Reverse engineering biological networks: computational approaches for modeling biological systems from perturbation data

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

YUNGIL KIM

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Chad L. Myers, Ph.D.

September 2013

Acknowledgements

I owe the greatest debt of gratitude to my advisor, Chad L. Myers, for inspiring me to study computational biology, guiding my development as a scholar, and advocating tirelessly for me. Fumiaki Katagiri motivated and guided the second half of the dissertation; I am particularly grateful for his attitude of questioning assumptions and for providing balance in my intellectual development. Thanks go to my other current and former committee members for giving me chances to discuss the topics in different perspectives. I am grateful to Kenichi Tsuda for his kind comments and suggestions in biology perspectives. I am also indebted to my co-workers, Jeremy Bellay and Rachel A. Hillmer. They have encouraged me to expand my viewpoints through valuable discussions. I was very fortunate to have kind group members; the comments and the technical helps from Benjamin VanderSluis, Raamesh Deshpande, and Roman Briskine were particularly helpful. I also appreciate Gaurav Pandey and Gang Fang for their valuable discussions and encouragements. Taehyun Hwang and Youn Joo Choi, to whom I deeply appreciate, always have been a continuous source of support and encouragement throughout graduate school. I would also express appreciation to all of my friends in both electrical engineering and computer science departments. They have sincerely supported me like family members. I apologize for many thankful persons not listed here. And I would like to express my deepest gratitude to my beloved girlfriend, Hwanhee Hong, who has been virtually working as hard as me on this thesis. I couldn't imagine how would I finish this thesis if it were not for her constant love and faith in me through all those hard times together, taking care of all my mental and physical needs.

Dedication

To my parents, Bong Hwan Kim and Namsoon Kwon, who instilled an enduring love of learning.

Abstract

A fundamental goal of systems biology is to construct molecule level models that explain and predict cellular or organism level properties. A popular approach to this problem, enabled by recent developments in genomic technologies, is to make precise perturbations of an organism's genome, take measurements of some phenotype of interest, and use these data to "reverse engineer" a model of the underlying network. Even with increasingly massive datasets produced by such approaches, this task is challenging because of the complexity of biological systems, our limited knowledge of them, and the fact that the collected data are often noisy and biased.

In this thesis, we developed computational approaches for making inferences about biological systems from perturbation data in two different settings: (1) in yeast where a genome-wide approach was taken to make second-order perturbations across millions of mutants, covering most of the genome, but with measurement of only a gross cellular phenotype (cell fitness), and (2) in a model plant system where a focused approach was used to generate up to fourth-order perturbations over a small number of genes and more detailed phenotypic and dynamic state measurements were collected. These two settings demand different computational strategies, but we demonstrate that in both cases, we were able to gain specific, mechanistic insights about the biological systems through modeling.

More specifically, in the yeast setting, we developed statistical approaches for integrating data from double perturbation experiments with data capturing physical interactions between proteins. This method revealed the highly organized, modular structure of the yeast genome, and uncovered surprising patterns of genetic suppression, which challenge the existing dogma in the genetic interaction community. In the model plant setting, we developed both a Bayesian network approach and a regularized regression strategy for integrating perturbations, dynamic gene expression levels, and measurements of plant immunity against bacterial pathogens after genetic perturbation. The models resulting from both methods successfully predicted dynamic gene expression and immune response to perturbations and captured similar biological mechanisms and network properties. The models also highlighted specific network motifs responsible for the emergent properties of robustness and tunability of the plant immune system, which are the basis for plants' ability to withstand attacks from diverse and fast-evolving pathogens. More broadly, our studies provide several guidelines regarding both experimental design and computational approaches necessary for inferring models of complex systems from combinatorial mutant analysis.

Table of Contents

List of Tables

List of Figures

1 Introduction

1.1 Background: Analyses of biological data with perturbations in systems biology

After the completion of the *H. influenza* genome sequence in 1995 [1], there have been great advances in high-throughput experimental technologies. These innovations in experimental methods have enabled the production of enormous genomic datasets and provided systems-level measurements for virtually all types of cellular components in model organisms. In particular, experimental perturbation, the observation of phenotypic variations resulting from specific genetic or environmental disruptions, has provided a wealth of powerful data for unraveling the complex interplay of genes and environment.

A fundamental goal of systems biology is to reverse engineer the molecular architectures of cellular organisms by constructing models to explain, predict, and manipulate their behavior. Even in information-rich environments with increasingly massive datasets, this task is still challenging because our knowledge of biological systems is incomplete and these data are often noisy and biased. To tackle these problems, we developed computational approaches for making inferences about biological systems from perturbation data in both yeast and plant. In particular, we explore statistical approaches for extracting and interpreting biological insights from genome-wide genetic interaction data for yeast protein complexes. We also develop modeling approaches for modeling a complex plant immune signaling network with the combination of two types of high-order perturbation data. The rest of this chapter is organized as follows.

First, we briefly give an overview of experimental perturbation approaches in systems biology and introduce various types of biological data available for systematic analyses. Second, we provide more specific biological backgrounds related to the chapters that follow. Third, we discuss several challenges in developing computational approaches regarding the breadth and order of experimental perturbation data. Finally, we describe the organization of the dissertation, including a brief summary of each chapter.

1.1.1 Experimental perturbations in systems biology

To understand the functional organization of cellular components as a whole system, there has been a concerted effort in systems biology to analyze comprehensive functional genomics datasets. The most widely used strategies for studying complex biological systems are experimental perturbations. These approaches focus on the quantitative measurements of the phenotypic changes of cellular components after some perturbation has been applied and analyze them in a systematic way. Based on the type of intervention, there are two general approaches for perturbing systems: external (conditional) perturbation and internal (genetic) perturbation. Using an external perturbation approach, one measures the differential sensitivities of cellular components in biological systems or specific pathways perturbed by environmental triggers. Using a genetic perturbation approach, on the other hand, one disables specific components of the cell and observes the resulting phenotype. The perturbation-based approaches can provide causal biological mechanisms as well as systems-level insights.

Recent attention has focused on combinatorial perturbation approaches, which is one strategy for elucidating higher-order complexity, an inherent property of biological systems. Based on early efforts to construct comprehensive mutant libraries, it was found a large fraction of genes in model organisms have little or no observable effect when deleted independently [2]. Thus, combinatorial genetic perturbations are prevalent for the analyses of network robustness [3] and studies of hidden biological mechanisms [4]. Moreover, the appropriate mixture of conditional and genetic perturbation provides useful information on the analysis of induced systems or specific pathways. Along with external stimuli, combinatorial genetic perturbation approaches enable the functional annotation of uncharacterized genes or chemical compounds and identify unknown relations between cellular responses and corresponding pathways [5]. Similarly, in population genomics, several researchers have studied systems that have been perturbed through genetic crosses that leverage to use natural genetic variations to study combinatorial effects [6],[7]. With the cost-effective and multi-factorial perturbations, they successfully identified potential drug targets and provided an integrated perspective on complex disease mechanisms [8].

1.1.2 High-throughput experimental data in systems biology analyses

1.1.2.1 Genomics

Genomics is a discipline in genetics that applies recombinant DNA, DNA sequencing methods, and bioinformatics to sequence, assemble, and analyze the function and structure of genomes. The advent of next generation sequencing (NGS) technologies enabled the discovery of fast, inexpensive, and accurate genome information, useful for many applications [9], [10]. For example, whole genome sequencing of known and uncharacterized genomes is highly capable of detecting single nucleotide polymorphisms (SNPs), and structural variations such as copy number variations (CNVs) [11], [12], [13]. These sequencing technologies are also widely applied in population studies [14], [15], [16], [17], [18], which allow more accurate imputation of variants in genome-wide association studies (GWAS) and better localization of disease-associated variants. Similarly, exome sequencing is being applied to many different types of medical applications including diagnosis and disease monitoring [19], [20].

1.1.2.2 Transcriptomics

Transcriptomic technologies provide information about the relative abundance of RNA transcripts, indicating the active components within a cell. In the late 1990s, microarrays and serial analysis of gene expression (SAGE) were frequently employed for quantifying the dynamics of gene expressions in many model organisms [21]. Recently, next generation sequencing (NGS) technology offers alternative solutions for these applications. RNA sequencing (RNA-seq), which refers to the use of NGS technology for the deep sequencing of mRNA has been used for the identification of novel transcripts [22], [23], [24], the detection of alternative splicing [25], [26], [27] and quantification of gene expression and analysis of differential expression [19], [23], [28], [29], [30].

1.1.2.3 Proteomics

Proteomic technologies enable the identification and quantification of the cellular levels of each protein encoded by the genome. The development of mass spectrometry (MS)-based approaches has enabled the high-throughput analysis of protein levels as well as post-transcriptional modifications with increasing sensitivity [31]. The MS-based approach has also been applied to the identification and quantitation of binding sites relating to post-translational modifications, including studies of phosphorylation as a function of stimulus, time, and subcellular location [32]. The systematic analysis of protein localization in a cell has also been accomplished by tagging green fluorescent protein and fluorescence resonance energy transfer [33].

1.1.2.4 Interactomics

Interactomic technologies focus on mapping different types of interactions between genes, and so far have focused on detecting two main types: physical and genetic interactions. Physical interactions represent the direct associations between two cellular components and can included both protein-DNA and protein-protein interactions. The genetic regulatory network controlling a cell is composed of protein-DNA interactions between transcription factors and their target promoters. Identifying this network structure and its properties is crucial to understanding how cells respond to intracellular, extracellular, and environmental signals. The combination of chromatin immunoprecipitation with whole-genome promoter or tiling arrays (ChIP-chip) have been applied for elucidating gene regulatory interactions [34]. Recently, chromatin immunoprecipitation with massively parallel DNA sequencing (ChIP-Seq) was developed and is now widely used for mapping gene regulatory networks and epigenetic mechanisms [19], [20], [29], [30].

Comprehensive protein-protein interaction networks provide a complementary perspective on the functional organization of a genome [35]. For instance, protein-protein interactions are the means of signal transduction by which a cell receives information about its external environment. In many organisms including *S. cerevisiae* [36], [37], *D. melanogaster* [38], *C. elegans* [39], and humans [40], these genome-scale interaction networks have been identified by yeast two-hybrid (Y2H) and co-affinity purification coupled with mass-spectrometry techniques.

Another highly informative type of interactions is genetic interaction. A genetic interaction is said to occur between two genes whose simultaneous mutation results in a phenotype different from what is expected given the phenotypes of their individual mutations [41]. Genetic interactions generally indicate co-involvement in the same complex, pathway, or parallel pathways leading to the same essential function, and thus are highly informative of gene function [41]. Systematic mapping of genetic interactions in yeast has been accomplished by synthetic genetic array (SGA) analysis [42] and diploid-based synthetic lethal analysis by microarray (dSLAM) analyses [43]. Similar experimental approaches based on RNA interference have also been developed for multi-cellular organisms [44], [45].

1.2 Background: genetic interaction maps in yeast and plant immune system

1.2.1 Genetic interaction maps in yeast

Systematic gene deletion analysis in the budding yeast, *Saccharomyces cerevisiae*, has demonstrated that only ~ 20 % of genes are required for viability in standard laboratory conditions [2], possibly reflecting the robustness of biological circuits and the extensive buffering of eukaryotic genomes against genetic and environmental perturbations [46]. Functional relationships between genes and their associated pathways have been revealed by genetic interactions in which double mutations lead to fitness phenotypes that significantly deviate from their expected multiplicative effect of combining two single mutants [41]. Based on the direction of the deviation, genetic interactions can be categorized into two distinct groups (positive or negative genetic interactions), which provide different information on the functional relationships between gene pairs [47], [48].

Over the last ten years, the systematic mapping of genetic interactions in *S. cerevisiae* have been have been conducted. Tong *et al.* used the SGA approach to screen 132 query strains associated with diverse biological processes against a complete library of ~ 4700 non-essential deletion strains and identified ~ 4200 genetic interactions involving ~1000 genes [49]. Costanzo *et al.* further screened ~ 1700 query strains against same non-essential arrays to report ~ 170,000 quantitative genetic interactions on a genome-scale [47]. Pan *et al.* used the dSLAM approach to screen 74 query strains involved in DNA replication and repair against the same deletion library and identified

~4900 synthetic sick/lethal interactions for 875 genes [43]. Technical alternatives such as eSGA [50], [51] and RNAi [44], [45] have been developed for large-scale genetic interaction maps in other organisms.

Similar genetic interaction profiles of genes belonging to the same pathway or biological process are used for constructing a global network which highlights their inherent functional organization [47]. Due to the complementary nature of genetic and physical interaction networks [47], [49], the integration of genetic and physical interaction maps provides a powerful model for mapping the functional characteristics of cellular components in a systematic way [48], [52], [53], [54]. Combined with chemical genomics, the study of using chemical probes on a genome-wide scale, genetic interaction networks also assist to identify the mode of action of drugs or uncharacterized compounds [47], [55]. The increasingly massive quantitative genetic interaction datasets for diverse organisms, from bacteria [50], [51] to mammalian cells [45], enable the study of general principles and structures of genetic networks, which play a key role in governing inherited phenotypes, including human disease [56].

1.2.2 Plant immune system

In nature, plants are continuously threatened by a wide range of harmful pathogens. Depending on their modes of attack, plant pathogens are categorized into biotrophs, nectrotrophs, and hemi-biotrophs [57]. Biotrophs feed on nutrients from living host tissues while necrotrophs feed on dead or dying cells. Hemi-biotrophs display different lifestyles throughout their life cycle. To defend themselves against these different types of pathogens, plants recognize pathogen attack and transduce this information through signaling networks within the cell, to different cells, and to distant tissues. Plants have evolved sophisticated strategies to facilitate innate immunity for efficient immune responses to such diverse pathogenic attackers [58].

The current model of the plant immune system consists of two major modes of pathogen recognition, called pattern triggered immunity (PTI) and effector triggered immunity (ETI) [58], [59]. In PTI, pattern recognition receptors (PRRs) of a plant recognize microbe associated molecular patterns (MAMPs) relatively conserved among similar types of microbes [60], [61]. The MAMP recognition triggers diverse downstream signaling events such as a MAP kinase cascade [62], ultimately activating a basal resistance [58]. During the coevolution between pathogens and their host plants, pathogens acquire effector molecules transported into the host cell, suppress PTI, and increase their virulence [63]. This results in effector-triggered susceptibility (ETS) [58]. Accordingly, plants initiate an alternative mode of pathogen recognition, called ETI [58]. In ETI, plant Resistance (R) gene products, generally located in the cytoplasm, recognize specific pathogen effectors directly, or indirectly by recognizing host proteins that have been damaged by effectors [64]. The recognition of an effector by an R protein leads to activation of a suite of signaling and defense responses [65]. In most cases, these include an oxidative burst [66], [67], activation of a programmed cell death response called the hypersensitive response (HR) [68], increased levels of the signaling molecule salicylic acid (SA), and a suite of gene expression changes [69]. Evidence of the effectiveness of these responses is provided by observations that mutants with defects in SA-dependent signaling allow increased growth of pathogens recognized by R proteins.

1.3 Dissertation focus

As explained in Section 1.1, high-throughput experimental technologies are generating increasingly massive biological datasets, which allow for the systematic investigation of biological phenomena in a relatively unbiased way. Given the enormous amount of large-scale data, our general goal is to address two biological questions:

1. What are the general principles governing biological systems?
2. How are genetic linkages between gene products or molecular machineries translated into phenotypic variations in response to diverse input signals?

In systems biology, combinatorial experimental perturbations are extensively used in these problems with two broad approaches. One approach is to systematically collect large-scale experimental data covering an entire genome and reveal emergent properties of cellular networks [47], [48]. For example, Costanzo *et al.* collected genome-wide unbiased genetic interactions from quantitative synthetic genetic array (SGA) analysis and constructed a functional map of the cell, organizing genes and higher-order bioprocesses according to their related roles [47] (Figure 1-1(a)). A complementary approach to the genome-scale systems biology studies has been instead to focus on the behavior of specific pathways [4], [70]. These studies typically generate quantitative datasets for the pathway and target components of interest and apply modeling approaches to understand specific biological mechanisms rather than broad emergent properties. As an example, Karen *et al.* used multi-dimensional flow cytometry data and applied Bayesian network models to infer a high-accuracy human primary T cell signaling causality map [70] (Figure 1-1(b)).

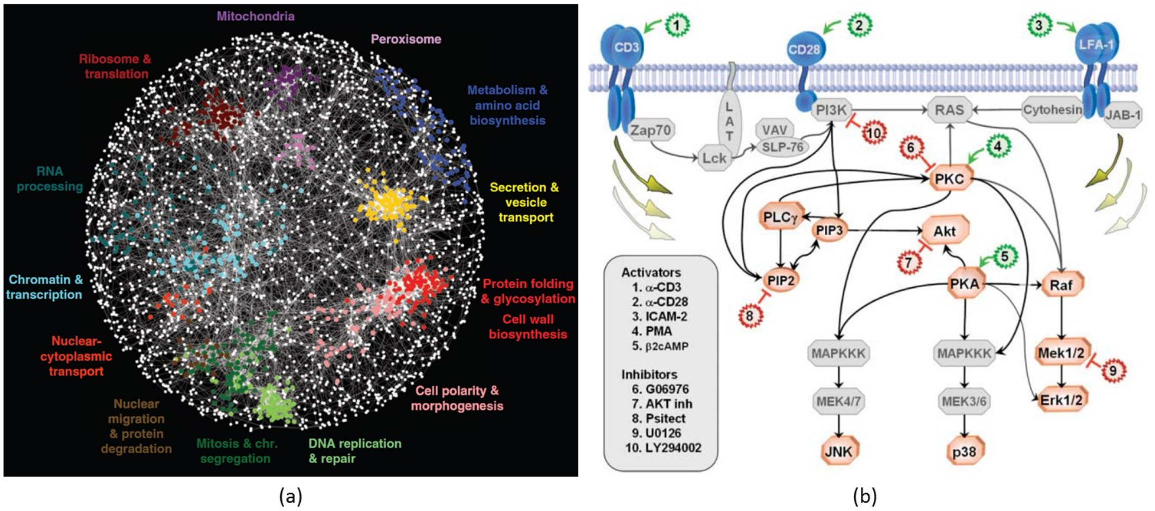


Figure 1-1 Examples of genome-wide and pathway-specific approaches with combinatorial experimental perturbations in systems biology. (a) A comprehensive genome-wide network analysis. A correlation-based network connecting genes with similar genetic interaction profiles in yeast [47]. Colored regions indicate sets of genes enriched for biological processes summarized by the indicated terms. (b) A candidate pathway analysis. Human immune signaling network and points of intervention for modeling the causal protein-signaling networks [70]. The interventions classified as activators are colored green and inhibitors are colored red.

These two strategies have been used successfully to turn repositories of experimental data into the new biological insights. However, both approaches face their own challenges. The comprehensive analysis of large-scale experimental data requires much attention to the collection of noisy datasets generated from heterogeneous settings. In the small-scale analyses for a complex system, the fact that specific components or pathways are selected for perturbation experiments can dramatically affect the accuracy of modeling, and knowledge is still relatively limited in most areas of biology. In this dissertation, we address several issues on these two approaches to systems biology from a computational perspective. In Chapter 2, we use the first genome-scale di-genic interaction network generated from SGA analysis to produce a functional catalogue of all yeast protein complexes. The single experimental framework can relax the difficulty in the integration of heterogeneous datasets from different experiments. Our effort focuses on recapitulation of pathway-based models, previously proposed from biased datasets, and the functional characterization of genetic interactions by protein complex-level analyses. We also consider that the computational approaches be relatively simple enough to be easily applicable to other organisms once large-scale genetic interactions in the organisms are available. In Chapters 3 and 4, we combine both mRNA expressions of four target genes at two time points and immunity levels in all combinatorial mutants of four targets and build dynamic models based on two different modeling frameworks. By comparing the obtained models from two distinct approaches, we address their advantages and disadvantages for modeling a complex biological system. We also discuss how biological assumptions and prior knowledge can be correctly incorporated into the selection of initial model structures and the interpretation of the results. More importantly, we investigate how much combinatorial perturbation experiments contribute to the accurate inference of the causal relationships between target components in a complex system.

1.4 Dissertation organization

This dissertation consists of two major parts based on analytical strategies with experimental perturbation datasets regarding their breadth and order of combinatorial mutants. Chapter 2 describes relatively simple statistical approaches for a comprehensive analysis of genome-wide di-genic interaction networks with yeast physical interaction networks. In contrast to Chapter 2, the next two chapters focus on two computational approaches for modeling a small-scale but complex system by combining two different types of high-order combinatorial perturbation datasets. In chapter 3, we present a modeling approach for a plant immune signaling network based on a Bayesian network. Chapter 4 introduces a multiple regression model for the plant immune system, describes the mechanistic basis for the properties of robustness and tunability of the plant immune signaling network and provides guidelines for modeling a complex system from combinatorial perturbation data. Chapter 5 concludes this dissertation with both a summary of our results and an outlook on possible future work.

2 Functional organization of yeast protein complexes from genome-wide genetic interactions

2.1 Chapter overview

This dissertation starts with an illustration of how quantitative genome-wide genetic interactions can be used to elucidate the systems-level organization of protein complexes in yeast. Mapping genetic interactions in model organisms such as yeast is crucial for investigating how genetic robustness and buffering mechanisms are organized in biological systems. Since genetic interactions and physical interactions contain complementary information, the interpretation of genetic interactions in a physical context is a powerful approach to understanding how molecular components are functionally organized in a cell. However, even in yeast (a single-cell model organism), the total number of combinatorial double mutations is quite large (about 18 million in total) such that current genetic interaction space could be biased towards specific biological processes and functions and is still far from being completely interrogated. Moreover, the large collection of genetic interactions from several different studies needs more attention since these data generally contain surprisingly low overlap for the same gene pairs and different properties in terms of rates of false positives and negatives. Thus, more systematic approaches for the collection of genetic interactions on a genome-scale as well as the statistical analysis of genetic interactions combined with physical interactions are needed.

In this chapter, we re-examine the relationship between quantitative genetic interactions and protein-protein interactions based on a recent genome-wide synthetic genetic array (SGA) analysis in yeast [47], [48] and a comprehensive set of yeast protein complexes [37], [71]. We first find protein complexes enriched for genetic interactions between any pairs of gene members of a complex. For theses complexes, we then measure the relative frequency of positive and negative genetic interactions, defined as a monochromatic purity score (MP-scores). Given the MP-scores, we functionally characterize the protein complexes regarding their essentiality and the presence of a direct physical interaction [36] and notably predict the frequency of interaction with genes in other complexes with the predominant type of genetic interaction within a protein complex. We further examine genetic suppression, a specific class of positive interactions, between protein complexes and build a yeast complex-level genetic suppression network. The large-scale analyses confirm parts of established dogma from small-scale studies and theoretical observations as well as suggest aspects that were not correct. The complex-based analysis with the genome-wide unbiased interaction data captures a more complex nature of the functional signatures in genetic interactions than what a previous pathway model describes. Moreover, the suppression analysis shows that genetic suppression between complexes is surprisingly common. The statistical approaches for interpreting quantitative genetic interactions in a physical context can be applied to functionally coherent modules or other organisms. The work presented in this chapter was published in [47], [48] and includes contributions from Anastasia Baryshnikova, Michael Costanzo, Ji-Young Youn, Bryan-Joseph San Luis, and Chad L Myers. Anastasia and Michael suggested some ideas for analyses. Ji-Young, Bryan-Joseph, and Michael completed experimental validations. Chad supervised the project.

2.2 Background: the relationship between genetic and physical interaction networks

Physical and genetic interaction networks are two major resources important for systematic understanding of the functional organization of a cell [72]. In principle, the physical interaction map with protein-protein and protein-DNA interactions describes the architecture of a cell by capturing direct associations between gene products. Genetic interaction maps, on the other hand, provide functional associations between genes which indicate how the physical architecture produces phenotypes. A previous study confirmed that the overlap between genetic interactions and protein-protein interactions was relatively rare (< 1 %), and suggested that these two types of interactions can be said to be orthogonal [49]. Due to the complementary views from these two types of data regarding cellular structures and functions, obtaining a complete picture of the cell necessitates the integration of both aspects from the two interaction maps in a systematic way. Large-scale experimental measurement of both genetic and physical interaction data have been successfully accomplished over the last decade. Networks of protein-protein interactions have been constructed using yeast two-hybrid (Y2H) [73] or tandem affinity purification coupled with mass spectrometry (TAP-MS) [37],[71]. Similarly, networks of protein-DNA interactions have been built using yeast one-hybrid (Y1H) assays [74] or chromatin immunoprecipication coupled with DNA microchips (ChIP-chip) [86],[87] or sequencing (ChIP-seq) [13],[22]. Large sets of genetic interactions in yeast were measured through the techniques of synthetic genetic arrays (SGA) [33],[37] and diploid-based synthetic-lethality analysis on microarray (dSLAM) [77]. In worms and higher eukaryotes, genetic interactions were measured by RNAi technology-based screening approaches [90],[91].

