stress responses in plants. *Front Plant Sci*. 2016;7(760):1-14. doi:10.3389/fpls.2016.00760
40. Xie M, Sun J, Gong D, Kong Y. The roles of arabidopsis C1-2i subclass of C2H2-type zinc-finger transcription factors. *Genes*. 2019;10(9). doi:10.3390/genes10090653
41. Li J, Wang Y, Zhang Y, Wang W, Irish VF, Huang T. RABBIT EARS regulates the transcription of TCP4 during petal development in Arabidopsis. *J Exp Bot*. 2016;67(22):6473-6480. doi:10.1093/jxb/erw419
42. Belkhadir Y, Yang L, Hetzel J, Dangl JL, Chory J. The growth-defense pivot: Crisis management in plants mediated by LRR-RK surface receptors. *Trends Biochem Sci*. 2014;39(10):447-456. doi:10.1016/j.tibs.2014.06.006
43. Crua Asensio N, Munõz Giner E, De Groot NS, Torrent Burgas M. Centrality in the host-pathogen interactome is associated with pathogen fitness during infection. *Nat Commun*. 2017;8:1-6. doi:10.1038/ncomms14092
44. Caldu-Primo JL, Alvarez-Buylla ER, Davila-Velderrain J. Structural robustness of mammalian transcription factor networks reveals plasticity across development. *Sci Rep*. 2018;8(1):1-15. doi:10.1038/s41598-018-32020-1
45. Buchanan BB, Gruissem W, Jones RL. Chapter 18. Signal Transduction. In: *Biochemistry and Molecular Biology of Plants*. Second. Wiley Blackwell; 2015:838.

Appendix 1: parameters table

Table A 1. Network parameters computed in this work to compare the networks.

Parameter	Computation	Meaning
Degree centrality	Mean of all values returned by networkx.degree_centrality	For a node number of neighbors or connections. Therefore, is the average value for all the nodes in the network.
Closeness centrality	Mean of all values returned by networkx.closeness_centrality	For a node, measure of centrality in a network, calculated as the reciprocal of the sum of the length of the shortest paths between the node and all other nodes in the network. Therefore, is the average value for all the nodes in the network.
Betweenness centrality	Mean of all values returned by networkx.betweenness_centrality	Proportion of the shortest paths that pass through a node. Therefore, is the average value for all the nodes in the network.
Average clustering	Mean of all values returned by networkx.clustering	For a node, proportion of connections its neighbors have among them in comparison to all the possible ones. How close its neighbors are to be a clique (complete graph). Average of all the node sin the graph.
Average eigenvector centrality	Mean of all values returned by networkx.eigenvectorcentrality	Measure of the influence of a node in a network.
Clique number	networkx.graph_clique_number	Number of subgraphs that are complete (all its nodes are connected with each other)
Average shortest path length	networkx.average_shortest_path_length	Average number of steps along the shortest paths for all possible pairs of network nodes.
Maximum k-core	Maximum value of networkx.core_number after removing self-loop edges from the network.	Maximal connected subgraph
Node with highest degree	Maximum value returned by networkx.Graph.degree	Degree of the node with the largest number of neighbors.
Node with highest betweenness	Maximum value returned by networkx.betweenness_centrality	Betweenness of the node that is in the largest proportion of shortest paths.
Network order	Number of element in object returned by networkx.Graph.nodes	Number of nodes in the network.
Network size	Number of element in object returned by networkx.Graph.edges	Number of edges (connections) in the network.
Betweenness-degree R^2	scipy.Stats.linregress(degree, betweenness).rvalue	Reflects if the nodes that tend to be hubs also tend to be bottlenecks.
Degree distribution R^2	scipy.Stats.linregress(degree distribution).rvalue	Reflects information about the topology of the network
Average eccentricity	Mean of all values returned by networkx.eccentrivity	It is defined as the maximum distance of one vertex from another vertex. In this case the average of all the nodes.
Network radius	networkx.radius	The minimum eccentricity in a network
Network diameter	networkx.diameter	The maximum eccentricity in a network
Network density	networkx.density	The density is 0 for a graph without edges and 1 for a complete graph