Many recent studies have sought to interpret the genetic interactions in a physical context. Initially, the functional reorganization of genes in yeast metabolism were elucidated by incorporating the concept of "monochromaticity" into function-enriched modules using the framework of flux balance analysis (FBA) [80]. By using a probabilistic model with protein-protein and protein-DNA interactions to identify both between-module and within-module explanations for genetic relationships, Kelly and Ideker proposed that a between-module model predicts negative genetic interactions between functionally related clusters better than a within-module model [52]. Ulitsky *et al.* further used an alternative approach with relaxed definition of physical modules and interpreted genetic interactions with regard to pathway redundancy and protein essentiality [53]. Bandyopadhyay *et al.* integrated quantitative genetic interactions and TAP-MS data and identified a functional map of 91 complexes involving yeast chromosome organization [81]. St. Onge *et al.* conducted a comprehensive and quantitative analysis of genetic interactions among genes conferring resistance to the DNA-damaging agent methyl-methanesulfonate (MMS). They classified a set of identified positive interactions into subclasses including masking, coequal, and suppression on the basis of the relative MMS sensitivity, and reconstituted DNA repair pathways with the genetic evidence [82]. Battle *et al.* further developed a Bayesian learning method to reconstruct activity pathway networks over large sets of genes based on quantitative genetic interactions [83].

In this work, we explicitly re-examine the relationship between quantitative genetic interactions and physically associated proteins. Based on unbiased genome-wide genetic interactions derived from SGA analysis [47], [48] and a yeast physical interaction network, we first focus on genetic interactions within protein complexes and measure the frequency of both positive and negative genetic interactions. More specifically, with enrichment analyses and monochromatic purity scores for protein complexes, we observe that there exists relationship between the monochromatic purity and the essentiality in highly enriched protein complexes. We also find that the predominant type of genetic interactions observed in a protein complex is predictive of its frequency of interaction with genes in other complexes. Moreover, relative comparison of single and double mutant phenotypes, previously applied in small-scale studies, is adopted to examine genetic suppression relationships between complexes on a genome-scale. Significant numbers of suppression interactions connecting complex pairs are observed and combined into a complex-level suppression network, showing many instances where loss-of-function mutations in one complex rescue growth defects from loss-of-function mutations in another complex.

2.3 A quantitative definition of genetic interaction

The quantitative deviation of the double mutant fitness from the expected double mutant fitness (a multiplication of two single mutant fitnesses [84]) indicates the different types of genetic interactions such as negative and positive genetic interactions [41]. Negative genetic interactions describe double mutants exhibiting a more severe phenotype than expected (red colors in Figure 2-1) [41]. The most extreme case of a negative genetic interaction is synthetic lethality, in which the combination of two mutants, each of which causes little or no phenotypic defect, induces cell death (i in Figure 2-1 and Table 2-1). Another type of negative genetic interactions is synthetic sickness, which shows s greater than expected phenotypic defect, but still results in a viable phenotype (ii in Figure 2-1 and Table 2-1). Positive genetic interactions, on the other hand, describe double mutants exhibiting a less severe phenotype than expected from the multiplicative model (green colors in Figure 2-1) [84]. Given the development of high-resolution quantitative scoring methods with high-throughput screening technology such as SGA [47], [48], positive interactions can be further sub-classified into categories associated with different biological mechanisms including masking, suppression, and coequal [82]. For example, as growth is better than the expected double mutant fitness and is close to the fitness of the sickest single mutant, double mutants exhibit masking interactions (iii in Figure 2-1 and Table 2-1). Conversely, a double mutant with increased fitness close to the healthier single mutant exhibits genetic suppression (iv in Figure 2-1 and Table 2-1). In other words, a mutation of one can suppress the severe phenotypic defect due to mutation of the other. If the non-wildtype phenotypes associated with two single mutants and the resultant double mutant for members of the same nonessential protein complex are quantitatively indistinguishable (symmetric), the genes are said to exhibit a specific type of positive interaction, called coequality [82] (v in Figure 2-1 and Table 2-1). In general, colony-based growth rate has been measured as a fitness in yeast [47], [48]. Other quantitative phenotypes such as stress responses [85] can be used for deriving quantitative genetic interactions in different contexts or organisms [45], [86].

Table 2-1 The Summary of the definition of specific genetic interactions according to fitness values.

	GI types	Epsilon	subclasses	
(i)	Negative interactions	$\varepsilon \ll 0$	Synthetic lethal	$f_{a\Delta b\Delta} \approx 0$
(ii)			Synthetic sick	$f_{a\Delta b\Delta} \ll f_{a\Delta} f_{b\Delta}$
(iii)	Positive interactions	$\varepsilon \gg 0$	Masking	$f_{a\Delta b\Delta} \cong f_{a\Delta}, f_{a\Delta} < f_{b\Delta}$
(iv)			Suppression	$f_{a\Delta b\Delta} \cong f_{b\Delta}, f_{a\Delta} < f_{b\Delta}$
(v)			Coequal	$f_{a\Delta b\Delta} \cong f_{a\Delta}, f_{a\Delta} \cong f_{b\Delta}$

Negative genetic interactions typically reflect the functional associations of two genes in parallel pathways [41], [87]. Many genes in DNA damage response, for example, have synthetic lethal interactions, suggesting the evolutionary importance of the functional redundancy to sustain the systematic integrity [43]. Ordering genes within pathways is possible through epistasis analysis using, for example, asymmetric positive interactions [88] while coequality can indicate the members of protein complexes that function as cohesive units [82]. The spectrum of positive interactions can be extended beyond these categories if the data is high resolution enough to capture these differences [89]. The depth of information afforded by quantitative genetic interaction analyses has been illustrated previously by several different studies [82], [90], [91].

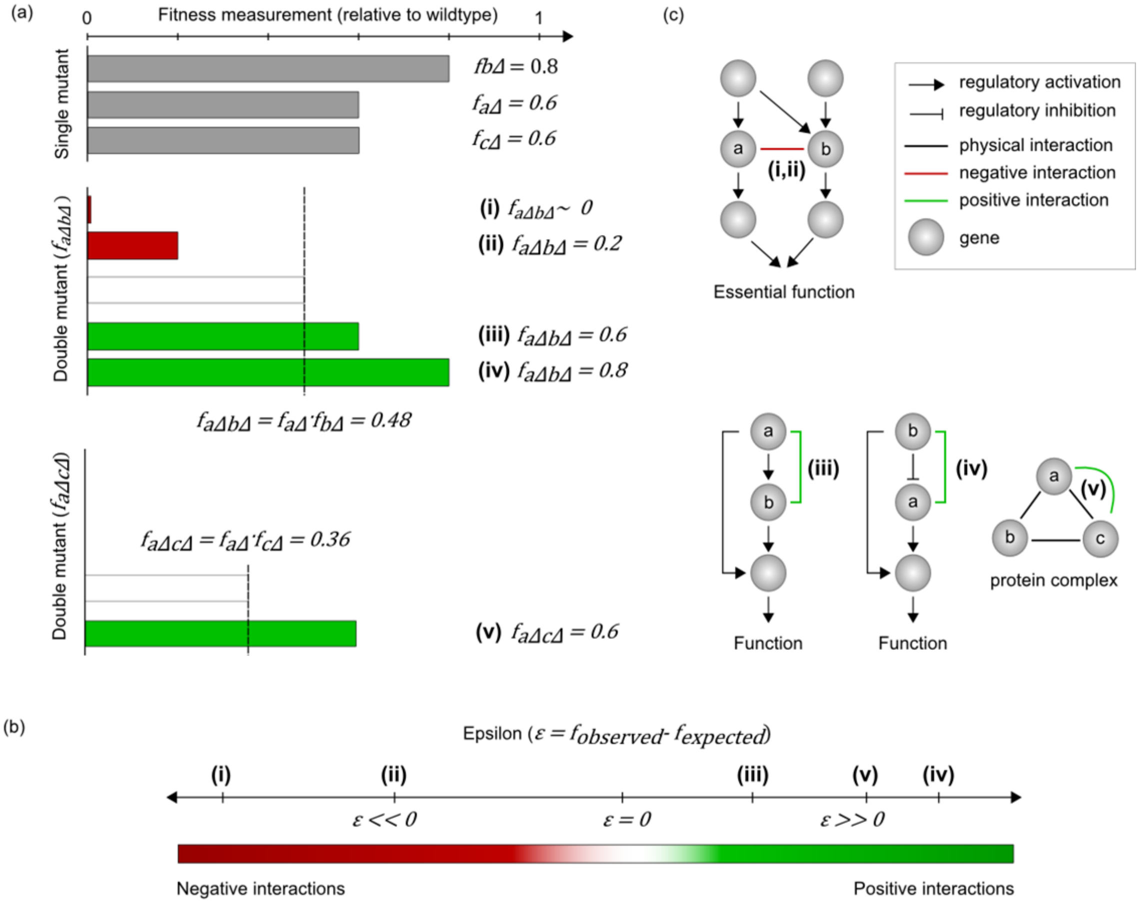


Figure 2-1 A graphical representation of how genetic interactions are inferred from a measurable phenotype (fitness measures from colony sizes) and interpreted in biological systems. (a) Single and double mutant fitness measures (b) Genetic interaction scores from observed and expected double mutant fitness measures in an epsilon space. (c) Possible biological interpretations of genetic interactions in a network space.

2.4 Dissecting the relationship between quantitative genetic and physical interactions

In our genome-wide SGA analysis [47], our collaborators screened 1712 *S. cerevisiae* query genes for a total of ~ 5.4 million gene pairs (~ 30 % of all combinatorial double mutants in *S. cerevisiae*) spanning all biological processes. Comparing fitness estimates of single mutants with their corresponding double mutant phenotypes identified ~170,000 genetic interactions. We compared the genetic interactions to the yeast physical interaction network as defined by TAP-MS [30],[97], Y2H [36] or protein-fragment complementation assay (PCA) [92]. In particular, we focused on genetic interactions within 161 annotated protein complexes in which more than one protein pair was screened for genetic interactions and measured the frequency of both positive and negative interactions within each complex. Consistent with smaller-scale studies [81], a large portion (92/161) of these complexes showed significant enrichment for genetic interactions ($p < 0.05$, hyper-geometric test). Within the enriched complexes, the majority of genes were linked to one another either by pure positive (46%) or pure negative (37%) genetic interactions (Figure 2-2) confirming previous theoretical observations [80].

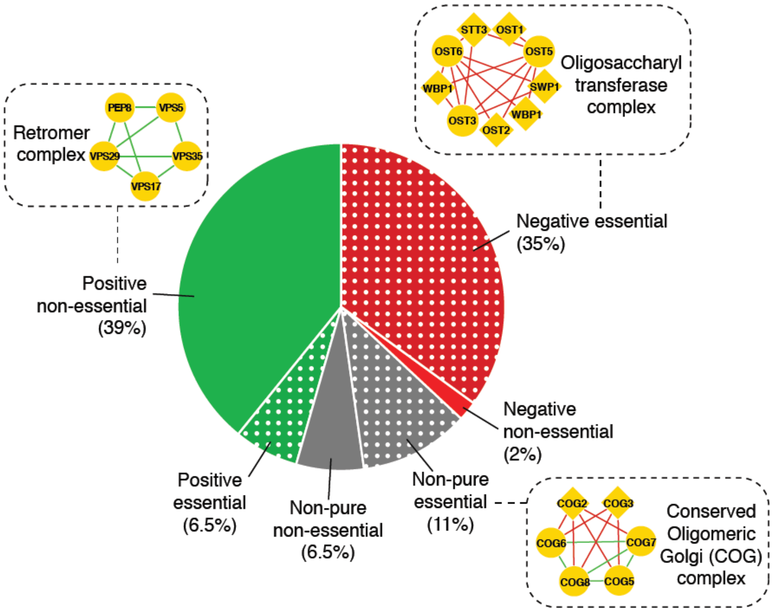


Figure 2-2 A set of 92 complexes enriched for negative and/or positive genetic interactions was assembled. Most complexes are biased towards purely positive interactions (green) or purely

negative interactions (red); grey denotes complexes composed of a mixture of positive and negative interactions. Purely negative and mixed protein complexes show a significant bias towards those containing an essential gene, as indicated by the dotted texture.

Strikingly, virtually all (94%) of the protein complexes that comprised exclusively of negative interactions contained at least one essential gene (Figure 2-2); which is much higher than appreciated previously[81]. Many of the genetic interactions within these negatively interacting complexes occur between non-essential genes, which comprise the majority of the genes on our network. This observation suggests that essential protein complexes may typically contain internal redundancy allowing the cell to tolerate the loss of a single non-essential component, whereas additional perturbation tends to result in the complete loss of complex function and impaired cell growth. In contrast, only 14% of all protein complexes associated with positively interacting gene pairs contained an essential gene (Figure 2-2). In addition to essentiality, another factor that appeared to determine the type of genetic interaction within a protein complex is the presence of a direct physical interaction, which can be found by targeted interaction screens such as yeast two-hybrid. We found that while co-complexed proteins with evidence of a direct physical interaction showed a nearly 3-fold bias towards positive genetic interactions, co-complex members with no evidence of a direct physical interaction showed a modest preference toward negative genetic interactions (Figure 2-3).

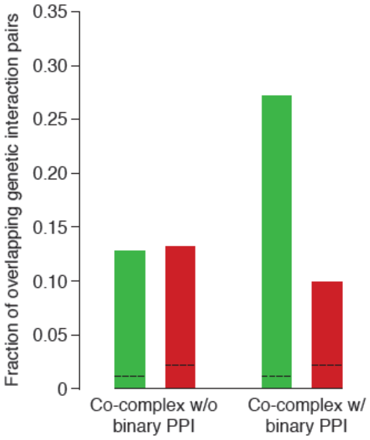


Figure 2-3 Genetic interaction frequency among co-complex members with and without a direct protein-protein interaction. Gene pairs within complexes annotated to our protein complex

standard were separated into two groups based on whether they also exhibited evidence for a direct protein-protein interaction as detected in a recent two-hybrid-based study [73]. For each group, the fraction of pairs exhibiting either positive (green) or negative (red) genetic interactions was measured after applying an intermediate confidence threshold ($|\varepsilon|$>0.08, p-value < 0.05). The background rate of positive and negative interactions among random gene pairs at the same confidence threshold is indicated by the black lines.

It is likely that even individual mutation of complex members with binary protein-protein interactions can induce almost complete loss of function of the protein complexes. It is also possible that these complexes with binary interactions can be divided into multiple subunits, which are transiently interacting each other in linear pathways.

Interestingly, complexes enriched for mixed interaction types (17.5%), some of which may be composed of multiple subunits whose genes show distinct subunit-specific double-mutant phenotypes, also showed a bias towards essential complexes (63%). An example of a mixed interaction complex is illustrated by the architecture of the Cog complex (Figure 2-2). Electron micrographs of this eight-member protein complex show a bi-lobed symmetry consisting of one non-essential (Cog5-8) and one essential (Cog1-4) lobe[93]. Indeed, we identified both negative and positive intra-complex interactions reflecting the highly organized structure of the Cog complex (Figure 2-2).

We also found that the predominant type of genetic interaction observed within a protein complex was predictive of its frequency of interaction with genes in other complexes. Nonessential genes within complexes connected by positive genetic interactions have an average of 2-fold more genetic interactions compared to nonessential genes within complexes connected by negative interactions (Figure 2-4, inset, rank-sum *p* $< 2 \times 10^{-3}$). The decreased number of interactions involving essential complexes may be explained by the fact that nonessential genes belonging to the same essential complex are functionally redundant. Thus, a single perturbation is less likely to compromise the activity of the protein complex and, as a result these nonessential genes may exhibit fewer genetic interactions within the context of the rest of the genome. Conversely, deletion of a nonessential gene within a positive complex may have a more severe impact on protein complex activity and, consequently, exhibits more genetic interactions.

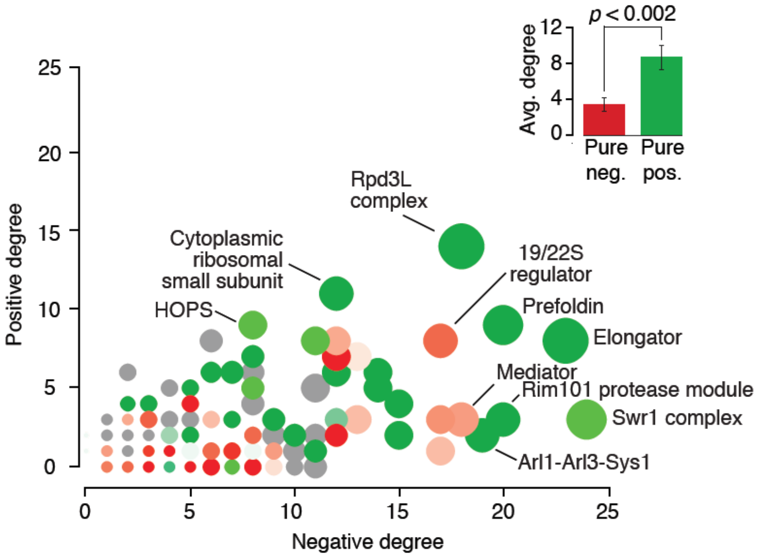


Figure 2-4 Degree analysis of the complex-complex genetic interaction network where the color of the node reflects the prevalence of positive (green) or negative (red) interactions within a given complex. Gray nodes represent complexes where too few gene pairs within the complex were screened to assess within-complex interactions. The size of the node indicates the number of proteins associated with the complex. (**inset**) The number of between-complex interactions was measured for purely negative and purely positive complexes to assess the relationship between within-complex interaction purity and number of interactions observed between different protein complexes.

Even though both positive and negative genetic interactions are enriched within protein complexes, the large majority of both negative and positive interactions did not overlap with physical interactions from high-throughput assays [27],[96]. Previous studies have focused on the relatively minor subset of rare positive genetic interactions that occur within protein complexes and pathways [81]; however, a genome-wide survey[47] revealed that the vast majority of positive interactions occur between complexes or pathways. In fact, only 0.5% of gene pairs with a significant positive interaction ($\varepsilon > 0.08$; $p < 0.05$) also shared a physical interaction. A number of the positive interactions that do not overlap with a protein-protein interaction may reflect functional relationships between complexes and indeed, we found 1182 complex pairs with significant enrichment for positive interactions connecting them (FDR 5%) (Figure 2-5).

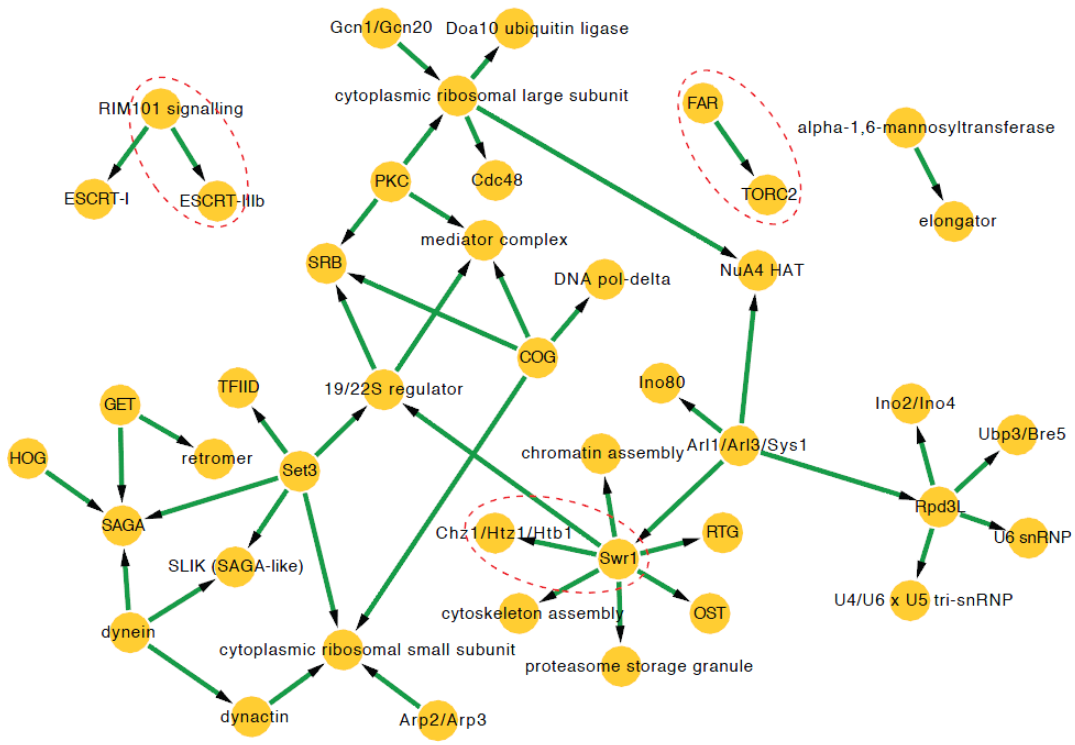


Figure 2-5 A network illustrating suppression interactions between protein complexes. The nodes represent protein complexes and the edges indicate positive SGA interactions classified as suppression based on comparison of colony size-based single and double mutant fitness of all gene-gene pairs annotated to the complexes. The arrows point to the complex whose fitness defect is suppressed.

2.5 Cross-complex genetic suppression networks

Given their prevalence in the genome-scale network, we examined some of the specific characteristics of positive interactions that link complexes. Relative comparison of single and double mutant phenotypes has been applied in small-scale studies to identify specific classes of positive interactions, including genetic suppression [82], [89]. Since the SGA scoring procedure produces single and double mutant fitness measurements, we adopted a similar strategy to examine genetic suppression between protein complexes on a global scale. In particular, we observed a surprising number of suppression interactions connecting across complex pairs, which suggest instances where loss-of-function mutations in one complex rescue growth defects associated with loss-of-function mutations in a second complex. We constructed a network to illustrate these protein complex interactions and provide a global view of such suppressor relationships (Figure 2-5).

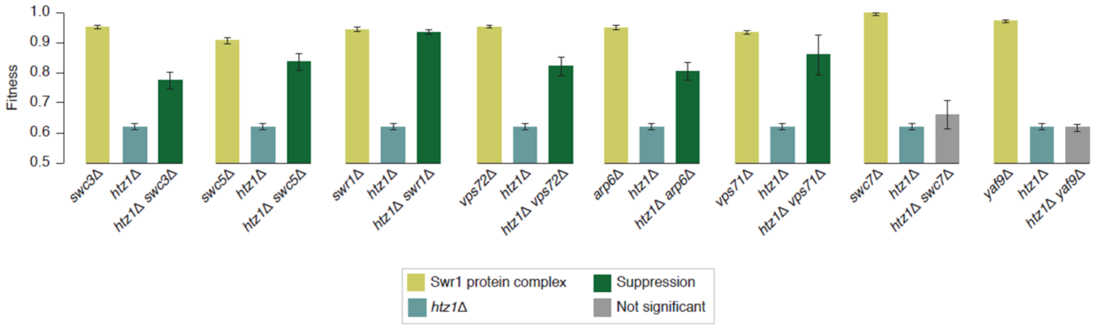


Figure 2-6 Comparison of colony size-derived single and double mutant fitness measures suggests that mutations in genes encoding members of the Swr1 protein complex (*swc3Δ*, *swc5Δ, swr1Δ, vps72Δ, arp6Δ* and *vps71Δ*) suppress growth defects associated with deletion of *HTZ1*.

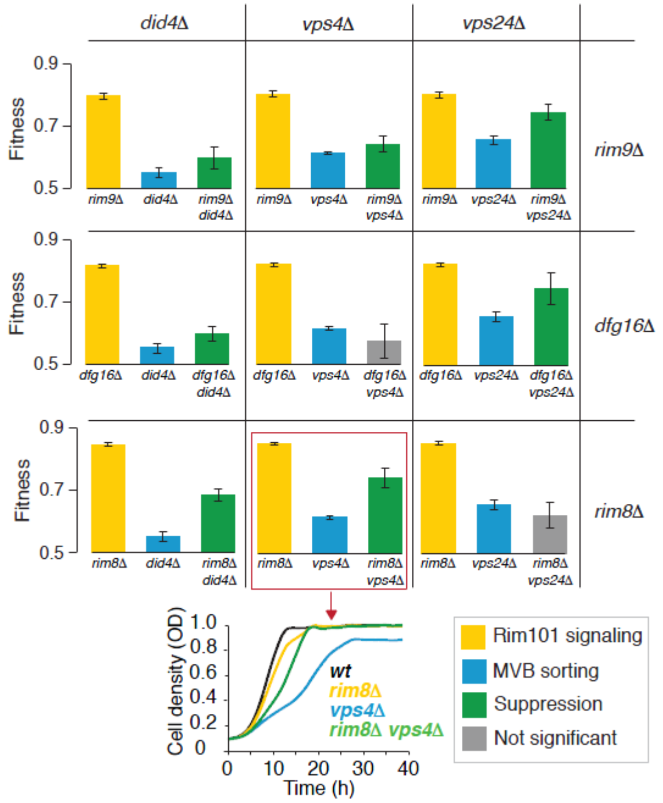


Figure 2-7 Comparison of colony size-derived single and double mutant fitness measures suggests that *rim8Δ*, *rim9Δ* and *dfg16Δ* suppress growth defects associated with deletion of *DID4*,

VPS4 or *VPS24*. (**inset**) Colony size-based fitness measures were confirmed by liquid growth profiling as described previously [82].

2.6 Independent experimental validation of suppression interactions

Several cases of cross-complex suppression interactions were experimentally confirmed, including rescue of *htz1*Δ slow growth by deletion of genes encoding components of the Swr1 protein complex (Figure 2-5 and Figure 2-6), an observation that was confirmed previously in a high-resolution growth competition assay [91]. We also validated novel suppression interactions within our network. For example, we confirmed loss-of-function suppression interactions involving the Rim101 signaling pathway genes, *RIM8*, *RIM9* or *DFG16,* and genes encoding multi-vesicular body (MVB) sorting proteins, including *DID4* and *VPS24*, which code for members of an ESCRT-III sub-complex, and the AAA-type ATPase gene, *VPS4* (Figure 2-5 and Figure 2-7). A functional relationship between Rim101 signaling and MVB sorting has been established previously [94]. *RIM101* encodes a zinc-finger transcription factor activated in response to alkaline growth conditions via proteolytic cleavage [95]. Rim101 processing is regulated by a signaling pathway that associates with endosome membranes through interactions with a specific ESCRT-III sub-complex, Snf7-Vps20 (Figure 2-8(a)). Our suppression network highlighted that deletion mutations in genes encoding upstream signaling components of the Rim101 pathway suppress the fitness defect associated with deletion alleles of *DID4*, *VPS24* or *VPS4,* which suggests that the suppression relationship occurs because a defect in upstream signaling prevents constitutive activation caused by loss-of-function of downstream negative regulators. Consistent with this possibility, we found that *did4Δ*, *vps24Δ* and *vps4Δ* mutants were sensitive to *RIM101* overexpression and this sensitivity was rescued by deletion of *RIM8* (Figure 2-8(b))*,* a putative upstream component of the Rim101 pathway (Figure 2-8(b)) [96].

Interestingly, under conditions where the Rim101 pathway is activated and required for viability, in the presence of lithium chloride (LiCl), suppression is observed in the opposite direction[96]. In LiCl, deletion alleles of *RIM8* and *RIM9* lead to fitness defects presumably because they block pathway signaling, whereas mutation of *DID4*, *VPS24* or *VPS4* suppress this phenotype [96]. These observations highlight the importance of considering condition-specificity when inferring pathway architecture based on genetic interactions. As noted previously [88], understanding when a pathway is active and/or required for cell growth is essential for accurate interpretation of genetic suppression or masking relationships.

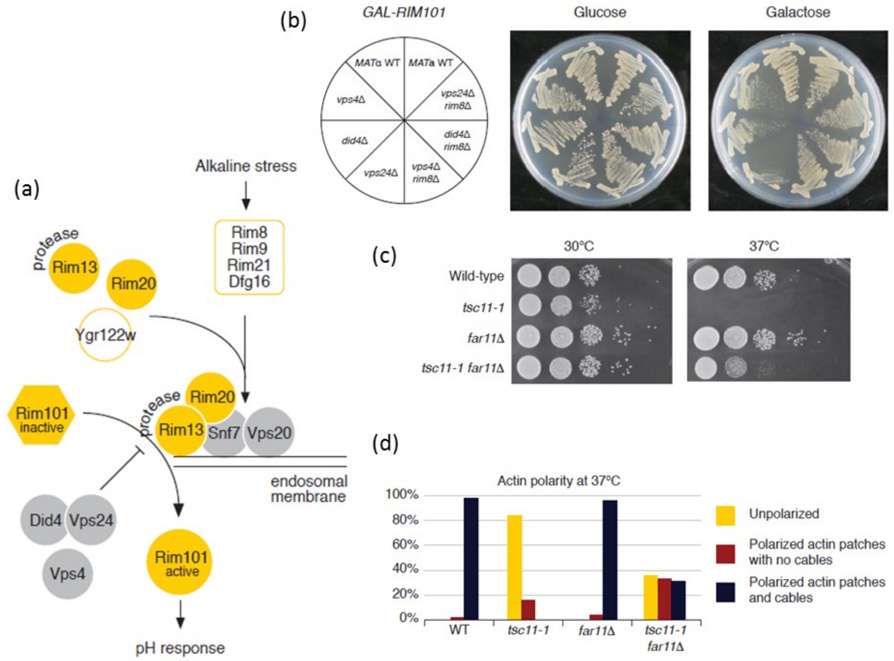


Figure 2-8 (**a**) In response to alkaline stress, the Rim101 pathway is activated by an unknown mechanism and recruits ESCRT complexes to the endosomal membrane. Rim13 and Rim20 are then recruited to the endosome membrane by an ESCRT-III sub-complex, Vps20-Snf7, which leads to Rim101 processing. Following Rim101 activation, a second ESCRT-III sub-complex, Did4-Vps24, along with Vps4 promote dissociation of protein complexes from endosomal membranes thereby inhibiting Rim101 proteolytic processing. (**b**) *rim8Δ* suppresses *did4Δ*, *vps24Δ* and *vps4Δ* sensitivity to *RIM101* over-expression. (**c**) Serial dilution growth assays confirm that *tsc11-1* growth defects are suppressed by deletion of *FAR11* under semi-permissive (30ºC) and non-permissive (37ºC) conditions. (**d**) A *tsc11-1* temperature-sensitive mutant exhibits an abnormal actin morphology that is suppressed by *FAR11* deletion. The extent of actin polarization was quantified in wild type and the indicated single and double mutants.

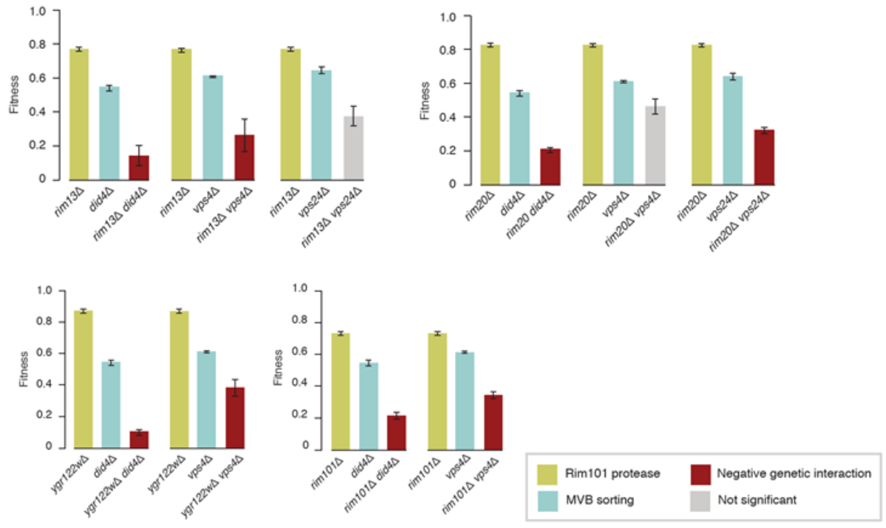


Figure 2-9 Comparison of colony size-derived single and double mutant fitness measures suggests that deletion of *RIM13*, *RIM20*, *YGR122W* or *RIM101* do not suppress growth defects associated with loss of *DID4*, *VPS4* or *VPS24*.

Unlike *rim8Δ* or *rim9Δ* mutants, other Rim101 pathway mutants, such as *rim13Δ* or *rim20Δ* mutants were not rescued by *did4Δ*, *vps24Δ* or *vps4Δ* mutations [96]. Similarly, we failed to observe *rim13Δ*– or *rim20Δ*–dependent suppression under our SGA conditions (Figure 2-9) indicating a functional distinction between Rim8/Rim9 and Rim13/Rim20 in the Rim101 signaling pathway (Figure 2-8(a)) [96]. Consistent with previous findings[97], our genetic interaction analysis suggests that the Rim pathway component, Dfg16, functions upstream in the pathway with Rim8 and Rim9 while, the previously uncharacterized gene, Ygr122w, likely functions further downstream in the pathway closer to the Rim13 protease (Figure 2-8(a) and Figure 2-9).

Another example showed that disruption of the FAR complex, originally implicated in cell cycle control[98], rescued growth defects associated with mutant alleles of TORC2 kinase complex gene members, *tor2-29* and *tsc11-1* (Figure 2-5, Figure 2-8(c), and Figure 2-10(a)-(b)). Moreover, deletion of *FAR11* suppressed actin polarization defects of a *tsc11-1* mutant at a non-permissive condition (37°C, Figure 2-8(d)). These results suggest that the FAR complex may function downstream to negatively regulate the role of the TORC2 complex in actin organization. Similar to TORC2, FAR complex members are conserved from yeast to humans, and mammalian Far protein orthologs belong to a large multi-protein complex that contains the serine/threonine protein phosphatase, PP2A [99]. Interestingly, suppression of TORC2 growth and actin polarity defects was also achieved by loss of *PPG1* (Figure 2-10). Thus, it is possible that the FAR complex mediates its function by working together with Ppg1 to dephosphorylate and inactivate proteins that normally control actin-based cell polarity.

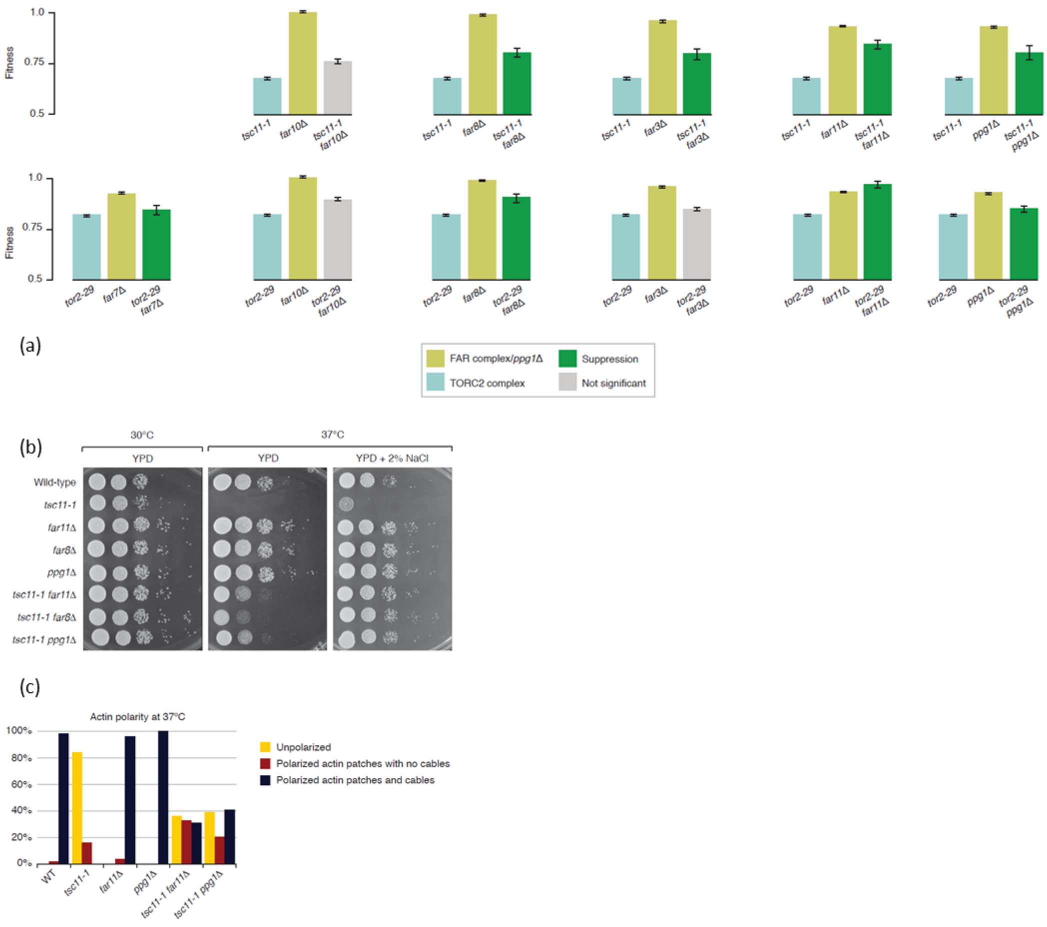


Figure 2-10 **(a)** Comparison of colony size-derived single and double mutant fitness measures suggests that mutations in genes encoding members of the Far3-11 protein complex suppress growth defects associated with mutant alleles of TORC2 kinase complex gene members, *tor2-29* and *tsc11-1*. Similarly, *tor2-29* and *tsc11-1* mutants are also suppressed by deletion of *PPG1*, a PP2A-related serine/threonine phosphatase. **(b)** Serial dilution growth assays confirm that *tsc11-1* growth defects are suppressed by deletion of *FAR11*, *FAR8* and *PPG1* under semi-permissive

(30ºC) and non-permissive (37ºC) temperatures. Suppression interactions are enhanced in the presence of 2% (0.4M) NaCl. **(c)** A *tsc11-1* temperature-sensitive mutant exhibits an abnormal actin morphology that is suppressed by loss of *FAR11* and *PPG1*. The extent of actin polarization was quantified in wild type and the indicated single and double mutants.

2.7 Conclusion

The combination of scale and resolution afforded by SGA analysis and the SGA interaction score allowed us to examine genetic interactions among ~5.4 million gene pairs spanning all biological processes [47]. This genome-scale, fitness-based analysis identified ~170,000 genetic interactions [47], with approximately two-fold more negative than positive interactions, and enabled a systematic comparison of quantitative genetic and physical interaction networks revealing a more complex relationship than previously appreciated. Contrary to the generalized between-within-pathway model [52] in which positive genetic interactions connect members of the same non-essential protein complex while negative interactions occur between non-essential protein complexes, the overlap with protein-protein interactions was similar for both positive and negative genetic interactions [47]. Although the between-within-pathway model provides initial insight for interpreting genetic interactions in a physical context, it does not completely explain the observed genome-wide data. More in-depth analyses on the interpretation of genetic interactions with other physical interaction data including protein-DNA interactions can help characterize the genetic interactions in higher resolution. Moreover, genome-wide genetic interaction data in other organisms available in the near future will increase our ability to develop more finely-tuned models, which can be applicable for the functional reorganization of overall biological processes and pathways.

The successful generation of the 1st genome-wide unbiased genetic interaction data in yeast also motivated us to systematically characterize the nature of positive genetic interactions. With the comprehensive genetic epistasis analyses from quantitative genetic interactions, we successfully revealed many suppression relationships between cellular components, some of which were confirmed by high-quality experiments. These results suggest that quantitative positive interactions are highly capable of understanding the genetic basis of the cellular network in a genome scale. A network of loss-of-function suppression interactions (Figure 2-5) illustrated the ability of positive interactions to capture broad phenotypic relationships by connecting genes belonging to functionally diverse protein complexes. Consistent with the findings here, an *in silico* study focused on essential metabolic genes found that negative interactions occur more frequently between genes with overlapping function while positive interactions are observed between functionally distinct metabolic pathways. However, unlike the genome-wide survey indicating that negative interactions are twice as prevalent within the yeast genetic network [47], this theoretical analysis suggests that positive interactions are surprisingly more abundant than negative interactions in both yeast and *E. coli* metabolic networks. Whether this is a specific characteristic of metabolic networks or a more global network property related to gene essentiality remains to be explored.

In addition to distant functional relationships, cross-complex suppression analysis can also identify genes or pathways that act downstream and normally negatively regulate another pathway. Preliminary analysis suggests the resolution of SGA scored interactions may be sufficient for reconstructing pathways based on the relative strength of single and double mutant phenotypes (Figure 2-7 and Figure 2-8(a),(b)). Hence, genome-wide application of this method may be an important step toward moving beyond relatively abstract prediction of gene function toward construction of specific, mechanistic models from interaction data.

2.8 Supporting methods

2.8.1 Quantitative genetic interaction data from genome-wide SGA analyses

SGA methodology enables rapid and systematic construction of yeast double mutants by mating a strain harboring a ‘query’ mutation of interest to input arrays of strains carrying different ‘array’ mutations, which are composed either of nonessential deletion mutants or conditional alleles of essential genes [100]. After several robot-facilitated selection steps (Figure 2-9(b)), the final output arrays (Figure 2-9(c)) consisting of haploid double-mutant colonies are imaged at a single time point [48], [49] (Figure 2-9(a)). We developed a model for relating the area of a double-mutant colony image to the fitness of the constituent single mutants by assuming that, in the absence of genetic interactions, double-mutant fitness is a multiplicative combination of single-mutant fitness and experimental factors [84]. Then we measured genetic interactions as deviations from the expected double-mutant fitness.

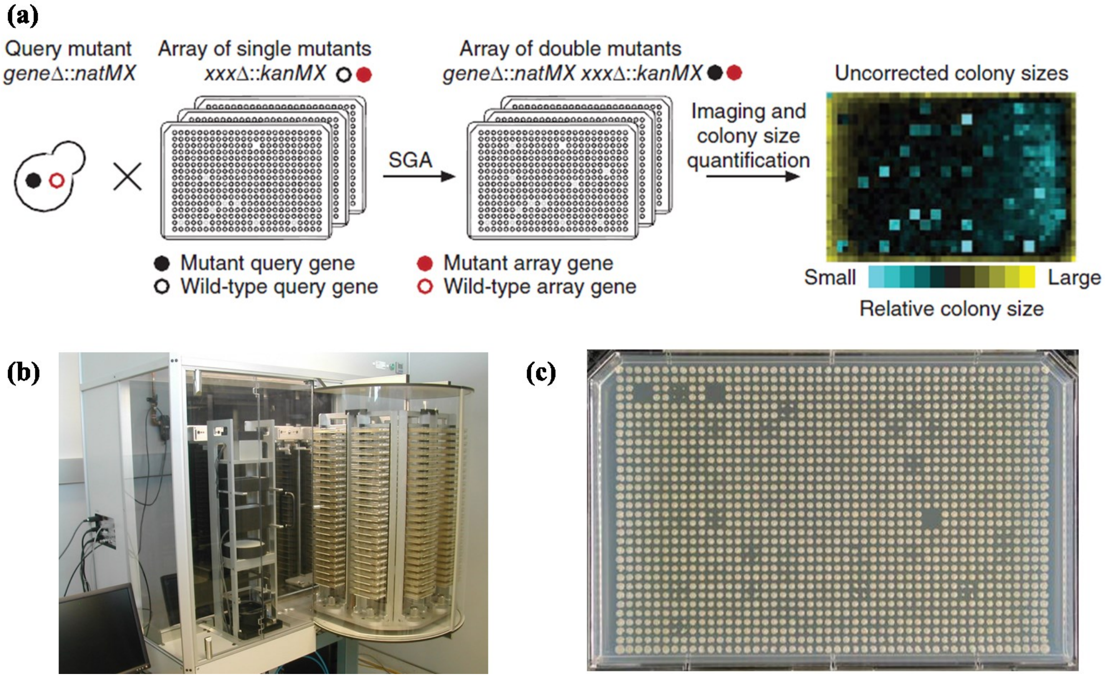


Figure 2-11 A high-throughput SGA experiment. (a) a SGA experiment crossing a strain carrying a query mutation to an input array of single mutants, each of which carries a wild-type copy of the query gene and a unique array strain mutation. A final output array of double mutants is generated after several SGA selection steps, photographed and processed using software that measures colony areas in terms of pixels. Relative colony size, determined by measuring deviation of individual colonies from the median size for the same colony across 1,712 different experiments [47] is shown. (b) Next-generation robotics for SGA analysis (Charles Boone's Lab) (c) A plate of double mutants from a SGA screen. This plate is the result of a single query crossed into a plate of array single mutants. Each double mutant appears with four replicates, and there are 1536 total colonies on each plate.

To derive quantitative genetic interactions, we modeled colony size as a multiplicative combination of the double mutant fitness, time, and experimental factors. Specifically, for a double mutant carrying mutations of genes i and j, colony size C_{ij} can be expressed as $C_{ij} = f_{ij} \cdot t \cdot s_{ij} \cdot e$, where f_{ij} is the double mutant fitness, t is time, s_{ij} is the combination of all systematic factors (Figure 2-12(a)), and e is log-normally distributed random noise. The double mutant fitness f_{ij} can be further expressed as $f_{ij} = f_i f_j + \varepsilon_{ij}$, where f_i and f_j are the fitnesses of the two single mutants and ε_{ij} is a quantitative measure of the genetic interaction between them. To derive accurate estimates of single mutant fitness, we applied our correction method to a set of control SGA screens, where the queries carried a mutation in a neutral genomic locus (Figure 2-12(b)). The obtained single mutant fitnesses (f_i and f_j) were combined with the double mutant fitnesses (f_{ij}) estimated from regular SGA experiments to derive genetic interaction measures as $\varepsilon = f_{ij} - f_i \cdot f_j$. A statistical confidence measure (p-value) was assigned to each interaction based on a combination of the observed variation of each double mutant across four experimental replicates and estimates of the background lognormal error distributions for the corresponding query and array mutants. The Estimates of error distribution for each array and query mutant were estimated separately from the set of all other double mutants carrying the corresponding array or query mutation, respectively.

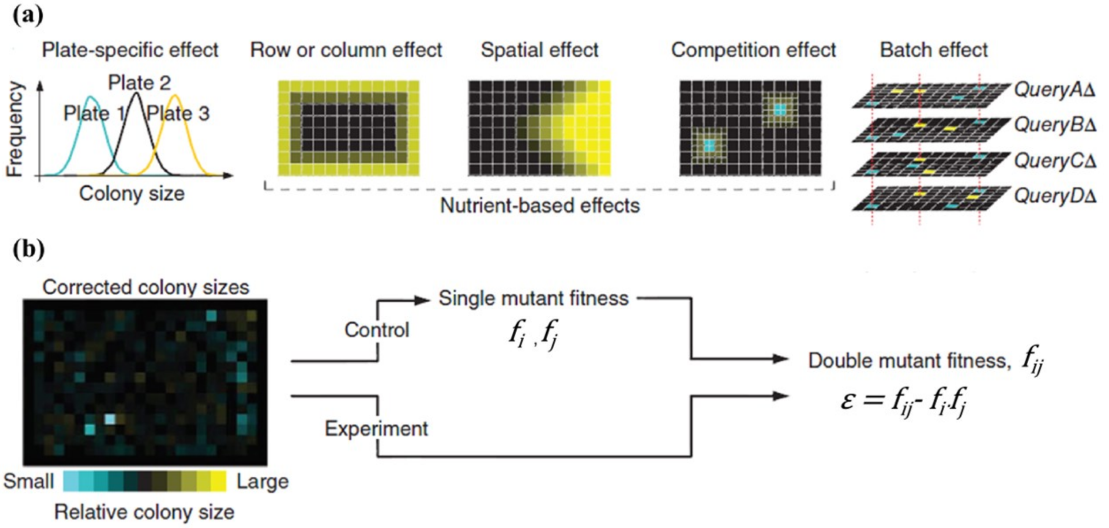


Figure 2-12 The SGA scroe of measuring quantitative genetic interactions. (a) Schematic depiction of the five factors that contribute to experimental variance of colony size. (b) Relative colony size after normalization. Single-mutant fitness (f_i, f_j) and double-mutant fitness (f_{ij})

derived from normalized colony size measurements were used to identify and measure genetic interactions (SGA score; ε)

The SGA genetic interaction dataset is composed of 1712 queries (including 334 conditional or hypomorphic alleles of essential genes) crossed to 3885 array strains. The dataset contains raw genetic interaction scores for ~5.4 million gene pairs spanning all biological processes (Figure 2-13(b)) and approximately 170,000 interactions are identified (Figure 2-13(a)). For all analysis of genetic interactions within and between protein complexes, we used interactions at the lenient cutoff ($p < 0.05$) to maximize coverage of physically interacting pairs.

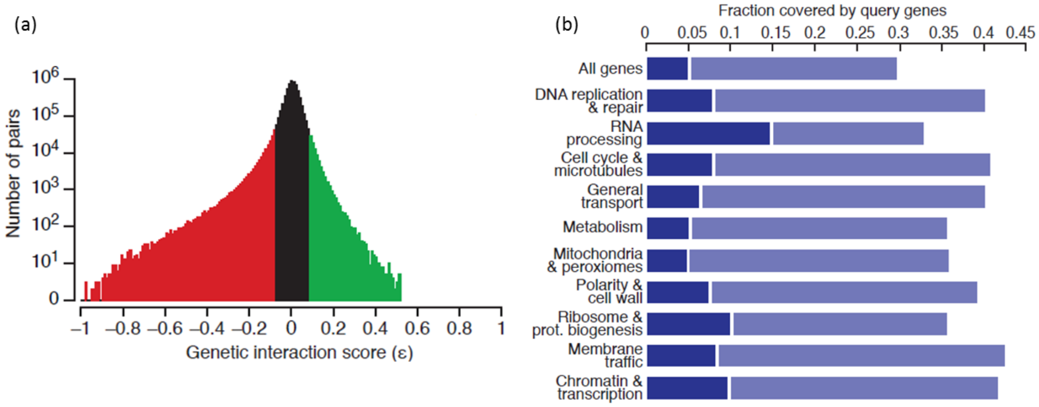


Figure 2-13 A SGA dataset. (a) Distribution of genetic interaction scores on a logarithmic scale. Most double mutant pairs show no genetic interaction (low *ε* values, black), while fewer pairs exhibit negative (red) and positive (green) interactions. Negative interactions are approximately 2-fold more prominent than positive interactions. (b) Functional coverage evaluated on the basis of the extent to which query mutant strains screened in SGA covered 10 broad functional categories. The 10 groups were defined using a Bayesian framework for data Integration [101]. Approximately 30% of all genes were screened genome-wide by SGA, which translates to at least 35% coverage of each partially overlapping functional category (light blue). Dark blue bars indicate the proportion of essential genes screened in each category.