Appendix 2: attacks functions code

```
def random_attack(G, m):

    '''''
    This function takes as input a NetworkX graph object and an integer m. Return
s a graph
    that has been attacked in m number of nodes randomly.
    '''

    assert len(G.nodes) >= m, "m cannot be higher than the number of nodes in the
 graph"

    j = 1
    while j <= m:
        node = sample(list(G.nodes), 1)
        G.remove_node(node[0])
        j += 1

    return G
```

```
def degree_attack(G, m):

    '''''
    This function takes as input a NetworkX graph object and an integer m. Return
s a graph
    that has been attacked in m number of nodes based on degree
    '''

    assert len(G.nodes) >= m, "m cannot be higher than the number of nodes in the
 graph"

    j = 1
    while j <= m:
        node = sorted(G.degree, key = lambda x: x[1], reverse = True)[0][0]
        G.remove_node(node)
        j += 1

    return G
```

```
def minpath_attack(G, m):

    '''''
    This function takes as input a NetworkX graph object and an integer m. Return
s a graph
    that has been attacked in m number of nodes based on the proportion of shorte
st paths that cross them
    '''

    assert len(G.nodes) >= m, "m cannot be higher than the number of nodes in the
 graph"

    j = 1
    while j <= m:
        load_cent = nx.load_centrality(G)
        node = max(load_cent, key = load_cent.get)
        G.remove_node(node)
        j += 1

    return G
```

```
def random_attack_size(G, m, att):

    '''''
    This function takes as input a NetworkX graph object with an attribute called

    'size' which values are either 'small' or 'large'. Also takes an integer m an
d
    the value of the attribute you want to filter. The function returns a graph
    that has been attacked in an m number of small or large nodes randomly
    '''

    nodes = [x for x,y in G.nodes(data=True) if y['size'] == att]

    assert len(nodes) >= m, "m cannot be higher than the number of nodes with the
 given attribute"

    j = 1
    while j <= m:
        node = sample(nodes, 1)
        G.remove_node(node[0])
        nodes.remove(node[0])
        j += 1

    return G
```

```
def degree_attack_size(G, m, att):

    '''''
    This function takes as input a NetworkX graph object with an attribute called

    'size' which values are either 'small' or 'large'. Also takes an integer m an
d
    the value of the attribute you want to filter. The function returns a graph
    that has been attacked in an m number of small or large nodes based on their
degree
    '''

    nodes = [x for x,y in G.nodes(data=True) if y['size'] == att]

    assert len(nodes) >= m, "m cannot be higher than the number of nodes with the
 given attribute"

    j = 1
    while j <= m:
        node = sorted(G.degree(nodes), key = lambda x: x[1], reverse = True)[0][0
]
        G.remove_node(node)
        j += 1

    return G
```

```
def minpath_attack_size(G, m, att):

    '''''
    This function takes as input a NetworkX graph object with an attribute called

    'size' which values are either 'small' or 'large'. Also takes an integer m an
d
    the value of the attribute you want to filter. The function returns a graph
    that has been attacked in an m number of small or large nodes based
    on the proportion of shortest paths that cross them
    '''
    nodes = [x for x,y in G.nodes(data=True) if y['size'] == att]

    assert len(nodes) >= m, "m cannot be higher than the number of nodes in the g
raph"

    j = 1
    while j <= m:

        load_cent = nx.load_centrality(G)

        new = {k: load_cent[k] for k in nodes}

        node = max(new, key = new.get)
        G.remove_node(node)
        nodes.remove(node)
        j += 1

    return G
```

Appendix 3: cut-off value computation code

```
def cutoff_comp(df):

    '''''
    This function takes as input argument a data frame with the count matrix
    from a RNA-Seq experiment, that contains the expression values in FPKM.

    It returns a cut-off expression value between technical noise and
    expressed genes.
    '''

    df_log = np.log1p(df)

    L = []

    for i in range(len(df_log.columns)):

        data = np.array(df_log.iloc[:, i]).reshape(-1, 1)
        gmm = GMM(n_components = 2, covariance_type = 'tied', random_state = 0).f
it(data)
        labels_gmm = gmm.predict(data)
        norm1 = data[labels_gmm == 0]
        norm2 = data[labels_gmm == 1]
        L.append(np.mean([max(norm1)[0], min(norm2)[0]]))

    return np.expm1(np.mean(L))
```