2.8.2 Yeast physical interaction data

2.8.2.1 Protein-protein interaction standard

Protein-protein interactions were downloaded from BIOGRID[102] on Sept. 17, 2009.

2.8.2.2 Construction of a protein complex standard

A literature-curated protein complex standard was compiled by combining the two most recent protein complex standards available for yeast. This standard consists of 430 complexes derived from SGD Macromolecular Complex GO standard (www.yeastgenome.org), CYC2008 protein complex catalog [103] and 26 manually curated complexes/pathways by biology experts. Redundant protein complex annotations were minimized by eliminating all but one complex with identical components and by excluding smaller complexes if all their members also belong to a larger complex. Partially overlapping complexes were treated as separate complexes. For the enrichment analysis of genetic interactions within protein complexes, each complex for which at least 2 pairs were screened for genetic interactions was assessed for enrichment of either positive or negative interactions among its members. All dubious ORFs were ignored and the union of genetic interactions for each gene with multiple alleles screened was considered as a final set of genetic interactions associated with it.

2.8.3 Enrichment analysis of genetic interactions within protein complexes

Each complex for which at least 2 pairs were screened for genetic interactions was assessed for enrichment of either positive or negative interactions among its members. We ignored all dubious ORFs and considered the union of genetic interactions for each gene with multiple alleles screened. Significance was evaluated using the hyper-geometric distribution as follows:

$$\text{P-value} = 1 - \sum_{n=0}^{X-1} \frac{\binom{K}{n}\binom{M-K}{N-n}}{\binom{M}{N}} \qquad \text{Eq. 2-1}$$

where *M* is a total number of screened gene pairs in the genome-wide genetic interaction data[48], *K* is a total number of gene pairs with positive/negative genetic interactions, *N* is a total number of screened gene pairs within protein complex, and *X* is a total number of interacting (positive/negative) gene pairs within protein complex.

2.8.4 Monochromatic analysis of genetic interactions within protein complexes

To evaluate the monochromaticity of interactions within protein complexes, we measured the purity of interactions for any complex with enrichment for interactions by the criteria described above (p-value < 0.05). Specifically, we defined a monochromatic purity score (MP-score) as follows:

$$\mathrm{MP}(C_i) = \frac{\frac{1}{N_i}\sum_{j,k\in C_i} e_{jk} - \alpha}{1 - sign\left(\frac{1}{N_i}\sum_{j,k\in C_i} e_{jk}\right) * \alpha} \qquad \text{Eq. 2-2}$$

where $\alpha = \frac{1}{N_{tot}}\sum_{\forall j,k} e_{jk}$, N_i is a number of screened pairs within complex i, N_{tot} is a number of total screened pairs, C_i is a set of all possible proteins (genes) in complex i, e_{jk} is +1 or -1 if genes j and k have a positive or negative genetic interaction, respectively. This score is designed such that a complex with pure positive genetic interactions will have an MP-score equal to +1, while a complex having all negative pairs will have an MP-score equal to –1. A complex reflecting the background ratio of positive to negative interactions will present a MP-score equal to 0. Monochromatic complexes in Figure 2-2 and Figure 2-4 were complexes satisfying $|MP(C_i)| > 0.5$ (Eq. 2-2). A distribution of all complexes monochromatic purity scores is shown in Figure 2-14 (a).

2.8.5 Construction of a complex-complex network of genetic interactions

To enable a study of interactions at the module level, we defined a complex-complex network of genetic interactions. Complex-complex pairs were assessed for enrichment of genetic interactions as well as monochromaticity. Enrichment of interactions within complex-complex pairs was assessed as follows:

$$\text{P-value} = 1 - \sum_{n=0}^{X-1} \frac{\binom{K}{n}\binom{M-K}{N-n}}{\binom{M}{N}} \qquad \text{Eq. 2-3}$$

where M = |{screened partners outside from complex i}∪{screened partners outside from complex j}|, K = |{genetic interaction partners outside from complex i} ∪ {genetic interaction partners outside from complex j}|, N is a total number of screened gene pairs between complex *i* and *j*, and *X* is a total number of genetic interaction pairs between complex *i* and *j*.

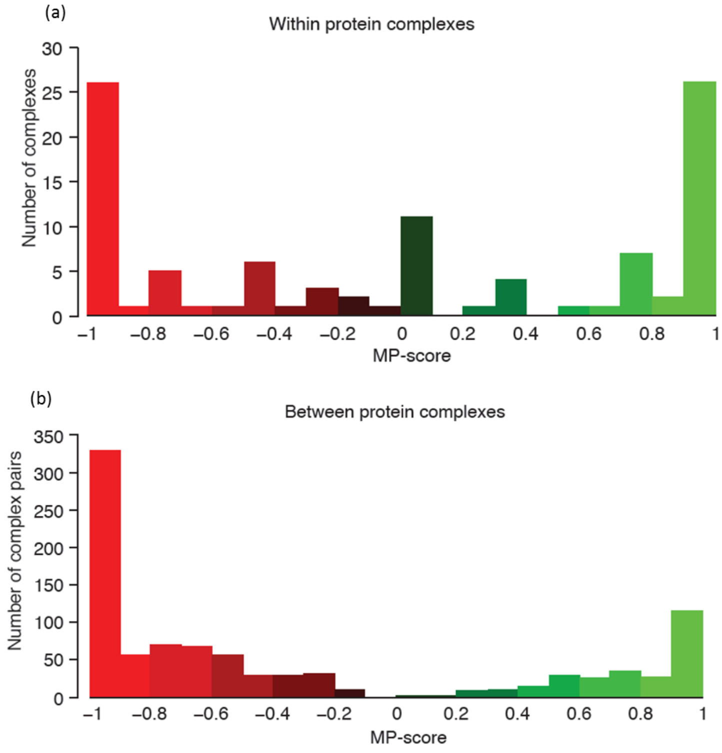


Figure 2-14 (a) Distribution of complexes with respect to within-complex monochromatic purity scores (MP-scores, Eq. 2-2). Only complexes having at least two genetic interactions based on the lenient cutoff (p-value < 0.05) are considered here. (b) Distribution of complex-complex pairs with respect to between-complex monochromatic purity scores (MP-scores). Edges between complexes are retained when the pair is enriched for genetic interactions (FDR = 5%)

Between-complex interaction analysis was restricted to positive and negative interactions identified at the lenient cutoff ($p < 0.05$). Monochromaticity was assessed similarly to the approach used for within-complex monochromaticity (Section 2.8.4) but based on all gene-pairs spanning across each pair of complexes. The complete distribution of between-complex MP scores is shown in Figure 2-14 (b). For the complex-complex network degree analysis in Figure 2-4, a between-complex edge in the modular network was conservatively assumed to exist where the pair had significant enrichment (false discovery rate 5% based on hyper-geometric distribution and Westfall and Young step-down procedure for multiple hypothesis correction) and the between-complex MP score satisfied $|MP(C_i–C_j)| > 0.75$. The complex-complex network degree of purely negative (within MP-score = –1) and purely positive (within MP-score = 1) complexes was measured and compared (Figure 2-4, inset), and the Wilcoxon rank-sum test was used to assess the significance of the difference in degree between complexes connected by pure negative and pure positive genetic interactions.

2.8.6 Protein complex suppression network

To construct a complex-complex suppression network, we first identified all complex-complex pairs with significant enrichment for positive interactions and a minimum of 5 shared positive genetic interactions. We then categorized the between-complex positive interactions into suppression and masking subclasses. Assuming that f_a, f_b and f_{ab} are the fitness measures for single mutants a, b and the double mutant ab, respectively, and that $f_a = \max(f_a, f_b)$, we used the following rules: (1) if $f_{ab} > f_b + \sigma_b$ (where σ_b is the standard deviation of f_b), mutant a was determined to suppress the phenotype of mutant b; (2) if rule (1) was not true, but if $f_{ab} < f_a - \sigma_a$, . then mutant b was determined to mask the phenotype of mutant $a f_{sm\_max}(= f_{sm1}) - \sigma_{sm\_max}(= \sigma_{sm1}) > f_{dm}$. Prior to this analysis, all positive interactions were filtered in the following way: a) $\varepsilon > 0.08$; b) $p < 0.05$; c) $f_a, f_b < 1$; d) $|f_a - f_b| \geq \sigma_a + \sigma_b$. We identified all complex-complex pairs with greater than 80% directional and suppression consistency among the corresponding gene pairs and visualized them as a network (Figure 2-5).

3 Modeling a plant immune signaling network with combinatorial perturbation data: Bayesian networks

3.1 Chapter overview

Plants constantly interact with a wide variety of microbial pathogens with different lifestyles and infection strategies. The evolutionary arms race between plants and their invaders resulted in a highly sophisticated defense system in plant. Plants recognize pathogenic attack and transduce the information through signaling networks and defense responses are induced. The outcomes from the interactions between plants and their attackers can dramatically affect crop yields since the induced defense responses entail fitness costs in plant. Thus, plants must be equipped with elaborate regulatory mechanisms to efficiently coordinate the activation of intricate signaling pathways and attain the optimal resistance with minimal fitness costs. Recent advances in plant immune research have provided new insights into the underlying defense system. The signaling pathways of diverse phytohormones in plants play pivotal roles in cross-communication between different subsystems to finely regulate immune responses. However, it is still unclear what and how key mechanisms facilitate optimal responses against such diverse attackers. Addressing these problems requires computational methods, both for supporting the development of network models and for the analysis of their functionality.

In this chapter, we develop a probabilistic modeling approach based on a Bayesian network to understand the dynamics of complex signaling network in the plant immune system. In particular, we build a dynamic network model to infer cross-communications among four major signaling sectors which provide much of the network backbone during pattern-triggered immunity (PTI) in Arabidopsis. Our experimental collaborators constructed Arabidopsis mutants in which all combinations of genes required for four signaling sectors were disrupted (16 combinatorial genotypes, including wild-type) [3]. In each of 16 genotypes, both the expression levels of four representative marker genes for the signaling sector activities at two time points and the levels of PTI against two strains of bacterial pathogens were measured after treatment with Microbe-

Associated Molecular Patterns (MAMPs), including flg22, elf18, and chitosan, or mock. Given these data, we searched candidate network structures by applying Elastic Net, and assessed the model structures with inferred parameters by applying a Bayesian network model. In the Bayesian network framework, we then iteratively sought plausible structures from model consensus and evaluate their models with their predictive power and the structure to find a final model. The resulting model not only confirmed well-known mechanisms such as a cross-activation between PAD4 and SA but revealed previously unappreciated mechanisms and network properties. The work presented in this chapter was submitted to Cell Host & Microbe and includes contributions from Kenichi Tsuda, Daisuke Igarashi, Rachel A. Hillmer, Fumiaki Katagiri, and Chad L. Myers. Kenichi and Daisuke generated experimental data, and Rachel suggested some ideas for the statistical analysis. Fumiaki supervised preparations for the data and data preprocessing and Chad supervised the Bayesian network modeling.

3.2 Background: modeling a plant immune signaling network based on a Bayesian network approach

Inducible immunity is a major component of plants' immunity against pathogens in which plants recognize pathogen attack and transduce this information through signaling networks within the cell, to different cells, and to distant tissues, and defense responses, some of which affect fitness of the pathogen *in planta*, are induced [58]. Although this sequence of recognition, signal transduction, and response is the common theme in any biological response to external or internal stimuli, plant inducible immunity is unique in that pathogens not only initiate the signaling event but also attack the signaling network. In the PTI system, pattern recognition receptors (PRRs) of a plant recognize molecular patterns that are relatively conserved among similar types of microbes, called microbe-associated molecular patterns (MAMPs) [58], [104]. For example, a part of bacterial flagellin (flg22), a part of bacterial elongation factor-Tu (elf18), and an oligosaccharide part of fungal cell wall (chitin) are recognized in *Arabidopsis thaliana* by the receptor-like kinase PRRs [105], [106], [107]. Such recognition of MAMPs by the cognate PRRs initiates PTI signaling.

Quantitative relationships among four principal signaling sectors during PTI in Arabidopsis, the jasmonate (JA), ethylene (ET), PAD4, and salicylate (SA) sectors, were previously investigated [3]. JA, ET, and SA are phytohormones important for immune signaling, and their signaling can be abolished by mutations in the genes, *DDE2*, *EIN2*, and *SID2*, respectively [108], [109], [110]. The *PAD4* gene affects the SA level as well as many SA-independent responses [111]. In a quadruple mutant *dde2/ein2/pad4/sid2*, the level of immunity triggered by flg22 (flg22-PTI) against *P. syringae* pv. *tomato* DC3000 (*Pto*) was diminished to 20% of the wild-type level [3]. The contributions to flg22-PTI of four signaling sectors and their 2-sector, 3-sector, and 4-sector interactions were examined by signaling allocation analysis [3]. The signaling allocation analysis conceptually reconstitutes the signaling network from the near-ground state of the quadruple mutant by measuring the immunity level in all 16 Arabidopsis combinatorial genotypes regarding the four sectors. In flg22-PTI the interactions that involve both the PAD4 and SA sectors had synergistic contributions to immunity, but the other interactions were compensatory. The compensatory interactions indicate robustness against perturbations, which could provide a mechanism for the network to withstand attack from fast-evolving pathogens.

To mechanistically understand how the four sectors contribute to the plant immune responses to various external signals, we applied combinatorial perturbations of the four sectors and collected measurements of activity levels in each sector. We then constructed immune signaling network models based on a Bayesian network approach by combining two different types of data: sector activities of the four sector markers and log10-transformed bacterial counts with 4 different conditions, 3 MAMPs (flg22, elf18, and chitosan) and mock, for all possible combinatorial genetic perturbations. The Bayesian network, a graphical model that encodes probabilistic relationships among variables of interest, has three principal advantages in our setting: easy incorporation of diverse data with both discrete and continuous variables, favorable Bayesian statistical methods for avoiding overfitting, and ideal representation for combining prior knowledge and perturbation data. The two components of a Bayesian network model are the network structure, which defines conditional dependence relationships between the variables being modeled, and the conditional probability tables, which quantify these dependencies. In this case, the objective is to infer both the structure of the network (under the structural constraints listed above) and the optimal conditional probability tables given the proper structure. For defining the structure, rather than using a Bayesian structure learning approach [112], a regularized linear modeling approach, specifically Elastic Net [113], was first used to reduce the search space to a small set of reasonable structures that capture relationships between the network sectors. Starting from these candidate structures, the conditional probability parameters were fit to the observed data, using cross-validation to assess the performance of each inferred model. By evaluating the candidate models with the prediction accuracies and selecting the significant model structures, we chose the final model which reflects the dynamics of the four principal sectors and predicts variant immune responses under different genotypes and conditions. The final model not only captures well-known cross-talk structures but reveals novel mechanistic relationships among the sectors such that our model is indeed a useful tool for navigating the emergent properties of the plant immune system.

3.3 A probabilistic graphical model for plant immune signaling network

The structure of the Bayesian network for our setting consists of four layers: (1) an input layer with one ternary node for three MAMPs including flg22, elf18, and chitosan, (2) an activation layer with four binary nodes of four sectors (JA, ET, PAD4, and SA) at 3 hours post treatment (hpt), (3) a cross-talk layer with four binary nodes of the four sectors at 9 hpt, and (4) an output layer with two continuous nodes for log-transformed bacterial counts of two bacterial strains (*Pto* and *Pma*) (Figure 3-1).

Given the network structure as a baseline framework, to model a plant immune system, we implemented several modeling approaches summarized in Figure 3-2. We first collected two different types of quantitative measurements of both dynamic activity levels of the four sectors and log10-transformed bacterial counts with all 16 combinatorial genetic perturbations and four treatments including three MAMPs or mock (see Section 3.6 in detail). After we properly normalized the raw mRNA expression values of the four sectors and preprocess the two types of data, Elastic Net was initially used to reduce the search space of hidden structures ($2^{(3x4)}$ = 4096 cases in total, Dashed lines in Figure 3-1). Staring from the candidate structures, we applied Bayesian network models by inferring the parameters given the structures and evaluating the models with regard to prediction accuracy and edge significance. We also extracted additional network structures from model consensus. The model evaluation was then iteratively continued until the final model with high predictive power and only significant edges is identified. In the following sections, we explain each modeling procedure in detail.

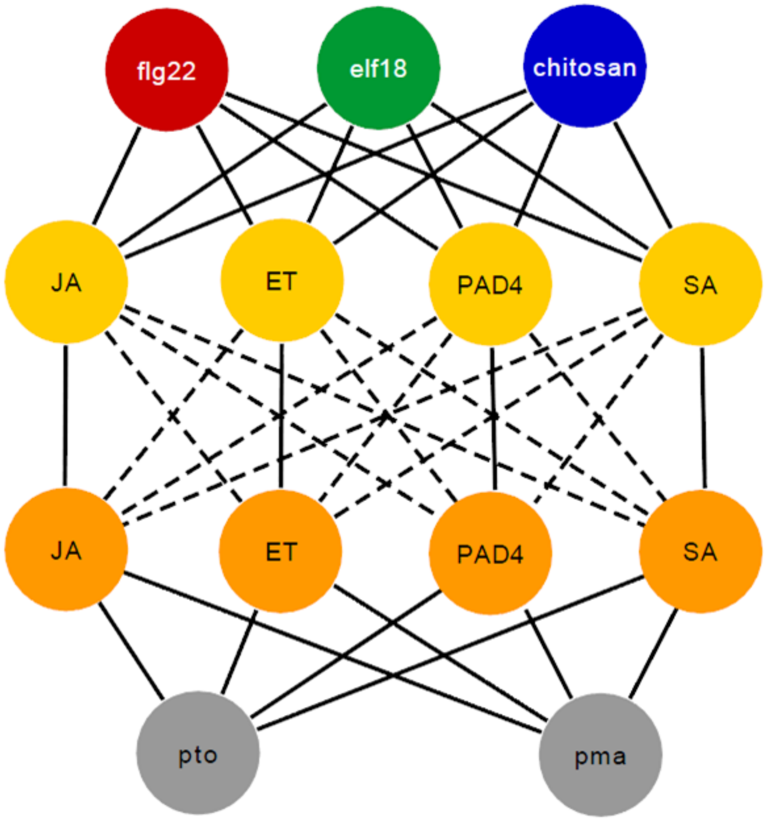


Figure 3-1 The configuration of a starting model. Red, green, and blue colored nodes in an input layer, the 1st layer from the top, represent flg22, elf18, and chitosan treatments, respectively. Yellow, orange, and gray colors describe an activation layer, a cross-talk layer, and an output layer, respectively. All variables (edges) are continuous in Elastic Net while all variables (nodes) except for two continuous variables (Gaussian) in the output layer are discrete in Bayesian network. Hidden structures between the activation and cross-talk layer are represented as dashed lines.

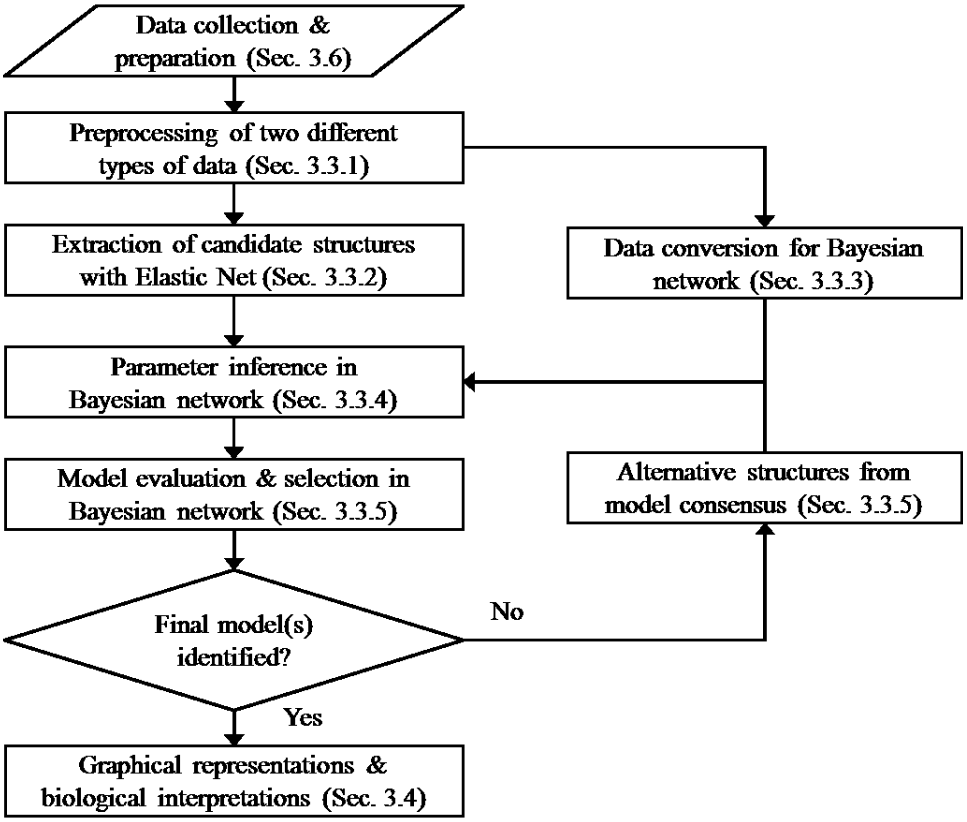

Data collection & preparation (Sec. 3.6)
Preprocessing of two different types of data (Sec. 3.3.1)
Extraction of candidate structures with Elastic Net (Sec. 3.3.2)
Data conversion for Bayesian network (Sec. 3.3.3)
Parameter inference in Bayesian network (Sec. 3.3.4)
Model evaluation & selection in Bayesian network (Sec. 3.3.5)
Alternative structures from model consensus (Sec. 3.3.5)
Final model(s) identified?
No
Yes
Graphical representations & biological interpretations (Sec. 3.4)

Figure 3-2 The flow chart of modeling procedures in a Bayesian network framework for obtaining a final immune signaling network.

3.3.1 Preprocessing of two different types of data

With mutant-adjusted values of mRNA expressions of the four sector markers after two-step normalizations (Section 3.6.1.3), the values of each of the markers were converted into activation probabilities using a linear mapping. For example, in the case of the JA sector, the minimum and maximum of the values of the JA sector marker across 64 different treatment:genotype:time conditions (4 treatments x 8 genotypes x 2 time points) were first computed and the values were mapped into activation probabilities using a linear mapping, which is defined as

$$p_{JA_t}^{act} = a \cdot s_{JA_t}^{exp} + b \qquad \text{Eq. 3-1}$$

where $p_{JA_t}^{act}$ is an activation probability of the JA sector marker at t time, $s_{JA_t}^{exp}$ is an adjusted expression value of the JA sector marker at t time, and a and b are the slope and intercept of the line, respectively (a = 0.173, b = -0.202). Each of lines (Eq. 3-1) for three other sectors is as follows: a = 0.184, b = -0.021 for ET sector; a = 0.174, b = 0.031 for PAD4 sector; a = 0.065, b = 0.092 for SA sector. Given an input, we observed considerable variations of mRNA expressions of four sectors at 3 hpt across different mutations, dependent of the states of other sectors. To simplify structures, we considered only wild-type at 3 hpt as an effective genotype for early states of the sector from the entire set of activation probabilities of each sector [112]. Note that all activation probabilities of each sector with genotypes including sector mutation were set to zero. Only the data from the three MAMP treatments (flg22, elf18, and chitosan) were used for modeling.

For the immunity levels, we chose the same number of replicates across all treatment:genotype:strain combinations (24 replicates in each treatment:genotype:strain condition). In any conditions with more than 24 replicates, we first filtered replicates based on z-scores, satisfying $z_i = \frac{y_i - \mu}{\sigma} \leq 2$ where μ and σ are a mean and a standard deviation of the original replicates in the treatment:genotype:strain condition, respectively and z_i is a z-score of y_i bacterial count. 24 replicates were then randomly chosen from the set that passed the filter.

3.3.2 Extraction of candidate structures with Elastic Net

To extract a set of candidate Bayesian network structures, Elastic Net [113] was applied. The Elastic Net structure was the same as that described above for the Bayesian network constraints except that the three MAMP nodes were used in the input layer instead: $V = \{V_{in}, V_{act}, V_{cross}, V_{out}\} = \{v_1, \dots, v_{13}\}$: $V_{in} = \{v_1, v_2, v_3\}$; $V_{act} = \{v_4, v_5, v_6, v_7\}$; $V_{cross} = \{v_8, v_9, v_{10}, v_{11}\}$, and $V_{out} = \{v_{12}, v_{13}\}$. The Elastic Net problem is defined as

$$\min_{(\beta_0, \boldsymbol{\beta}) \in \mathbb{R}^{p+1}} \frac{1}{2N} \sum_{i=1}^{N_E} (y_i - \beta_0 - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 + \lambda \left((1 - \alpha) \cdot \frac{1}{2} \|\boldsymbol{\beta}\|_{l_2}^2 + \alpha \|\boldsymbol{\beta}\|_{l_1} \right)$$

where N_E is the total number of instances for the Elastic Net, $\boldsymbol{\beta}$ is a $p \times 1$ vector with p parameters, $\boldsymbol{x}_i$ is a $p \times 1$ vector with p variables capturing variant network structures for an instance i of treatment:genotype:time:replicate or treatment:genotype:strain:replicate combinations, y_i is an actual measurement of the i instance (either an activation probability or a bacterial count), λ is an Elastic Net penalty factor, and α is a balancing factor controlling the compromise between ridge regression ($\alpha = 0$) and lasso regression ($\alpha = 1$) [113]. More specifically, $\boldsymbol{x}_i$ has 42 variables, representing states of all possible links: $x_{i,j}$ ($j \in \{1, \ldots, 12\}$), links between V_{in} and V_{act}; $x_{i,j}$ ($j \in \{13, \ldots, 28\}$), links between V_{act} and V_{cross}; $x_{i,j}$ ($j \in \{29, \ldots, 36\}$), links between V_{cross} and V_{out}; $x_{i,j}$ ($j \in \{37, \ldots, 42\}$), links between V_{in} and V_{out}. If y_i is either an activation probability at 3 hpt or a bacterial count, $\boldsymbol{x}_i$ is a binary vector containing 1 or 0 according to the i treatment:genotype condition. If y_i is an activation probability at 9 hpt, $\boldsymbol{x}_i$ is a continuous vector which contains $1/m$ (m = 4, one wild-type; m = 3, four singles; m = 2, six doubles; m = 1, four triples) or 0 for $x_{i,j}(j \in \{1, \ldots, 12\})$ and 1 or 0 for $x_{i,j}(j \in \{13, \ldots, 42\})$ according to the i treatment:genotype condition. To integrate the two different types of data which had different ranges, we rescaled all activation probabilities before solving the problem by multiplying by a scaling factor (1.403). This factor was derived by maximizing a correlation between original and rescaled instances.

To estimate the model parameters, a 6-fold cross-validation (CV) approach was used across the set of treatment:strain combinations [114]. All 6 datasets for cross-validation contained quadruple deletion such that the prediction accuracy of the six models, only affected by four sectors, was computed. For example, if we held out bacterial counts of the *pto* strain with 15 genotypes excluding quadruple deletion under flg22 treatment as test data, we fit the model with the rest: all activation probabilities of the four sectors in all conditions, bacterial counts with 16 genotypes in the other five treatment:strain conditions, and bacterial counts with quadruple deletion in flg22:pto. With an α ($\forall \alpha \in$ [0.01, 1]), we applied a coordinate descent algorithm for finding a set of appropriate λs to be searched ($\forall \lambda \in (0.005, 100)$) based on the entire data [113]. Given a setting α and λ, in each round we used 5 out of 6 folds of the data to fit the model and tested the model with the held-out data. With a selection of λ resulting in the smallest mean square test error from 6CV for a given α, we refit Elastic Net to obtain model structures and the parameters with the penalty constraints [113]. Among these fitted models with different levels of sparsity, we extracted 10 candidate structures ($\boldsymbol{G}_B^m$, $m \in \{1, \dots, 10\}$) retaining distinct sets of non-zero parameters for the inter-connectedness between four sectors as an initial set of model structures for the Bayesian network.

3.3.3 Data conversion for Bayesian network

The Bayesian network model has 11 network components: one ternary node (X_1) for three MAMPs as input signals; four binary nodes ($X_i, i \in \{2, \dots, 5\}$) at early states (3 hpt) and four binary nodes ($X_i, i \in \{6, \dots, 9\}$) at late states (9 hpt) describing dynamics of the four sectors; two continuous nodes ($X_i, i \in \{10, 11\}$) for two bacterial strains. The Bayesian network model was designed to capture the dependency between input signals (three MAMPs) and outcomes (immunity levels against two bacterial strains) only through the four sectors. Thus, given a MAMP treatment, remainder effects were estimated by taking the differences between two means of bacterial counts with quadruple mutations under mock and the MAMP treatment. In each treatment:genotype condition, we generated 24 instances, each of which consists of a set of states of each of the 11 components of the network as follows:

$$\begin{aligned}\boldsymbol{x} &= [x_1, \dots, x_{11}]\\ &= [x_{MAMP}^t, x_{JA_{3h}}^b, x_{ET_{3h}}^b, x_{PAD4_{3h}}^b, x_{SA_{3h}}^b, x_{JA_{9h}}^b, x_{ET_{9h}}^b, x_{PAD4_{9h}}^b, x_{SA_{9h}}^b, x_{pto}^c, x_{pma}^c]\end{aligned}$$

where x_{MAMP}^t is a ternary variable ($x_{MAMP}^t \in \{1,2,3\}$: 1, flg22; 2, elf18; 3, chitosan), $x_{s_t}^b$ is a binary variable representing an on or off state of *s* sector at time *t* ($x_{s_t}^b \in \{0,1\}$: 0, inactivation state or deletion; 1, activation state), x_p^c is a continuous variable representing the bacterial count of p strain ($x_p^c \in \mathbb{R}$). For example, an instance for chitosan:wildtype condition is described as

$$\boldsymbol{x} = [3, 1, 1, 1, 0, 1, 1, 1, 1, 4.08, 4.54].$$

To derive the binary states from the observed continuous data, we first generated the corresponding number of zeros and ones according to the activation probabilities (e.g. if $p_{s_t}^{act}$ = 0.25, the $x_{s_t}^b$ variables in 24 instances under the condition consist of 6 ones and 18 zeros). Within 24 instances under each condition, we randomly assigned the zeros and ones into the binary variables ($x_{i,j}, j \in \{2, \dots, 9\}$). In total, we generated a set of instances covering all data 100 times ($\boldsymbol{x}_i, i \in \{1, \dots, 100\}$) to prevent from any bias due to the discretization process.

3.3.4 Parameter inference in Bayesian networks

A set of n_B random variables ($\boldsymbol{X} = \{X_1, \dots, X_{n_B}\}$, $n_B = 11$) is a Bayesian network with respect to a directed acyclic graph ($\boldsymbol{G}_B$) if its joint probability density function can be written as

$$p(\boldsymbol{X}) = \prod_{i=1}^{n_B} p(X_i | X_j) \text{ for } j \in \{I[pa(X_i)]\}$$

where $pa(X_i)$ is a set of parents of X_i and $I[\,]$ is an index function, satisfying a local Markov property: each variable is conditionally independent of its non-descendants given its parent variables,

$$X_i \perp X_j | X_k \text{ for } j \in \{I[\boldsymbol{X} \backslash de(X_i)]\} \; and \; k \in \{I[pa(X_i)]\} \qquad \text{Eq. 3-2}$$

where $de(X_i)$ is a set of descendants of X_i [112].

To infer the parameters, conditional probabilities of all variables given a network structure $\boldsymbol{G}_B^m$, we first created a network with random parameters. The parameter sets for nine discrete nodes ($\boldsymbol{\theta}_{i,j}, j = \{1, \dots, 9\}$) were generated based on a likelihood equivalent uniform Bayesian Dirichlet prior [112] and the parameter sets for last two continuous nodes ($\boldsymbol{\theta}_{i,j}, j = \{10,11\}$) were generated from normal distributions. We then found the maximum likelihood estimates (MLEs) of the parameters,

$$\widetilde{\boldsymbol{\theta}}_{i,j}^m = \max_{\boldsymbol{\theta}_{i,j}} p(\boldsymbol{x}_{i,j} | \boldsymbol{G}_B^m, \boldsymbol{\theta}_{i,j}) = \max_{\boldsymbol{\theta}_{i,j}} \sum_{k=1}^{N_B} p(x_{i,j,k} | \boldsymbol{G}_B^m, \boldsymbol{\theta}_{i,j}) \qquad \text{Eq. 3-3}$$

where $\boldsymbol{x}_{i,j}$ is a set of instances of j random variable in an i set of instances, $\boldsymbol{\theta}_{i,j}$ is a conditional probability distribution of j random variable ($j \in \{1, \dots, 11\}$) with i set of instances ($i \in \{1, \dots, 100\}$), $\boldsymbol{G}_B^m$ is a given m structure ($m \in \{1, \dots, 10\}$), $\widetilde{\boldsymbol{\theta}}_{i,j}^m$ is the estimated $\boldsymbol{\theta}_{i,j}$, maximizing the posterior conditional probability distribution, $p(\boldsymbol{x}_{i,j} | \boldsymbol{G}_B^m, \boldsymbol{\theta}_{i,j})$, N_B is the total number of instances and $x_{i,j,k}$ is a k instance in $\boldsymbol{x}_{i,j}$

($k \in \{1, \ldots, N_B\}$). In our case, we fully observed the states of all nodes such that the maximum likelihood estimates were counts of the number of instances with the given combination of states. Note that in case of a zero instance due to perturbation, we updated the parameters only for the descendent nodes of the perturbed node [112].

3.3.5 Evaluation criteria in Bayesian network models

We evaluated each Bayesian network model with the estimated parameters ($\widetilde{\boldsymbol{\theta}}_i^m$) given the structure ($\boldsymbol{G}_B^m$) in five different ways:

1. calculating a log-likelihood (ll_i, Eq. 3-4),
2. using Bayesian information criteria (BIC_i, Eq. 3-5),
3. calculating two Spearman's rank correlation coefficients to measure both a training set accuracy, r_s(3CV, trng, activation probabilities), and a test set accuracy, r_s(3CV, test, activation probabilities), of predicting activation probabilities of the four sectors with 3CV across the three different MAMPs,
4. calculating two Spearman's rank correlation coefficients to measure both a training set accuracy, r_s(6CV, trng, immunity levels), and a test set accuracy, r_s(6CV, test, immunity levels), of predicting log-transformed bacterial counts of two strains with 6-fold cross validation across the six different MAMP:strain combinations, and
5. computing the statistical significance of conditional dependencies for links between any two sector nodes (Eq. 3-6).

Each of the evaluation methods is described in detail below.

The log-likelihood with $\boldsymbol{x}_i$ instances was defined as

$$ll_i = \log \prod_{k=1}^{N_B} p\left(\boldsymbol{x}_{i,k} | \boldsymbol{G}_B^m, \widetilde{\boldsymbol{\theta}}_i\right) = \sum_{j=1}^{n} \sum_{k=1}^{N_B} \log p\left(x_{i,j,k} | \boldsymbol{x}_{i,pa(j),k}, \widetilde{\boldsymbol{\theta}}_{i,j}\right) \qquad \text{Eq. 3-4}$$

where $\boldsymbol{x}_{i,pa(j),k}$ is a set of instances of all the parent nodes of j random variable in an i set of instances.

Given the log-likelihood with $\boldsymbol{x}_i$ instances, the BIC_i was calculated as

$$BIC_i = \frac{1}{N_B}\sum_{k=1}^{N_B} \log p\left(\boldsymbol{x}_{i,k} | \boldsymbol{G}_B^m, \widetilde{\boldsymbol{\theta}}_i\right) - \frac{n_p^m}{2}\log N_B \qquad \text{Eq. 3-5}$$

where n_p^m is the total number of estimated parameters given the structure $\boldsymbol{G}_B^m$.

To explore the prediction accuracy of the activation probabilities of the four sectors with each Bayesian network model given $\boldsymbol{G}_B^m$ and $\boldsymbol{x}_i$, we used 3 CV by splitting entire instances into three parts based on the MAMPs,

$$\boldsymbol{x}_i = [\boldsymbol{x}_{i,flg22}, \boldsymbol{x}_{i,elf18}, \boldsymbol{x}_{i,chitosan}]^T.$$

For the *k* part (held-out test data, $k \in \{1,2,3\}$), we created a network with random parameters (Section 3.3.4), estimated the model parameters with the other k-1 parts of the instances (training data) by MLEs (Eq. 3-3), and predicted both the training and test data. Spearman's rank correlations for the training error and the test error were then calculated between observed and predicted instances. For the prediction accuracy of the log-transformed bacterial counts against two bacterial strains, the same procedures as above were executed to calculate two Spearman's rank correlations with 6CV in terms of MAMP:strain combinations,

$$\boldsymbol{x}_i = [\boldsymbol{x}_{i,flg22:pto}, \boldsymbol{x}_{i,elf18:pto}, \boldsymbol{x}_{i,chitosan:pto}, \boldsymbol{x}_{i,flg22:pma}, \boldsymbol{x}_{i,elf18:pma}, \boldsymbol{x}_{i,chitosan:pma}]^T.$$

Note that we computed median values for all four of the metrics explained above.

We also applied a Wilcoxon signed rank test to measure the statistical significance of the conditional dependencies among sectors [115]. More specifically, with 100 different measured weights ($\widetilde{\boldsymbol{\omega}}_{a,b}$) of the cross-talk regulation between *a* and *b* in 100 different models based on $\boldsymbol{x}_i$($i \in \{1, \dots ,100\}$), we performed a two-sided rank test of the hypothesis that the values in the vector come from a distribution whose median is zero. To do so, a test statistic $W,$ denoted as

$$W = \left| \sum_{i=1}^{n_{rp}} \left[sgn\left(\widetilde{\omega}_{2,i} - \widetilde{\omega}_{1,i}\right) \cdot R_i \right] \right| \qquad \text{Eq. 3-6}$$

where n_{rp} is the number of pairs, *sgn* is a sign function, $\widetilde{\omega}_{1,i}$ and $\widetilde{\omega}_{2,i}$ are any paired measurements of the weights, and R_i is the rank of *i* pair, was calculated to obtain a p-value. If the p-value is smaller, it is more likely that the hypothesis is rejected. In our case, if p-value $< 10^{-4}$, we denoted that the conditional dependence between a node and b node is significant with the sign.

Although all the 5 different criteria provide useful information for evaluating the network models, they assess the modeling performance in slightly different ways. Given each obtained model, the log-likelihood (Eq. 3-4) and two Spearman correlation coefficients focus on the accuracy of the parameter inference. The BIC (Eq. 3-5) captures both the precision of the estimated parameter (log-likelihood, Eq. 3-4) and the complexity of the model by additionally considering the number of the parameters. On the contrary, given the distributions of the estimated parameters (see Section 3.3.3), the confidence of conditional dependencies relating to links in the cross-talk layer is statistically assessed by a Wilcoxon signed rank test (Eq. 3-6). Thus, the edge significance is crucial for the selection of a final model from several candidate models if they perform similarly. In the next section, we show that many candidate models are indistinguishable from each other regarding their predictive power and the statistical evaluation on their structures plays a pivotal role for selecting our final model.

3.3.6 Selection criteria for fining final model(s) in Bayesian networks

Starting with 10 structures ($\boldsymbol{G}_B^{m_1},\ m_1 \in \{1, \dots, 10\}$) from the Elastic Net, we evaluated the parameters and the structures of each of the models in five different validation criteria (Section 3.3.5). Note that selected Bayesian network models with high prediction accuracy and reasonable complexity (high ll, high r_s(CV, trng) and r_s(CV, test) for both activation probabilities of four sectors and bacterial counts against two strains, and high BIC) shared most of significant links for cross-talks among four sectors but also had distinct links, some of which might be insignificant. Four models shown in Figure 3-3, for example, were first selected by comparing both their predictive power and model complexity among the candidate models. Most of significant links in one model also appeared in other models as confident links. However, there were insignificant links in every model ($x_3 - x_9$ and $x_5 - x_6$) and a significant edge in only one model ($x_4 - x_6$ in $\boldsymbol{G}_B^4$). The confidence of the ambiguous links might depend on the originated structures such that we further investigated 8 structures (possible combination of the links) in a similar fashion ($\boldsymbol{G}_B^{m_2}, m_2 \in \{1, \dots, 8\}$). As shown in Table 3-1, all the edges shown in the final model (Figure 3-4) were significant, meaning that the structure is likely to be real and robust against artificial noise from the experimental setup. Admittedly, we also found another network with all significant edges and similar performance (last model in Table 3-1). However, the network structure in this model is almost same as the structure in the final model except for the absence of one edge (from JA at 3 hpt to ET at 9 hpt).

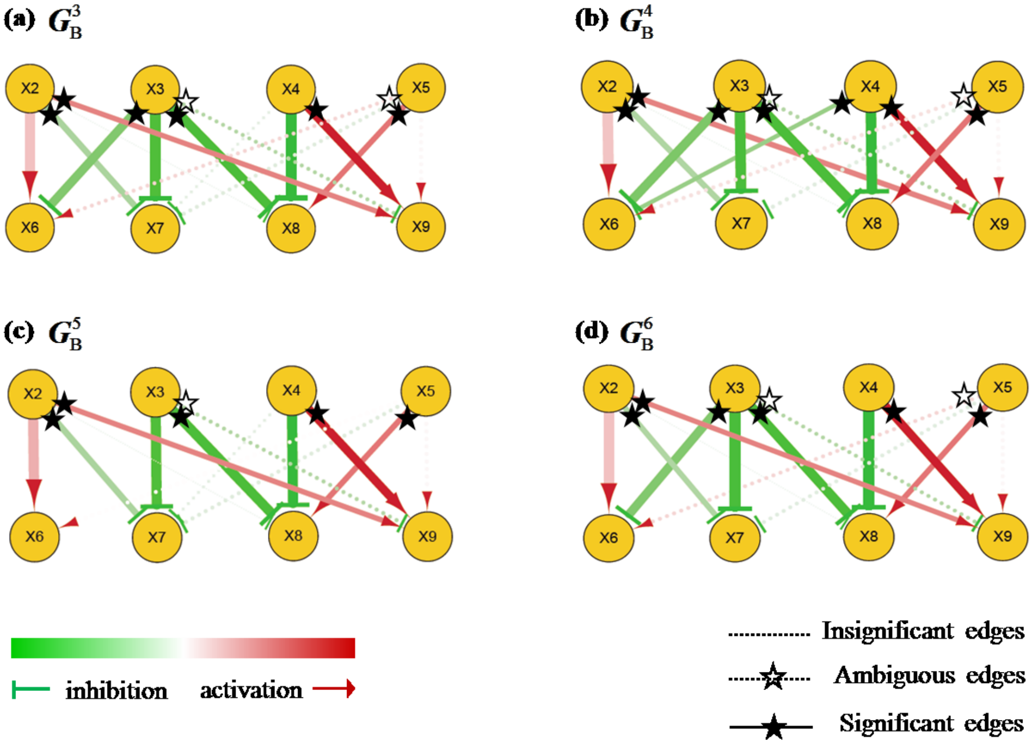


Figure 3-3 The four illustrative examples of selected models for model consensus. These models are shown in Table 3-1 with bold text for $\boldsymbol{G}_B^{m_1}$. Applying the consensus of these model, we further extracted 8 extra candidate structures (all possible combination of three ambiguous edges, $\boldsymbol{G}_B^{m_2}, m_2 \in \{1, \dots, 8\}$) to find a best model, shown in Table 3-1 with bold text for $\boldsymbol{G}_B^{m_2}$.

Although the edge had relatively weak compared to other edges, it was still significant. Thus, we concluded that it is reasonable to add the edge into our final model. Note that the structural uncertainty with the initial candidate models from the Elastic Net originated from different modeling frameworks between the linear regression model and the Bayesian network model. It can be generalizable as an iterative procedure for the model selection approach. The final Bayesian network model selected by the iterative approach is shown in Figure 3-4. All labeled values in the links are medians among 100 values from the models with $\boldsymbol{x}_i, i \in \{1, \dots, 100\}$. The size of nodes of the four sectors at both 3 hpt and 9 hpt is proportional to their marginal activation probabilities, $\tilde{p}_{s_t}^{act} = p(s_t = 1)$.

Table 3-1 The obtained values of metrics for evaluating Bayesian network models. $\boldsymbol{G}_B^{m_1}$ was a set of the candidate model structures from Elastic net approaches and $\boldsymbol{G}_B^{m_2}$ was a set of the selected model structures from the iterative approaches by model consensus. A subset of $\boldsymbol{G}_B^{m_1}$ with bold text was the selected models for model consensus and a $\boldsymbol{G}_B^{m_2}$ with bold text was the final Bayesian network model shown in Figure 3-4. '% sig.' in the last column described how many links were significant in terms of a Wilcoxon signed rank test [115] in percentage.

	$\overline{ll}$	$\overline{BIC}$	$\overline{r_{s,3CV}^{trng}}$	$\overline{r_{s,3CV}^{test}}$	$\overline{r_{s,6CV}^{trng}}$	$\overline{r_{s,6CV}^{test}}$	# links (% sig.)
$\boldsymbol{G}_B^{m_1}, m_1 \in \{1, \dots, 10\}$	-6513	-7013	.604	.280	.702	.503	12(58%)
	-6519	-6991	.599	.281	.703	.501	11(55%)
	-6518	**-6990**	**.603**	**.280**	**.702**	**.502**	**11(55%)**
	-6516	**-6988**	**.606**	**.280**	**.703**	**.504**	**11(64%)**
	-6521	**-6979**	**.597**	**.281**	**.702**	**.5**	**10(50%)**
	-6519	**-6963**	**.604**	**.281**	**.702**	**.503**	**10(60%)**
	-6523	-6953	.590	.268	.700	.498	9(56%)
	-6528	-6929	.579	.259	.696	.505	8(63%)
	-6531	-6905	.579	.257	.696	.503	7(57%)
	-6534	-6894	.570	.249	.694	.5	6(50%)
$\boldsymbol{G}_B^{m_2}, m_2 \in \{1, \dots, 8\}$	-6521	-6951	.606	.282	.703	.504	9(78%)
	-6524	-6926	.606	.282	.703	.505	8(88%)
	-6524	-6926	.604	.279	.702	.502	8(75%)
	-6524	-6926	.605	.288	.703	.504	8(88%)
	-6526	-6914	.605	.279	.702	.504	7(86%)
	-6528	**-6901**	**.606**	**.291**	**.703**	**.504**	**7(100%)**
	-6527	-6901	.604	.288	.702	.502	7(86%)
	-6529	-6889	.603	.288	.703	.503	6(100%)

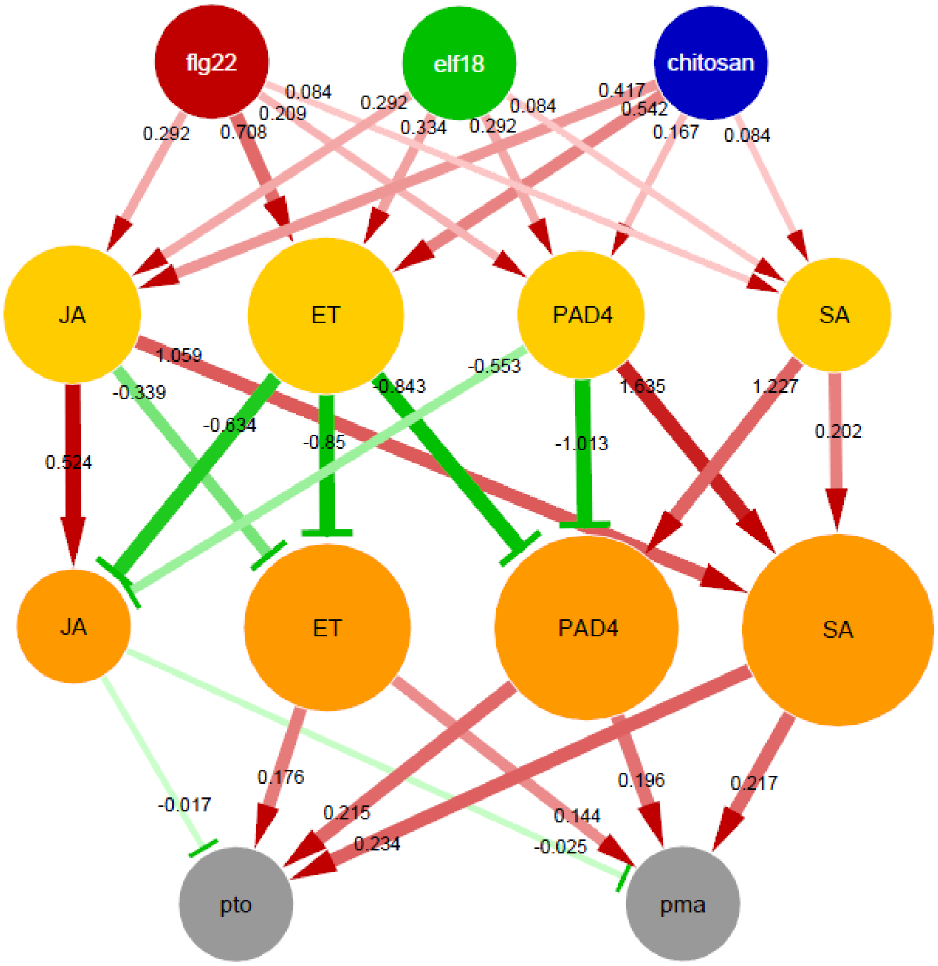


Figure 3-4 A model structure obtained by a Bayesian network approach. Red and green directed links represent activation and inhibition. The size of sector nodes represent the overall activation probabilities, calculated from the marginal probabilities regardless of the states of directly connected upstream nodes. The width and the color intensity of a link are proportional to the link parameter value. All the values assigned to links were calculated with the corresponding metrics in Section 3.4.1.

In summary, we first used the linear regression model to choose the initial structures. Staring from the candidate models in Bayesian frameworks, we iteratively evaluated the accuracy of estimated parameters and the significance of model structures and expanded the search space with model consensus approach. Throughout the iterative approach for the model evaluation and selection, we assessed candidate models based on five criteria explained above. In terms of the prediction accuracy given log-likelihood and correlation between observed and predicted values, the candidate models are indistinguishable from each other since the candidate models performed similarly. Although BIC value captures both the prediction accuracy and the model complexity, the criterion is also limited by Gaussian approximation [116]. The significance of links from the Wilcoxon signed rank test, on the other hand, provides clear information on the separation of good models from less accurate models. In general, it is likely that we can find several models with different structures but similar accuracy. In our model selection approach, by incorporating the information on structural stability, we successfully found a reliable model capturing the dynamics of the four sectors and predicting the immune responses reasonably well.