Appendix 4: heatmaps

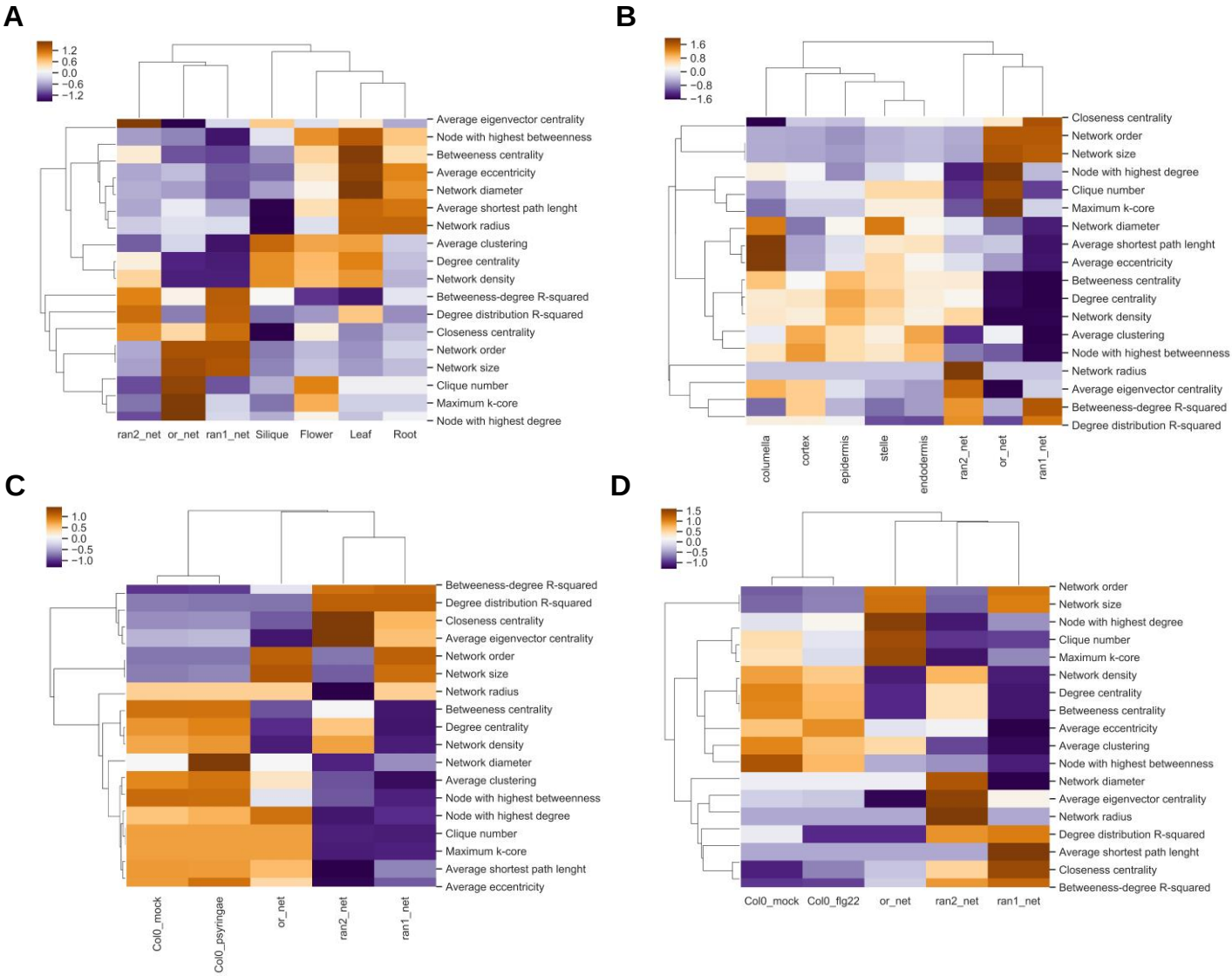

Figure 1A. Heatmaps of the parameter values for the networks used in this work. The 18 network parameters computed for each network in the work are represented with a heatmap. The figures were generated using the hierarchical clustering method of the function seaborn.clustermap. **(A)** Heatmap for the organs dataset. **(B)** Heatmap for the root dataset. **(C)** Heatmap for the *P. syringae* dataset. **(D)** Heatmap for the flg22 dataset. The colors represent the z-score of the parameters along the rows. Orange color indicates that the parameters value is larger, while purple indicates that the value is lower. To the heatmap were additionally added two random networks. ran1_net corresponds to the averaged values of all the parameters of 100 random networks with the size and order of CSI^{LRR}. ran2_net is the same, but the order and size are the mean of the generated networks from the Python script in chapter Results 3.3.

Appendix 5: expressed genes table from organs data set

Table A 2. Table with the percentage of expressed genes in each organ from the organs dataset and the order of the graph of each organ.

Tissue	Not expressed	Expressed genes	% of expressed genes	Graph order
Leave	21138	13150	38.35%	40
Root	19746	14542	42.41%	65
Flower	18904	15384	44.86%	55
Silique	20219	14069	41.03%	33