3.4 Graphical model representations and biological interpretation in Bayesian network model

In Section 3.3, we implemented five different criteria to evaluate the estimation accuracy of the parameters and structures and successfully selected a final Bayesian network model with high test accuracy and highly confident links. The structure of a Bayesian network model only describes the conditional dependency between two connected nodes. To biologically interpret the relative conditional effects on downstream nodes from connecting upstream nodes, several different types of regulatory effects are defined and used for visualizing the final network model (Figure 3-4). The interaction types between two sectors regulating a common sector are inferred from the conditional probabilities. These relationships between sectors are useful not only for visualizing the immune signaling network with cross-talk but for understanding the roles of the four sectors in different time and context.

3.4.1 Relative conditional dependencies

To explore strengths of conditional effects between two connected nodes, we defined several different types of regulatory effects related to links (Figure 3-4), which could be calculated directly from the learned Bayesian network models:

1. For the links between X_1 and X_i $(i \in \{2, \dots, 5\})$, the strength of the signaling activation of s_1 sector at 3 hpt under *t* MAMP treatment can be calculated as

$$\widetilde{\omega}_{MAMP, s_{1,3h}} = p\left(x^b_{s_{1,3h}} = 1 | x^t_{MAMP} = t, \boldsymbol{G}^m_B, \widetilde{\boldsymbol{\theta}}^m_i\right),$$

where $\widetilde{\omega}_{a,b}$ is a measured weight of the link between *a* and *b*.

2. For the links between X_i and X_{i+4} $(i \in \{2, \dots, 5\})$, the strength of the direct regulation of the $s_1$sector from 3 hpt to 9 hpt can be measured as

$$\widetilde{\omega}^i_{s_{1,3h}, s_{1,9h}} = \log_2\left(\frac{p\left(x^b_{s_{1,9h}} = 1 \middle| x^b_{s_{1,3h}} = 1,\ x^b_{s_{i,3h}} = 0, \boldsymbol{G}^m_B, \widetilde{\boldsymbol{\theta}}^m_i\right)}{p\left(x^b_{s_{1,9h}} = 0 \middle| x^b_{s_{1,3h}} = 1,\ x^b_{s_{i,3h}} = 0, \boldsymbol{G}^m_B, \widetilde{\boldsymbol{\theta}}^m_i\right)}\right),$$

$$s_{i,3h} \in pa\left(s_{1,9h}\right) \backslash s_{1,3h} \text{ if } \left|pa\left(s_{1,9h}\right)\right| > 1,$$

where $\widetilde{\omega}^i_{a,b}$ is a measured weight of the link between *a* and *b* based on $\boldsymbol{x}_i$.

3. For the links between X_i $(i \in \{2, \dots, 5\})$ and X_j $(j \in \{6, \dots, 9\})$, the strength of the cross-talk regulation of s_1 sector at 9 hpt from s_2 sector at 3 hpt can be measured as

$$\widetilde{\omega}^i_{s_{2,3h}, s_{1,9h}} = \log_2\left(\frac{p\left(x^b_{s_{1,9h}} = 1 \middle| x^b_{s_{1,3h}} = 1,\ x^b_{s_{2,3h}} = 1,\ x^b_{s_{i,3h}} = 0, \boldsymbol{G}^m_B, \widetilde{\boldsymbol{\theta}}^m_i\right)}{p\left(x^b_{s_{1,9h}} = 0 \middle| x^b_{s_{1,3h}} = 1,\ x^b_{s_{2,3h}} = 1,\ x^b_{s_{i,3h}} = 0, \boldsymbol{G}^m_B, \widetilde{\boldsymbol{\theta}}^m_i\right)}\right)$$
$$- \log_2\left(\frac{p\left(x^b_{s_{1,9h}} = 1 \middle| x^b_{s_{1,3h}} = 1,\ x^b_{s_{2,3h}} = 0,\ x^b_{s_{i,3h}} = 0, \boldsymbol{G}^m_B, \widetilde{\boldsymbol{\theta}}^m_i\right)}{p\left(x^b_{s_{1,9h}} = 0 \middle| x^b_{s_{1,3h}} = 1,\ x^b_{s_{2,3h}} = 0,\ x^b_{s_{i,3h}} = 0, \boldsymbol{G}^m_B, \widetilde{\boldsymbol{\theta}}^m_i\right)}\right),$$
$$s_{i,3h} \in pa\left(s_{1,9h}\right) \backslash s_{1,3h} \text{ if } \left|pa\left(s_{1,9h}\right)\right| > 1.$$

4. For the links between X_i $(i \in \{2, \dots, 5\})$ and X_j $(j \in \{6, \dots, 9\})$, if X_j has more than two connected nodes besides direct regulation (X_i, $i \in \{2, \dots, 5\}$, $j \neq i + 4$), the type of regulatory effects (either cooparativity or redundancy) from s_2 and s_3 sectors at 3 hpt on s_1 sector at 9 hpt can be inferred from

$$\widetilde{I}^{s_{1,9h}}_{s_{2,3h}, s_{3,3h}} = \log_2\left(\frac{p\left(x^b_{s_{1,9h}} = 1 \middle| x^b_{s_{1,3h}} = 1,\ x^b_{s_{2,3h}} = 1,\ x^b_{s_{3,3h}} = 1, \boldsymbol{G}^m_B, \widetilde{\boldsymbol{\theta}}^m_i\right)}{p\left(x^b_{s_{1,9h}} = 0 \middle| x^b_{s_{1,3h}} = 1,\ x^b_{s_{2,3h}} = 1,\ x^b_{s_{3,3h}} = 1, \boldsymbol{G}^m_B, \widetilde{\boldsymbol{\theta}}^m_i\right)}\right)$$
$$-\log_2\left(\frac{p\left(x^b_{s_{1,9h}} = 1 \middle| x^b_{s_{1,3h}} = 1,\ x^b_{s_{2,3h}} = 1,\ x^b_{s_{3,3h}} = 0, \boldsymbol{G}^m_B, \widetilde{\boldsymbol{\theta}}^m_i\right)}{p\left(x^b_{s_{1,9h}} = 0 \middle| x^b_{s_{1,3h}} = 1,\ x^b_{s_{2,3h}} = 1,\ x^b_{s_{3,3h}} = 0, \boldsymbol{G}^m_B, \widetilde{\boldsymbol{\theta}}^m_i\right)}\right)$$

$$-\log_2\left(\frac{p\left(x^b_{s_{1,9h}}=1\middle|x^b_{s_{1,3h}}=1,\ x^b_{s_{2,3h}}=0,\ x^b_{s_{i,3h}}=1,\boldsymbol{G}^m_B,\widetilde{\boldsymbol{\theta}}^m_i\right)}{p\left(x^b_{s_{1,9h}}=0\middle|x^b_{s_{1,3h}}=1,\ x^b_{s_{2,3h}}=0,\ x^b_{s_{i,3h}}=1,\boldsymbol{G}^m_B,\widetilde{\boldsymbol{\theta}}^m_i\right)}\right)$$

$$+\log_2\left(\frac{p\left(x^b_{s_{1,9h}}=1\middle|x^b_{s_{1,3h}}=1,\ x^b_{s_{2,3h}}=0,\ x^b_{s_{i,3h}}=0,\boldsymbol{G}^m_B,\widetilde{\boldsymbol{\theta}}^m_i\right)}{p\left(x^b_{s_{1,9h}}=0\middle|x^b_{s_{1,3h}}=1,\ x^b_{s_{2,3h}}=0,\ x^b_{s_{i,3h}}=0,\boldsymbol{G}^m_B,\widetilde{\boldsymbol{\theta}}^m_i\right)}\right),$$

$$s_{i,3h}\in pa(s_{1,9h})\backslash s_{1,3h}\text{ if }\left|pa(s_{1,9h})\right|>1.$$

If $\tilde{I}^{s_{1,9h}}_{s_{2,3h},s_{3,3h}} \ll 0$, s_2 and s_3 sectors are cooperative for regulating s_1 sector. If $\tilde{I}^{s_{1,9h}}_{s_{2,3h},s_{3,3h}} \gg 0$, s_2 and s_3 sectors are redundant for regulating s_1 sector. If $\tilde{I}^{s_{1,9h}}_{s_{2,3h},s_{3,3h}} \approx 0$, s_2 and s_3 sectors have simply additive effects on the regulation of s_1 sector.

5. For the links between X_i $(i \in \{6,\ldots,9\})$ and X_j $(j \in \{10,11\})$, the strength of the relative contributions of $s_1$sector to the immunity level against bacterial strain p can be measured as

$$\widetilde{\omega}_{s_{1,9h},p}=-\log_2\left(\frac{\widetilde{m}_p\text{ given }x^b_{s_{1,9h}}=1,\ x^b_{s_{i,9h}}=0,\boldsymbol{G}^m_B}{\widetilde{m}_p\text{ given }x^b_{s_{1,9h}}=0,\ x^b_{s_{i,9h}}=0,\boldsymbol{G}^m_B}\right),i\in\{2,3,4\},$$

where $\widetilde{m}_p$ is an estimated mean of X_i $(i \in \{n-1,n\})$ based on the condition.

3.4.2 Biological interpretation of the Bayesian network model

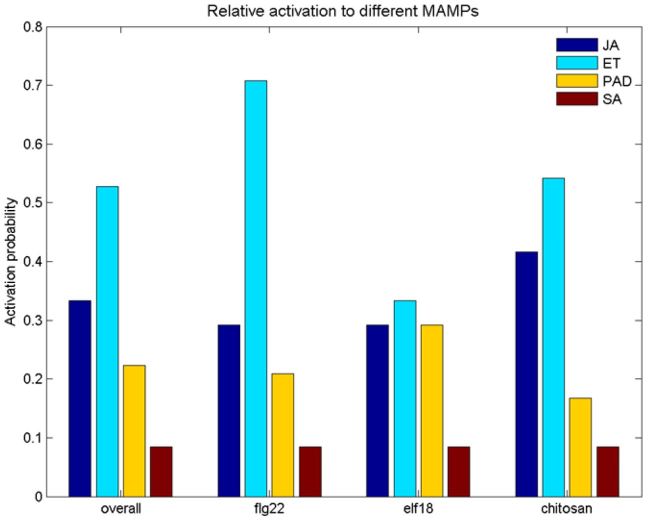


Figure 3-5 The activation probabilities of four signaling sectors. An overall activation probability of individual sectors was calculated from the marginal probability of the sector to be activated

regardless of the input signals. All the activation probabilities of the four sectors in specific MAMPs are based on the defined regulatory activation, shown in Section 0.

In the Bayesian network model, an activation layer describes how the activation of the four signaling sectors is induced by three MAMP treatments. In Figure 3-5, the overall activation probabilities of ET and JA sectors are relatively higher than PAD4 and SA sectors, suggesting that ET and JA sectors are required to be highly activated at the early stages of signal transduction for controlling the immune signaling at the later stage. As shown in Figure 3-4, the ET sector has a principal role in an inhibitory cross-talk and governs the activity levels of JA and PAD4 sectors at the later stage. The activation probability of the SA sector, on the other hand, is relatively lower, meaning that the SA sector is not a main contributor to the signal transmission at the early stage. If we consider differential activations from three MAMP treatments, ET and JA sectors probably have more important roles with high activation in flg22 and chitosan treatment, respectively. In elf18, JA, ET, and PAD4 sectors are similarly activated in a low level.

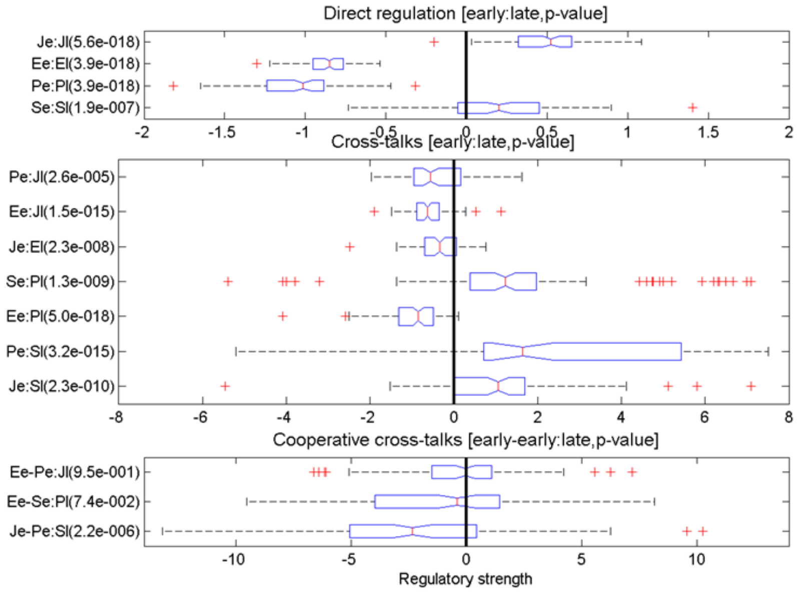


Figure 3-6 Relative strength of cross-talk relationships between two connected sector nodes and the regulatory interactions between two sectors at 3 hpt connected to the sector of interest. P-values were calculated from the Wilcoxon signed rank sign test.

In the cross-talk layer, the relative strength of self-regulations (direct regulation in Figure 3-6) and cross-talk relationships between two connected sector nodes (cross-talks in Figure 3-6) are calculated based on the estimated conditional probabilities in the final Bayesian network model. JA and SA sectors have self-activation while ET and PAD4 sectors appear to have self-inhibitory relationship. Based on the cross-talk relationships, there exists clear distinction between the two sector groups: cross-inhibition between JA and ET sector and cross-activation between PAD4 and SA sector. We also inferred the types of interactions between two upstream sectors at 3 hpt, which connected to the downstream sector at 9 hpt for regulations (cooperative cross-talks in Figure 3-6). Given this interaction information, PAD4 and JA sectors are likely to have compensatory roles to activate SA sector. ET and PAD4 sectors for JA sector and ET and SA sectors for PAD4 sector are rather independent each other. At the late stage of the signal transduction, JA and ET sectors are repressed through cross-talk regulation. On the other hand, SA sector is highly activated by either JA or PAD4. The activation of PAD4 sector is balanced at the relevant level via the opposite regulations from ET and SA. Altogether, through highly interactive communications among the four sectors, SA sector plays a more critical role than JA during late signal transmission.

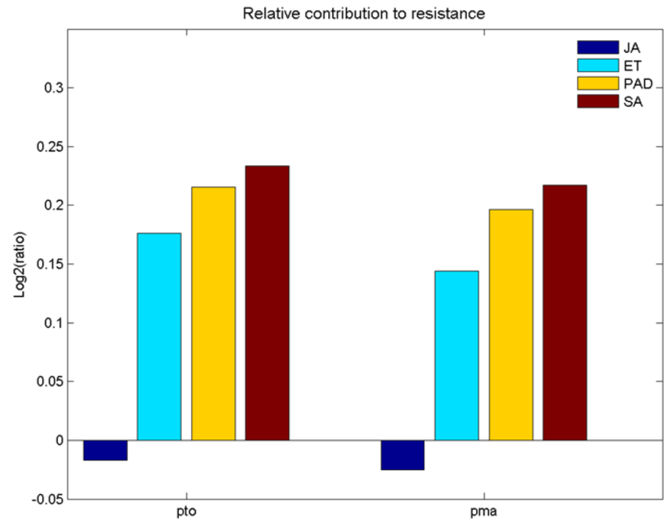


Figure 3-7 The relative contribution of four sectors to the resistance against two bacterial strains.

By using a log-ratio, we calculated the relative effects of the four sectors on immunity levels against two bacterial strains (Figure 3-7). In general, we observed that ET, PAD4, and SA sectors strongly contribute to the resistance and JA sector may not be involved in direct contribution to the immune response. In all, at the initial recognition of the signaling inputs (MAMP treatments), the plant PTI signaling network is mainly controlled by ET sector. The signals are transmitted with the intensive cross-talk relations among sectors, governing dynamics of the immune signaling network. At the late stage, PAD4 and SA sectors have significant roles and directly contribute to plant immunity.

3.5 Conclusion

Modeling the immune signaling network based on a Bayesian network produced a relatively simple and interpretable model with reasonable prediction accuracy (Figure 3-4 and Table 3-1). However, there are several limitations of predicting observed data across all possible experimental cases based on the current Bayesian network model setting. First of all, only one genotype (wild-type) was considered for modeling activities of the four sectors at the early stages due to the Markov property assumed by the Bayesian network. Based on all the measurements of sector activities, we observed that the activity values for some of the four sectors increased even before 3 hpt and varied across different genotypes. The conventional Bayesian network approach did not allow cyclic edges such that the information on relatively high interactions among sectors at very initial stage could not be modeled with the current approach [117]. Second, each of the continuous sector activity values was mapped into a set of binary values to obtain a corresponding activation probability used for modeling. This data conversion caused discretization errors so that there existed loss of quantitative information across different cases. Furthermore, it was assumed that only a single structure capturing cross-communications among four sectors could explain all differential dynamics of the four sectors over different conditions. However, it is also possible that there are treatment-specific interactions or mechanisms which cannot be easily captured by the current network model. This structural assumption should be explored with further analyses. Besides these, the modeling formulation based on the Bayesian network may not be the ideal approach for this problem in that there were unnecessary parameters to be estimated but not used for the interpretations (e.g., conditional probabilities of two output nodes given many of parents' states) and further calculations were required for interpreting the connections between nodes. These challenges motivate us to pursue another modeling strategy, a multiple linear regression approach, which is introduced in the next chapter.

3.6 Data collection and preparation

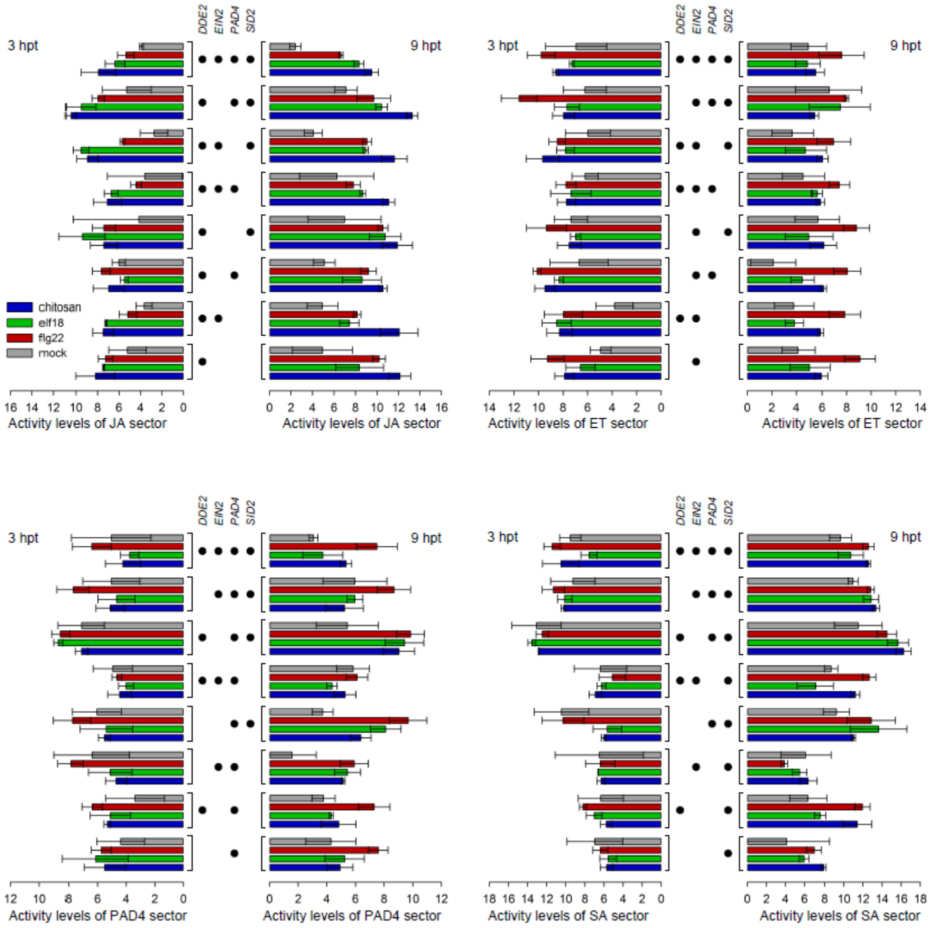


Figure 3-8 Measurements of the sector activities of the four sector marker genes at two different time points with 8 effective combinatorial genotypes under 4 different treatments.

Two types of data were collected: (1) the mRNA level of marker genes for each of four signaling sectors as a proxy of the sector activity (Figure 3-8); (2) apoplastic growth of

Pto and *P. syringae* pv. *maculicola* ES4326 (*Pma*) (Figure 3-9). The mRNA level and bacterial count measurements were performed after treatment with mock, flg22, elf18, or chitosan (a modified chitin [118]) in 16 combinatorial genotypes of Arabidopsis. The mRNA levels of the marker genes were measured 3 and 9 hpt. The bacterial strains were infiltrated into the leaf 24 hpt, and the bacterial counts were measured two days after infiltration. The $\log_2$-transformed mRNA level values were non-linearly scaled to normalize the variations of biological replicates across marker genes. These preprocessed mRNA level values were defined as the sector activity values and used in modeling. On the other hand, the $\log_{10}$-transformed bacterial counts were used in modeling and the decrease in the bacterial growth with MAMP vs. mock treatment of the same genotype was defined as the immunity level.

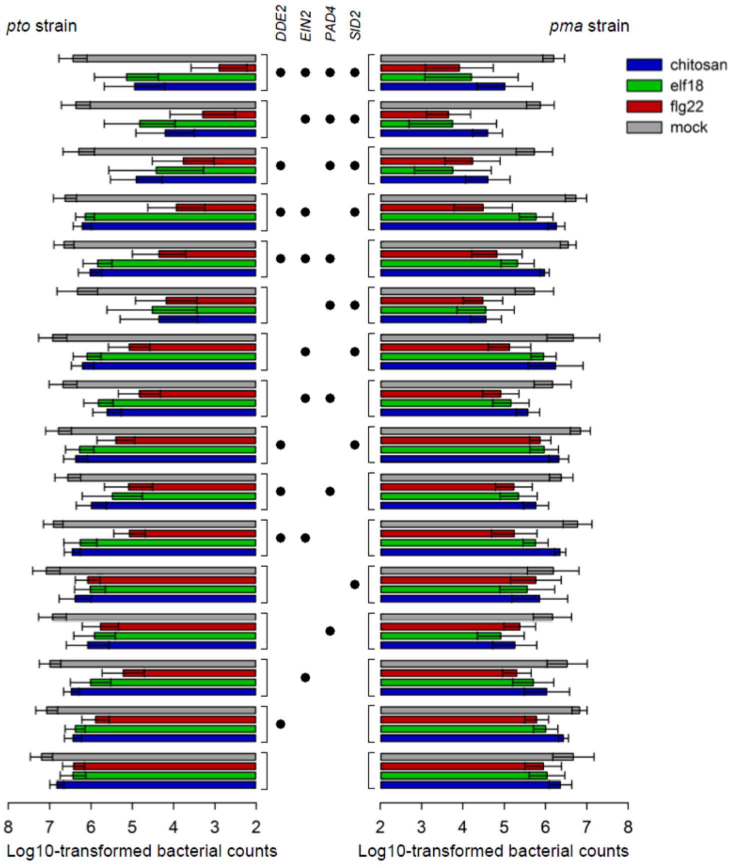


Figure 3-9 The measurements of log10-transformed counts against two bacterial strains (*Pto* and *Pma*) with 16 different combinatorial genotypes under 4 different treatments.

3.6.1 Activity levels of four sectors based on mRNA expressions of corresponding sector markers

3.6.1.1 Selection of the sector marker genes

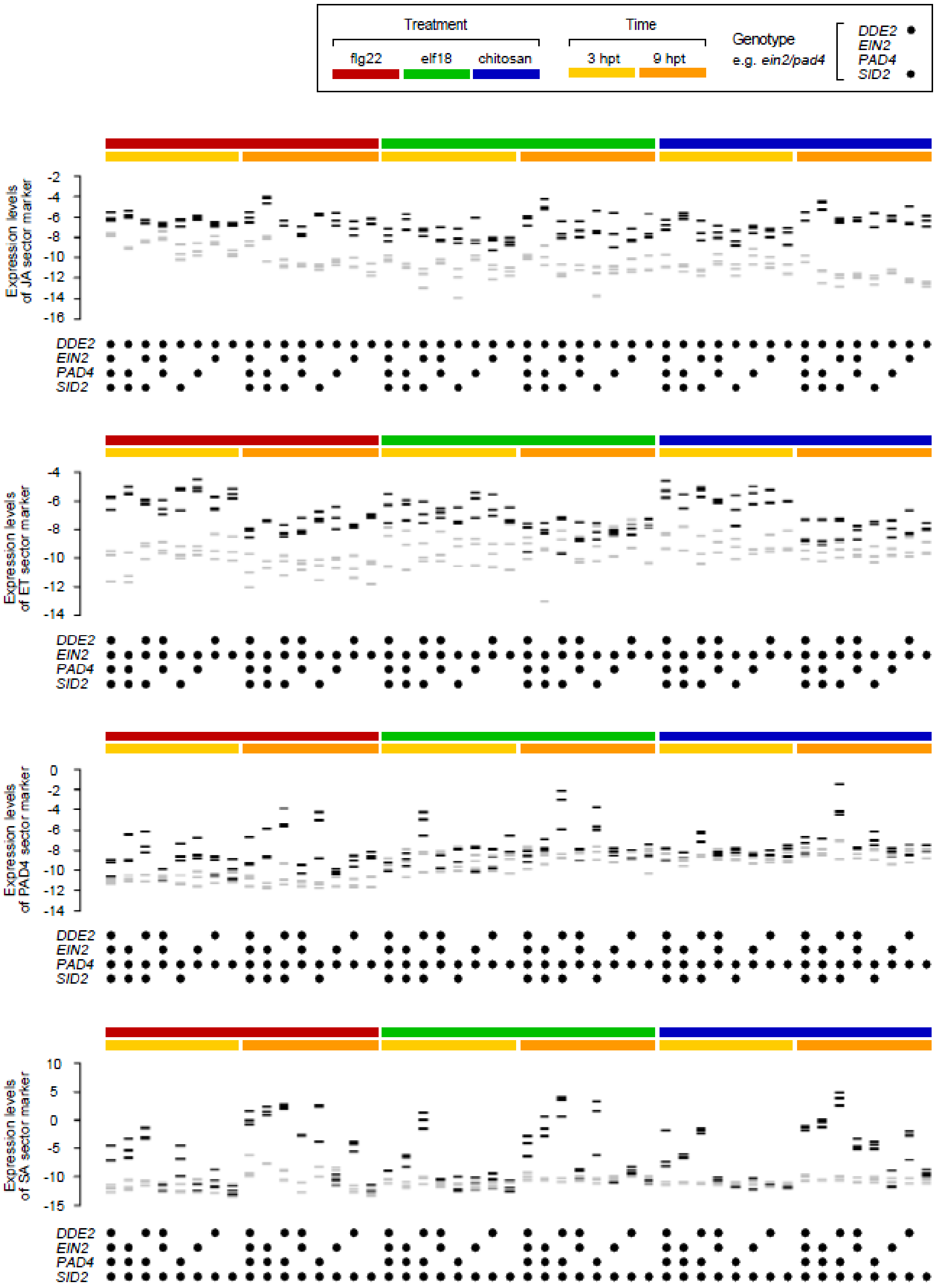


Figure 3-10 Sector marker gene expression levels. Each plot shows the expression levels, after between-samples normalization before mutant-adjustment, of the indicated sector marker gene.

The wild-type alleles in the genotype are shown by black dots. Each bar represents an observation: black, genotype containing the wild-type allele for the sector; gray, genotype containing the mutant allele for the sector. See key at top for the color-codes for the treatment and the time.

The mRNA levels of the marker genes, At3g50280, At2g41230, At5g46960, and At2g14610 (*PR1*), were used as the proxies of for the JA, ET, PAD4, and SA sector activities, respectively. The datasets used for the marker gene selection consisted of mRNA profile data that were available publicly. The mRNA profile data used were generated by DNA microarrays, mainly Affimetrix ATH1 Arabidopsis whole genome array. The marker genes were selected based on four criteria: (1) their mRNA levels increased during some pathogen interactions, (2) their induced mRNA levels were almost completely dependent on the genes that were known to be essential for the respective sector functions (such as *DDE2*, *EIN2*, *PAD4*, and *SID2* for the respective sectors), (3) their mRNA levels increased with exogenous application of the respective hormone molecules, and (4) the effects of the respective hormone molecules are highly specific. Criteria (3) and (4) were not applicable for the PAD4 sector marker selection, as no signaling molecule that directly activates the PAD4 sector in an SA-independent manner is known. Thus, in the case of the PAD4 sector marker, in addition to criteria (1) and (2), another criterion was used: (5) the induced mRNA level is substantially but not completely dependent on the function of the SA sector. This criterion was used because the mRNA level of the *PAD4* gene is up-regulated by the function of the SA sector under many conditions [111], which suggests that the PAD4 sector function is also up-regulated in an SA sector dependent manner under such conditions. As the normalization gene, At4g29480 was used in that its mRNA levels were very stable across many pathogen-related conditions in many datasets.

Figure 3-10 shows the mRNA levels of the sector marker genes, in the $\log_2$-scale relative to the normalization gene, across the treatment:genotype:time combinations. For the ET, PAD4, and SA sector marker genes, it is clear that (1) the marker gene mRNA levels are induced by MAMP treatments and (2) the induced mRNA levels are almost completely dependent on the respective sector. All the procedures associated with the selection of sector marker genes were supervised by our collaborators in Fumiaki's Lab.

3.6.1.2 Measurement of the sector marker gene expression levels

Three well-expanded leaves per plant were infiltrated with mock (H_2O), 1 μM flg22, 1 μM elf18, or 100 μg/ml chitosan using a needleless syringe. At 3 or 9 hpt treated leaves were harvested and flash frozen in liquid N_2 for RNA extraction later. Different individual plants were used for different time points because excision of some leaves at 3 hpt would affect the response in other leaves of the same individual at 9 hpt. Leaves from two to four individual plants were pooled for each biological sample. Total RNA was extracted from the tissue and subjected to quantitation by quantitative reverse transcription PCR (qRT-PCR) as previously described [119]. The obtained C_t value was used as raw data for the sector marker gene expression level. Measurements were made in three independent experiments; these are the three biological replicates.

3.6.1.3 Preprocessing of the sector marker gene expression levels to obtain the sector activity values

Between-samples normalization based on the normalized gene expression levels was first performed to make expression level values of the same gene from different samples comparable. The C_t value for each sector marker gene was subtracted from the C_t value for the normalization gene of the same RNA sample to obtain the $\log_2$-transformed expression level value for each marker gene as follows:

$$\log_2 exp.level_{sample}^{sector} = C_{t\,sample}^{\,norm.gene} - C_{t\,sample}^{\,sector}.$$

Second, each marker gene expression was weakly affected by signals mediated by signaling sectors other than the respective signaling sector. To obtain the expression level value specifically affected by the respective signaling sector, the $\log_2$-transformed expression level value for each genotype carrying the mutant allele for the sector in question was subtracted from the $\log_2$-transformed expression level value for each genotype carrying the wild-type allele for the sector for each treatment:time:replicate combination. For example, to obtain the "mutant-adjusted" $\log_2$-transformed JA marker gene expression level value for *ein2*/*sid2* at 3 hpt with flg22 in replicate 1, the $\log_2$-

transformed JA marker gene expression level value for *dde2*/*ein2*/*sid2* at 3 hpt with flg22 in replicate 1 was subtracted from the $\log_2$-transformed JA marker gene expression level value for *ein2*/*sid2* at 3 hpt with flg22 in replicate 1,

$$mutant.adjusted^{JA}_{ein2/sid2,flg22,3hpt,r1} = \log_2 exp.level^{JA}_{ein2/sid2,flg22,3hpt,r1} - \log_2 exp.level^{JA}_{dde2/ein2/sid2,flg22,3hpt,r1}.$$

This mutant-adjusting process resulted in expression levels of 0 for the respective sector being assigned to any genotypes containing the mutation corresponding to the respective sector (e.g., *dde2*-containing genotypes for the JA sector marker gene). The 0 values in these genotypes were kept as the fully preprocessed sector activity values for the genotypes, and the mutant-adjusted values in the other genotypes were subjected to further preprocessing. Third, the mutant-adjusted values were used to fit a linear mixed-effects model with the gene:treatment:genotype:time as the fixed effect and the replicate/gene as a random effect. The replicate/gene effect was subtracted from the mutant adjusted value to minimize the replicate effect in the data.

When the standard deviation (SD) values of the mutant.replicate-adjusted values within the same treatment:genotype:time levels were plotted against their means for each marker gene in the MAMP-treated data, the SD values of the replicates for one sector over the means of the replicates were different from them for the other on average.

To obtain the homogenous SD of 1 across the mean for all the marker genes, based on the variational patterns of the expressions across means for each of the marker genes, we chose different non-linear transformation functions, each of which is explained in details below.

(1) The JA and SA maker gene values tended to have higher SD values for the middle mean values (1st and 4th panels in Figure 3-11(a)). We assumed that these marker gene values follow logistic distributions. The first derivative $f'(x) = \frac{ace^{-a(x-b)}}{(1+e^{-a(x-b)})^2}$ of the logistic function $f(x) = \frac{c}{1+e^{-a(x-b)}}$ was fit to the SD vs. mean relationships using the nonlinear regression with least square. The fitted curve *f'(x)* (red curve) and the fitted values of a, b, and c are shown in Figure 3-11. The inverse $g^{-1}(x) =$

$\frac{1}{p}\log\left(\frac{x-s}{r-x+s}\right)+q$ of the logistic function $(x) = \frac{r}{1+e^{-p(x-q)}} + s$, where $r = \frac{2}{a}\log\frac{0.95}{0.05}$, $p = \frac{ac}{r}$, $q = \frac{r}{2}$, $s = b - q$ was used for transformation of the mutant.replicate-adjusted values. To moderate the transformation close to the logistic asymptotes, the parts of the inverse function corresponding to the logistic quantiles smaller than bottom 12.5% and larger than the top 12.5% were replaced with linear extensions of the inverse function at the boundaries.

(2) The PAD4 marker gene values tended to have an approximately linear relationship between the SD and mean values (3rd panel in Figure 3-11(a)) although the fitted derivative of the logistic function (red curve) and the associated parameter values as in procedure (1) are shown in the panel. We fitted a linear regression (intercept, *t*; slope *u*) instead, and the shifted log-transformation $h(x) = \frac{1}{u}\log(x + \frac{t}{u})$ was used to transform the mutant.replicate-adjusted values. To moderate the transformation near $x = -\frac{t}{u}$, the part corresponding to $x < x_{max}/4$ were replaced with linear extension of the shifted log-transformation at the boundary.

(3) The ET marker gene values appeared to already have a homogenous SD across the mean (2nd panel in Figure 3-11(a)). Therefore, the mutant-adjusted values were linearly scaled to make the average SD 1.

After the above transformation procedures, a constant was added to set the minimum value for each marker gene 0 where the minimum value was negative. These fully preprocessed values are referred to as the sector activity values and were used in the modeling process. To test the homogeneity of the SD of the sector activity, polynomial regressions of up to the fourth order were fit with the SD as the response and the mean as the explanatory variable, and the model with the lowest AIC was selected for each marker gene (Figure 3-11(b)). For each marker gene, the model with only an intercept, of approximately 1, had the lowest AIC (blue line), which indicates the transformed values have approximately homogenous SDs. Figure 3-11(d) shows the plots of the values after the transformation vs. the values before. The preprocessing of the sector marker gene expression levels were performed in the R environment.

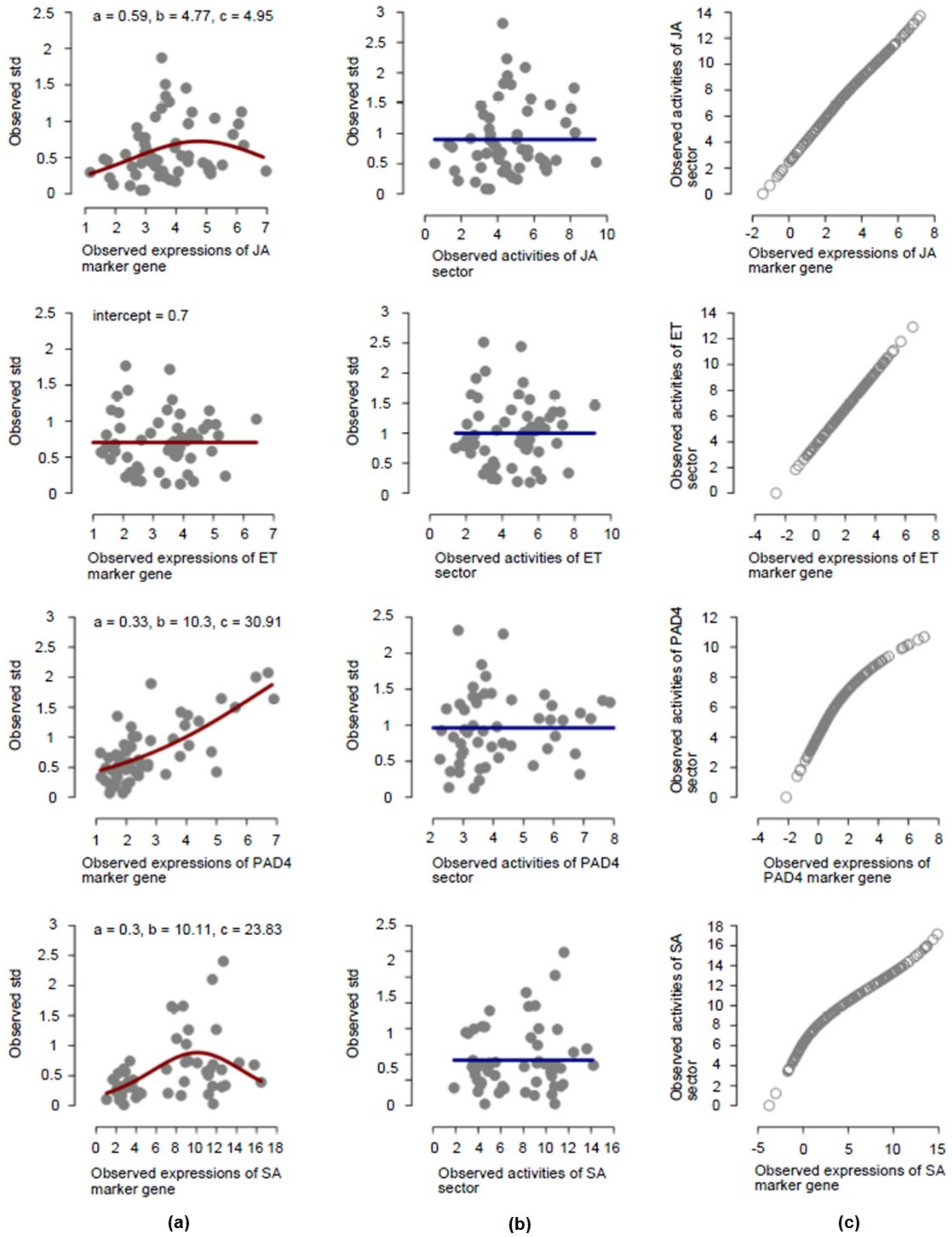


Figure 3-11 The non-linear transformations of mutant-adjusted sector marker gene expression levels. (a) The standard deviation vs. the mean of every treatment:genotype:time combination in each indicated sector before the transformation. The derivative of the logistic function was fit (red curve), and the parameter values, *a*, *b*, and *c*, are shown, except for the ET sector. For the ET

sector, a line parallel to the *x*-axis was fit, and its intercept is shown. (b) The standard deviation vs. the mean after the transformation. The blue line shows the best model determined by AIC among those up to the fourth-order polynomial: the best model was a line parallel to the *x*-axis with the intercept ~ 1 for every sector. (c) The transformation function is shown as the values after vs. the values before the transformation for each sector.

3.6.2 Measurement of MAMP-induced immunity levels

Two well-expanded leaves per plant were infiltrated with mock (H_2O), 1 µM flg22, 1 µM elf18, or 100 µg/ml chitosan using a needleless syringe. At 24 hpt, after mock or elicitor treatment, the same leaves were infiltrated with suspension of *Pto* or *Pma* with OD_{600} = 0.0001. Two days after inoculation, leaf discs were punched out from the inoculated leaves and subjected to bacterial counting by plating bacterial suspension from macerated leaf discs. For each strain:genotype:treatment combination, the bacterial count was measured in at least eight biological replicates in each of at least three independent experiments. The $\log_{10}$-transformed bacterial count (CFU/cm^2) was used in modeling.

4 Modeling a plant immune signaling network with combinatorial perturbation data: multiple regression models

4.1 Chapter overview

As explained in section 3.6, we quantitatively measured PTI levels and activities of four principal network sectors by exhaustive perturbation experiments. This large set of experimental data covered all possible combinatorial conditions in terms of genotypes, treatments, and time points which enabled a successful construction of a dynamic immune signaling network model as shown in Figure 3-4. In addition to the Bayesian network approach discussed in Chapter 3, we pursued a multiple regression modeling approach to see if we could derive models with higher prediction accuracy. The linear model is more flexible than the Bayesian approach as it allows cyclic connections within the same layer, which cannot be modeled with a Bayesian network. Our regularized multiple regression models show a high level of predictive power for dynamic sector activities and level of immunity across the large set of mutants explored here, and furthermore, it captured known and previously unappreciated signal flows in the network. The sole inhibitory sector in the model, the ethylene sector, was central to the network robustness via its inhibition of the jasmonate sector. The model's multiple input sites linked specific signal input patterns in strength and timing to different network response patterns, indicating a mechanism enabling tunability. We also predicted the sector activities and the immunity levels with the model relatively well when the data for a particular MAMP were held out, suggesting that the model would be able to reasonably predict the sector activities and the immunity levels after treatment with other MAMPs. The model also highlights specific network motifs (i.e. a tetra-stable switch) responsible for the robustness and tunability of the plant immune system, which are considered important properties of the network to withstand attacks from diverse and fast-evolving pathogens. More broadly, this study reveals several guidelines regarding the data and computational approaches necessary for modeling a complex system in various contexts.

This chapter starts by introducing the multiple regression formulation with Lasso regularization, explains model evaluation approaches including prediction accuracy and parameter stability against artificial noise, and structural invariance across different MAMP treatments. We then demonstrate an important strength of the regression approach, which is model interpretability in the plant immunity context. The chapter concludes by outlining several modeling guidelines based on our results. The work presented in this chapter was submitted to Cell Host & Microbe and includes contributions from Kenichi Tsuda, Daisuke Igarashi, Rachel A. Hillmer, Hitoshi Sakakibara, Fumiaki Katagiri, and Chad L. Myers. Kenichi and Daisuke generated experimental data, and Rachel suggested some ideas for the statistical analysis. Hitoshi generated hormone profiles. Chad and Fumiaki supervised the project.

4.2 The modeling approach: A multiple linear regression formulation with Lasso regularization

In Chapter 3, we implemented a Bayesian network approach to model a plant immune signaling network by combining two types of data: activity measurements of the four signaling sectors at 3 hpt and 9 hpt and bacterial counts of two strains after treatment with mock, flg22, elf18, and chitosan [120] in 16 combinatorial genotypes of Arabidopsis. These data are also used for multiple regression models. The regression model structure is composed of four layers of nodes. The top layer consists of three MAMP nodes, flg22 (red), elf18 (green), and chitosan (blue). All MAMP inputs are calculated relative to mock inputs, so the mock node is not included in the visualization. The yellow and orange nodes of the second and third layers represent the 3- and 9-hpt states of the four signaling sectors, respectively. The gray output node in the bottom layer represents the immunity level measured with either *Pto* or *Pma*.

As shown in Figure 4-1, a multiple regression model with each of the sector and output nodes as the response (targets of directed links) and the nodes in the preceding or same layers as the explanatory variables (sources of directed links) was set up as the starting model. The link from the 3-hpt to the 9-hpt sector nodes within each signaling sector (i.e., JA 3-hpt to JA 9-hpt in Figure 4-1(a)) is omitted from the starting model because it is not linearly independent of the link from the MAMP nodes to the 9-hpt node. Therefore, the link from a MAMP node to a 9-hpt node represents the sum of those from the MAMP node and from the early node of the same sector. The starting model was fit using Lasso regression with varying penalty (λ) values [113] in combination with a Bagging approach [114], and the predictive power of the model structure (i.e., non-zero link parameters) obtained for each λ value is evaluated. A compact model structure maintaining high predictive power was selected in this way and refit to the complete dataset using a least squares approach to estimate the parameter values. The detailed description of the modeling approach is shown in the following two sections.

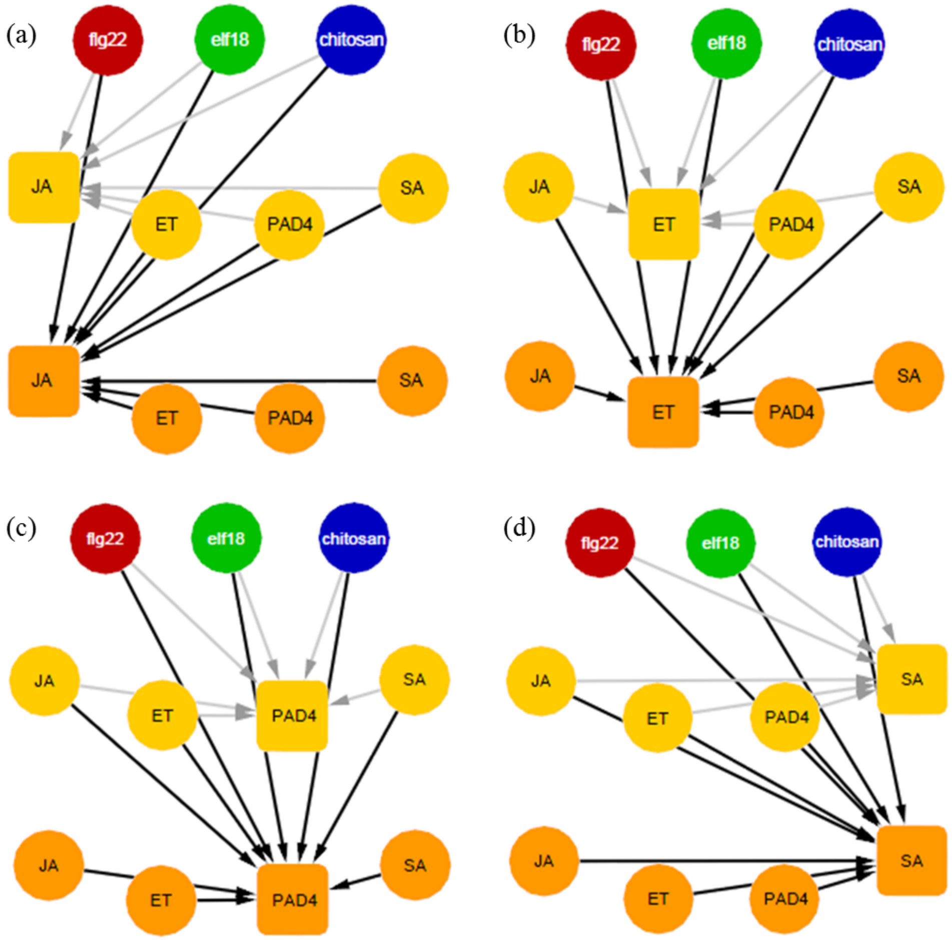


Figure 4-1 Starting network structures of regression models for the four sectors: (a) JA sector, (b) ET sector, (c) PAD4 sector, and (d) SA sector. The model consists of four layers, listed from the top: (1) three MAMP treatment inputs (flg22, red; elf18, green;chitosan, blue); (2) activities of

four signaling sectors at 3 hpt (yellow nodes); (3) activities of the signaling sectors at 9 hpt (orange nodes); (4) two immunity outputs (against *Pto* and *Pma* strains, gray nodes). Each link represents a directional dependency between an explanatory variable “source” node and a response “target” node. Gray and black links in (a) represent the models for 3 and 9 hpt, respectively.

4.2.1 Multiple regression models with four sectors

The linear regression models with L1-norm (Lasso) regularization [113] used for modeling the immune system here are formulated as

$$\hat{\boldsymbol{\beta}}^{lasso} = \min_{(\beta_0,\boldsymbol{\beta})\in\mathbb{R}^{p+1}} \left[\frac{1}{2N}\sum_{i=1}^{N}\left(y_i^m - \beta_0 - \boldsymbol{x}_{\boldsymbol{i}}^{\boldsymbol{T}}\boldsymbol{\beta}\right)^2 + \lambda\|\boldsymbol{\beta}\|_{l_1}\right] \qquad \text{Eq. 4-1}$$

where p is the total number of initial parameters without an intercept, N is the total number of cases covering all possible combinatorial conditions, β_0 is an intercept, $\boldsymbol{\beta}$ is the parameter vector to be estimated, $\boldsymbol{x}_{\boldsymbol{i}}$ consists of either binary indicators for treatment-specific variables or continuous variables directly from the corresponding activity values, y_i^m is either an actual activity value for *m* sector or an actual log10-transformed count of *m* bacteria and λ is a penalty factor balancing the prediction error and the model complexity [113].

We implemented separate linear models for the four individual sectors given that the range of activity values differed across sectors (Figure 3-11(a)). For each sector, we modeled its early activity value as a linear function of the single MAMP input and the activity values of the other sectors. The late activity value of each sector was modeled as a linear function of the early activity value of the same sector and the early and late activity values of the other sectors. A complete formulation of the regression model for the JA sector (Figure 4-1 (a), Eq. 4-2) is:

$$\begin{aligned} y^{JA} = {} & \beta_0^{JA} + 1_{3h}\cdot 1_f\cdot\beta_{f,JA_3} + 1_{3h}\cdot 1_e\cdot\beta_{e,JA_3} + 1_{3h}\cdot 1_c\cdot\beta_{c,JA_3} + 1_{3h}\cdot ET_3\cdot \\ & \beta_{ET_3,JA_3} + 1_{3h}\cdot PAD4_3\cdot\beta_{PAD4_3,JA_3} + 1_{3h}\cdot SA_3\cdot\beta_{SA_3,JA_3} + 1_{9h}\cdot 1_m\cdot\beta_{m,JA_9} + \\ & 1_{9h}\cdot 1_f\cdot\beta_{f,JA_9} + 1_{9h}\cdot 1_e\cdot\beta_{e,JA_9} + 1_{9h}\cdot 1_c\cdot\beta_{c,JA_9} + 1_{9h}\cdot ET_3\cdot\beta_{ET_3,JA_9} + 1_{9h}\cdot \\ & PAD4_3\cdot\beta_{PAD4_3,JA_9} + 1_{9h}\cdot SA_3\cdot\beta_{SA_3,JA_9} + 1_{9h}\cdot ET_9\cdot\beta_{ET_9,JA_9} + 1_{9h}\cdot PAD4_9\cdot \\ & \beta_{PAD4_9,JA_9} + 1_{9h}\cdot SA_9\cdot\beta_{SA_9,JA_9}. \end{aligned} \qquad \text{Eq. 4-2}$$

where β_0 is a constant; 1_{3h} and 1_{9h} are binary indicator variables for 3 hpt and 9 hpt respectively; 1_m, 1_f, 1_e, and 1_c are binary indicator variables for mock, flg22, elf18, and chitosan treatment, respectively; ET_3, and ET_9 represent the ET sector activity at 3 hpt and 9 hpt, respectively; $\beta_{s,t}$ is a parameter that reflects the influence (cross-talk) associated with between a source node s and a target node t for the JA sector activity. The models for ET (Figure 4-1(b), Eq. 4-3), PAD4 (Figure 4-1(c), Eq. 4-4), and SA (Figure 4-1(d), Eq. 4-5) sector are formulated in the same way as above:

$$y^{ET} = \beta_0^{ET} + 1_{3h} \cdot 1_f \cdot \beta_{f,ET_3} + 1_{3h} \cdot 1_e \cdot \beta_{e,ET_3} + 1_{3h} \cdot 1_c \cdot \beta_{c,ET_3} + 1_{3h} \cdot JA_3 \cdot \beta_{JA_3,ET_3} + 1_{3h} \cdot PAD4_3 \cdot \beta_{PAD4_3,ET_3} + 1_{3h} \cdot SA_3 \cdot \beta_{SA_3,ET_3} + 1_{9h} \cdot 1_m \cdot \beta_{m,ET_9} + 1_{9h} \cdot 1_f \cdot \beta_{f,ET_9} + 1_{9h} \cdot 1_e \cdot \beta_{e,ET_9} + 1_{9h} \cdot 1_c \cdot \beta_{c,ET_9} + 1_{9h} \cdot JA_3 \cdot \beta_{JA_3,ET_9} + 1_{9h} \cdot PAD4_3 \cdot \beta_{PAD4_3,ET_9} + 1_{9h} \cdot SA_3 \cdot \beta_{SA_3,ET_9} + 1_{9h} \cdot JA_9 \cdot \beta_{JA_9,ET_9} + 1_{9h} \cdot PAD4_9 \cdot \beta_{PAD4_9,ET_9} + 1_{9h} \cdot SA_9 \cdot \beta_{SA_9,ET_9}; \qquad \text{Eq. 4-3}$$

$$y^{PAD4} = \beta_0^{PAD4} + 1_{3h} \cdot 1_f \cdot \beta_{f,PAD4_3} + 1_{3h} \cdot 1_e \cdot \beta_{e,PAD4_3} + 1_{3h} \cdot 1_c \cdot \beta_{c,PAD4_3} + 1_{3h} \cdot JA_3 \cdot \beta_{JA_3,PAD4_3} + 1_{3h} \cdot ET_3 \cdot \beta_{ET_3,PAD4_3} + 1_{3h} \cdot SA_3 \cdot \beta_{SA_3,PAD4_3} + 1_{9h} \cdot 1_m \cdot \beta_{m,PAD4_9} + 1_{9h} \cdot 1_f \cdot \beta_{f,PAD4_9} + 1_{9h} \cdot 1_e \cdot \beta_{e,PAD4_9} + 1_{9h} \cdot 1_c \cdot \beta_{c,PAD4_9} + 1_{9h} \cdot JA_3 \cdot \beta_{JA_3,PAD4_9} + 1_{9h} \cdot ET_3 \cdot \beta_{ET_3,PAD4_9} + 1_{9h} \cdot SA_3 \cdot \beta_{SA_3,PAD4_9} + 1_{9h} \cdot JA_9 \cdot \beta_{JA_9,PAD4_9} + 1_{9h} \cdot ET_9 \cdot \beta_{ET_9,PAD4_9} + 1_{9h} \cdot SA_9 \cdot \beta_{SA_9,PAD4_9}; \qquad \text{Eq. 4-4}$$

$$y^{SA} = \beta_0^{SA} + 1_{3h} \cdot 1_f \cdot \beta_{f,SA_3} + 1_{3h} \cdot 1_e \cdot \beta_{e,SA_3} + 1_{3h} \cdot 1_c \cdot \beta_{c,SA_3} + 1_{3h} \cdot JA_3 \cdot \beta_{JA_3,SA_3} + 1_{3h} \cdot ET_3 \cdot \beta_{ET_3,SA_3} + 1_{3h} \cdot PAD4_3 \cdot \beta_{PAD4_3,SA_3} + 1_{9h} \cdot 1_m \cdot \beta_{m,SA_9} + 1_{9h} \cdot 1_f \cdot \beta_{f,SA_9} + 1_{9h} \cdot 1_e \cdot \beta_{e,SA_9} + 1_{9h} \cdot 1_c \cdot \beta_{c,SA_9} + 1_{9h} \cdot JA_3 \cdot \beta_{JA_3,SA_9} + 1_{9h} \cdot ET_3 \cdot \beta_{ET_3,SA_9} + 1_{9h} \cdot PAD4_3 \cdot \beta_{PAD4_3,SA_9} + 1_{9h} \cdot JA_9 \cdot \beta_{JA_9,SA_9} + 1_{9h} \cdot ET_9 \cdot \beta_{ET_9,SA_9} + 1_{9h} \cdot PAD4_9 \cdot \beta_{PAD4_9,SA_9}. \qquad \text{Eq. 4-5}$$

As shown in Eq. 4-1, the complexity of linear regression formulations was constrained by Lasso regularization. To measure the prediction error, we used a bootstrap aggregation approach, called Bagging [114], with 1000 rounds of resampling. More specifically, we first extracted 100 candidate penalty factors, $\lambda \in (0.001,1.5)$, based on an entire set of activity values of sector markers. For each bootstrap iteration, we randomly sampled a subset of the activity data from the entire set of treatment:genotype samples (32 different conditions: four treatments x eight genotypes) with replacement as a training dataset, estimated parameters with the candidate penalty factors, and predicted activity values of the rest of the data as a test set. With these 1000 different models for each penalty factor, a median value was selected from all predicted values for each treatment:genotype condition and theses medians were aggregated into a final test set. Among several models with different λs, based on the Pearson correlation coefficient (PCC) between observed and predicted (held-out) examples, we selected the model with the largest λ that yielded a PCC prediction performance within the 95 % confidence interval of the best measured PCC across all models [113] (see Figure 4-2 for the final network structures of the sector-specific models).

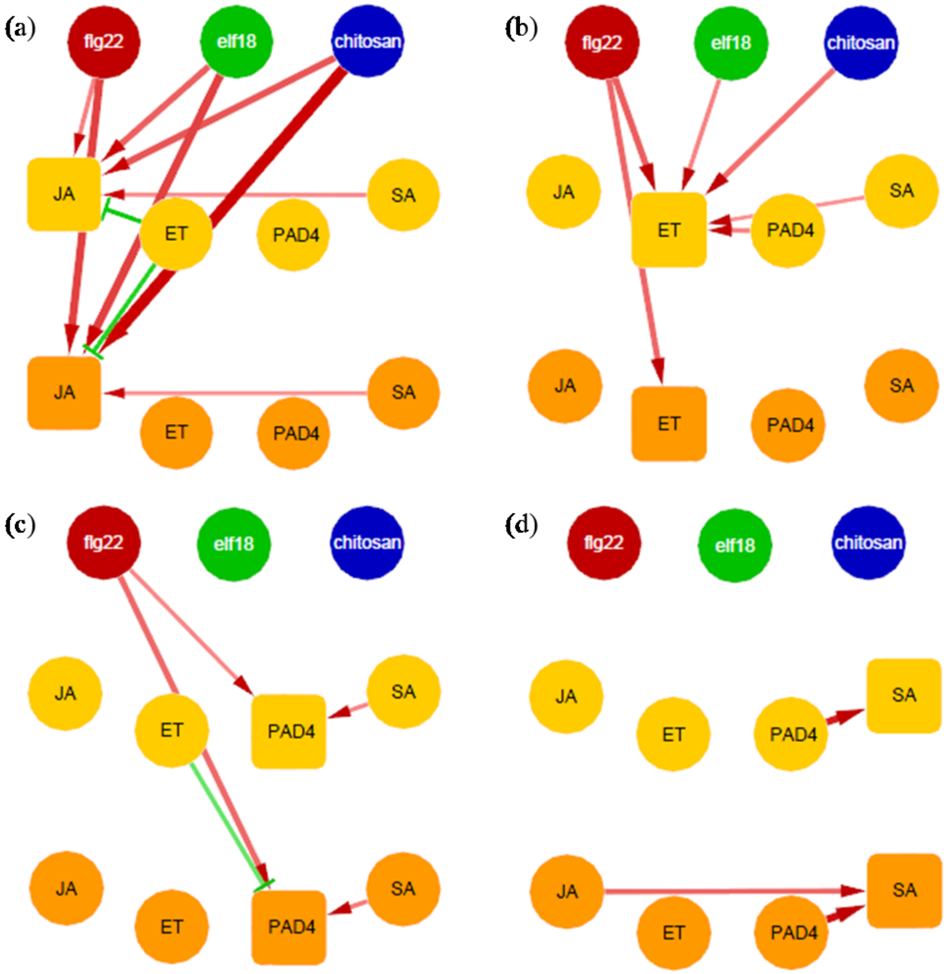


Figure 4-2 Final network structures of regression models with Lasso regularization for the four sectors: (a) JA sector, (b) ET sector, (c) PAD4 sector, and (d) SA sector. Red and green directed

links represent activation and inhibition. The width and the color intensity of a link are proportional to the link parameter value.

Given this model structure, we then refit the selected (non-zero) parameters using an ordinary least square approach with the entire activity dataset (see Table 4-1 and Table 4-2 for predicted means of parameters and their confidence intervals associated with the corresponding sectors).

Table 4-1 Parameter estimates for JA and ET sectors in the original model (Figure 4-5) and their 95 % confidence interval (TS parameters: Treatment-specific parameters; SS parameters: Sector-specific parameters).

Categories	Source	Target				Target			
		JA_{3h}		JA_{9h}		ET_{3h}		ET_{9h}	
		mean	conf. interval	mean	conf. interval	mean	conf. interval	mean	conf. interval
TS parameters	β_m	NA		0		NA		-0.288	(-0.311, -0.264)
	β_f	1.73	(1.706, 1.755)	4.32	(4.295, 4.344)	3.137	(3.144, 3.161)	2.822	(2.798, 2.846)
	β_e	3.062	(3.037, 3.086)	4.341	(4.316, 4.365)	1.632	(1.608, 1.656)	0	
	β_c	3.439	(3.414, 3.463)	6.92	(6.896, 6.945)	2.409	(2.385, 2.433)	0	
SS parameters	JA_{3h}	NA		NA		0		0	
	ET_{3h}	-0.158	(0.183, -0.134)	-0.15	(-0.174, -0.125)	NA		NA	
	$PAD4_{3h}$	0		0	0	0.142	(0.119, 0.166)	0	
	SA_{3h}	0.079	(0.055, 0.103)	0	0	0.052	(0.028, 0.076)	0	
	JA_{3h}	NA		NA		NA		0	
	ET_{3h}	NA		0	0	NA		NA	
	$PAD4_{3h}$	NA		0	0	NA		0	

	SA_{3h}	NA		0.048	(0.024, 0.072)	NA		0	

Table 4-2 Parameter estimates for PAD4 and SA sectors in the original model (Figure 4-5) and their 95 % confidence interval (TS parameters: Treatment-specific parameters; SS parameters: Sector-specific parameters).

Categories	Source	Target				Target			
		$PAD4_{3h}$		$PAD4_{9h}$		SA_{3h}		SA_{9h}	
		mean	conf. interval	mean	conf. interval	mean	conf. interval	mean	conf. interval
TS parameters	β_m	NA		0		NA		0	
	β_f	1.5	(1.485, 1.515)	2.678	(2.662, 2.693)	0		0	
	β_e	0		0		0		0	
	β_c	0		0		0		0	
SS parameters	JA_{3h}	0		0		0		0	
	ET_{3h}	0		-0.127	(-0.142, -0.112)	0		0	
	$PAD4_{3h}$	NA		NA		0.641	(0.634, 0.647)	0	
	SA_{3h}	0.113	(0.098, 0.128)	0		NA		NA	
	JA_{3h}	NA		0		NA		0.274	(0.267, 0.281)
	ET_{3h}	NA		0		NA		0	
	$PAD4_{3h}$	NA		NA		NA			(0.716, 0.730)
	SA_{3h}	NA		0.16	(0.144, 0.175)	NA		NA	

4.2.2 Multiple regression models with two bacterial strains

To model immunity levels against two bacterial strains, we fit a separate linear model relating the model-predicted activity values of the four sectors to the observed bacterial counts. First, all activity values were predicted using the process described above, except that zeros were directly assigned for any markers where the corresponding sector was deleted. In each regression model for a bacterial strain, a constant was included to capture the bacterial count in the quadruple mutant under mock treatment, and three binary indicators were used to capture MAMP-specific effects with quadruple deletion that could not be explained by the four sectors. All parameters except for the constant were multiplied by -1 so that positive values reflect positive contributions to plant immunity.

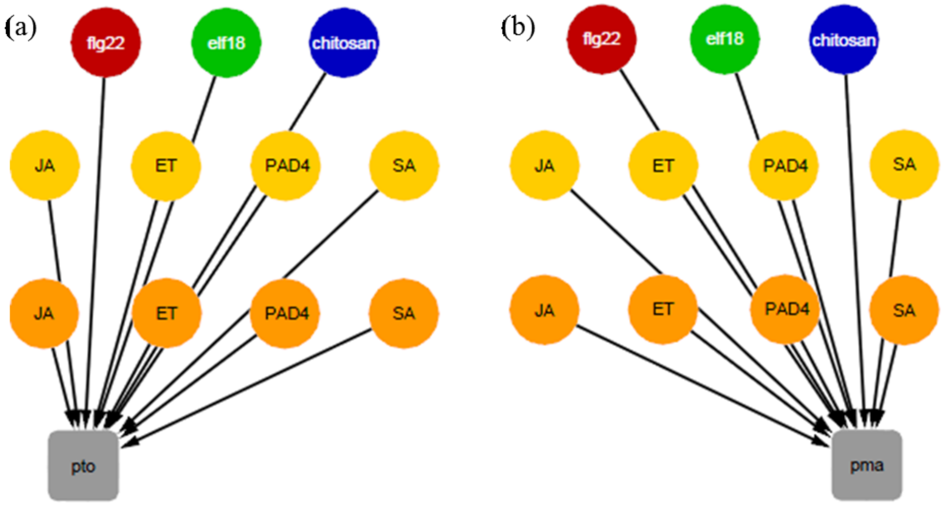


Figure 4-3 The starting network structures of regression models for (a) *Pto* and (b) *Pma* strains.

The pto-specific regression model (Figure 4-3(a), Eq. 4-6) was formulated as

$$\begin{aligned} y^{pto} = \beta_0 - (&1_f \cdot \beta_{f,pto} + 1_e \cdot \beta_{e,pto} + 1_c \cdot \beta_{c,pto} + \widehat{JA}_3 \cdot \beta_{JA_3,pto} + \widehat{ET}_3 \cdot \beta_{ET_3,pto} \\ &+ \widehat{PAD}_3 \cdot \beta_{PAD_3,pto} + \widehat{SA}_3 \cdot \beta_{SA_3,pto} + \widehat{JA}_9 \cdot \beta_{JA_9,pto} + \widehat{ET}_9 \cdot \beta_{ET_9,pto} \\ &+ \widehat{PAD}_9 \cdot \beta_{PAD_9,pto} + \widehat{SA}_9 \cdot \beta_{SA_9,pto}) \end{aligned} \qquad \text{Eq. 4-6}$$

Where β_0 is a constantan; 1_f, 1_e, and 1_c are binary indicator variables for flg22, elf18, and chitosan treatment, respectively; $\widehat{JA}_3$, and $\widehat{JA}_9$ represent the predicted activity values of JA sector marker at 3 hpt and 9 hpt, respectively; $\beta_{s,t}$ is a explanatory parameter for the effect from a sector node s to a target node t on the immunity level against pto strain.

Similarly, the pma-specific regression model (Figure 4-3(b), Eq. 4-7) was formulated as follows:

$$y^{pma} = \beta_0 - \left(1_f \cdot \beta_{f,pma} + 1_e \cdot \beta_{e,pma} + 1_c \cdot \beta_{c,pma} + \widehat{JA}_3 \cdot \beta_{JA_3,pma} + \widehat{ET}_3 \cdot \beta_{ET_3,pma} + \widehat{PAD}_3 \cdot \beta_{PAD_3,pma} + \widehat{SA}_3 \cdot \beta_{SA_3,pma} + \widehat{JA}_9 \cdot \beta_{JA_9,pma} + \widehat{ET}_9 \cdot \beta_{ET_9,pma} + \widehat{PAD}_9 \cdot \beta_{PAD_9,pma} + \widehat{SA}_9 \cdot \beta_{SA_9,pma}\right). \qquad \text{Eq. 4-7}$$

For each strain-specific model, bagging with 1000 samplings was applied to find the best model structure (see Figure 4-4 for the final network structures of the strain-specific models) and a least squares approach was used to estimate the selected parameters with the complete set of bacterial counts (see Table 4-3 for predicted means and confidence intervals).

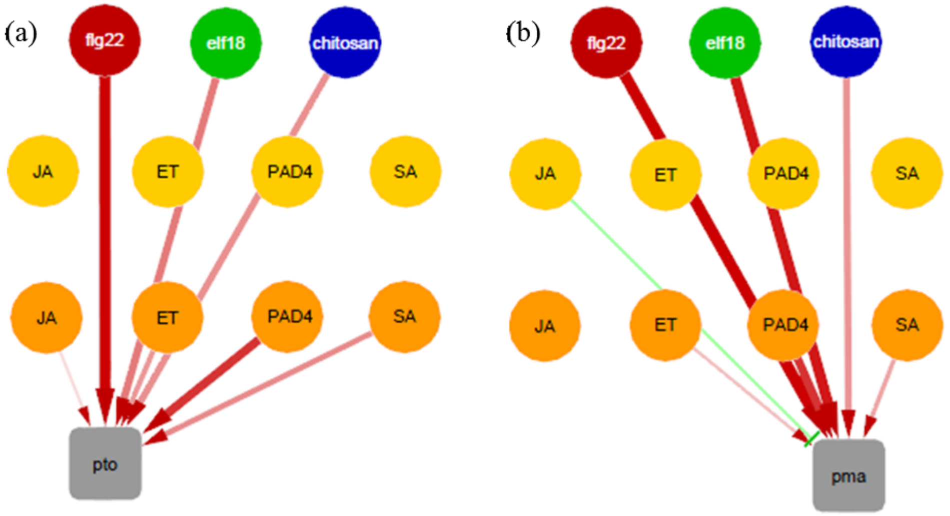


Figure 4-4 The final network structures of regression models with Lasso regularization for (a) *Pto* and (b) *Pma* strains.

Table 4-3 Parameter estimates for *Pto* and *Pma* strains in the obtained network model (Figure 4-5) and their 95 % confidence interval (TS parameters: Treatment-specific parameters; SS parameters: Sector-specific parameters).

Categories	Source	Target			
		Pto		*Pma*	
		mean	conf. interval	mean	conf. interval
TS parameters	β_f	1.711	(1.705, 1.717)	1.202	(1.197, 1.207)

	β_e	0.976	(0.970, 0.983)	1.124	(1.119, 1.129)
	β_c	0.85	(0.844, 0.856)	0.668	(0.663, 0.673)
SS parameters	JA_{3h}	0		-0.021	(-0.026, -0.016)
	ET_{3h}	0		0	
	$PAD4_{3h}$	0		0	
	SA_{3h}	0		0	
	JA_{3h}	0.008	(0.002, 0.014)	0	
	ET_{3h}	0.06	(0.054, 0.066)	0.03	(0.025, 0.035)
	$PAD4_{3h}$	0.138	(0.132, 0.144)	0.141	(0.136, 0.146)
	SA_{3h}	0.065	(0.059, 0.071)	0.05	(0.045, 0.055)

4.3 Evaluation of the modeled immune signal network: predictive power and structural stability

By combining the final regression models with the significant parameters selected by Lasso, we obtained a PTI signaling network model. In this section, the performance of our network model is assessed by evaluating how well our model predicts both dynamic sector activities and systematic immune responses to genetic perturbation. The importance of links associated with treatment-specific effects and cross-talk interactions for capturing the dynamic behavior of signaling networks is examined by fitting regression models with different initial structures. We also address the structural stability of our model by fitting regression models to noise-added data and comparing their structural to that of the original model (Figure 4-5). The structural invariance of our model is validated by fitting separate models for different treatments and comparing these treatment-specific models with the original model in terms of their predictive power and model complexity. The detailed descriptions of these results related to the validation of the modeling approach are shown in the following four sections.

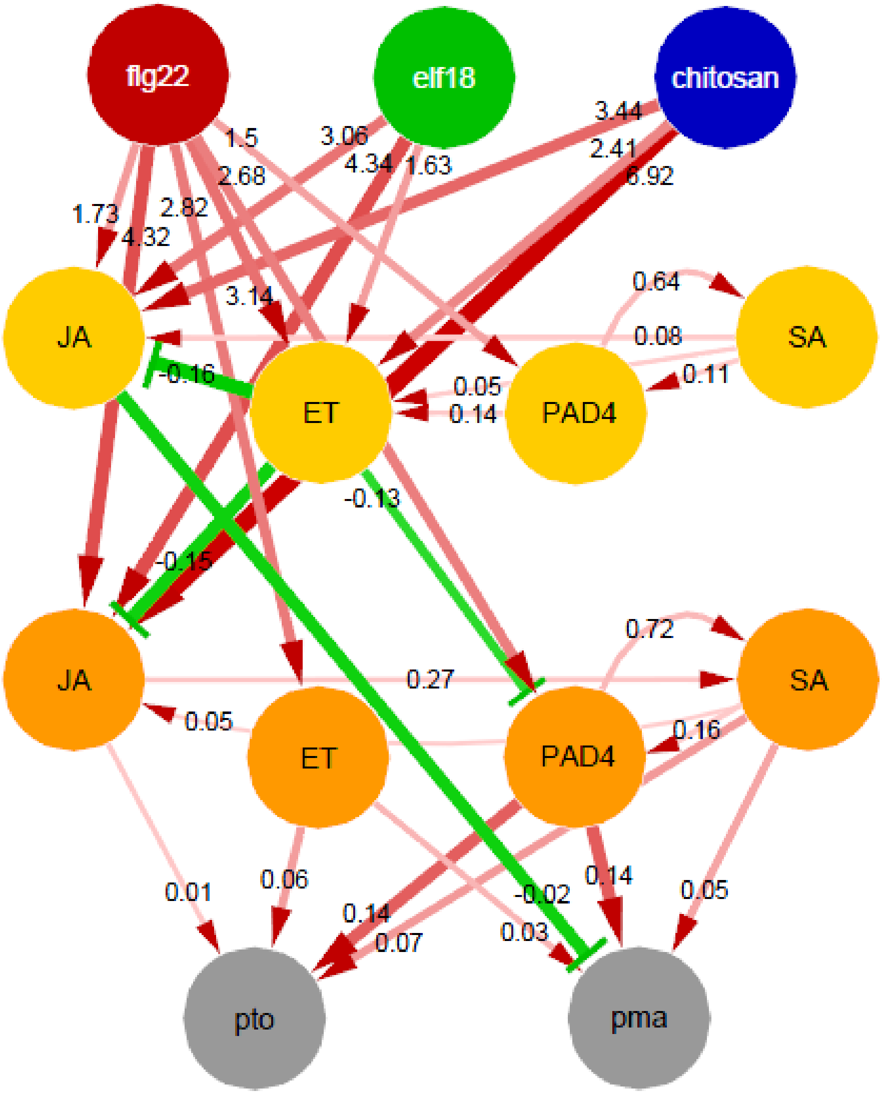


Figure 4-5 An obtained PTI signaling network model. Directional links in red and green represent the significant parameters, indicating activation and inhibition, respectively. The width and color intensity of the links represent parameter values. The links to the immunity nodes are scaled differently from those to the sector nodes for a better visualization. The links from the MAMP nodes to the immunity nodes are not shown as they represent the immunity level that is not explained by the four sectors. The estimated mean parameter value for each link is indicated.

4.3.1 Predictive power

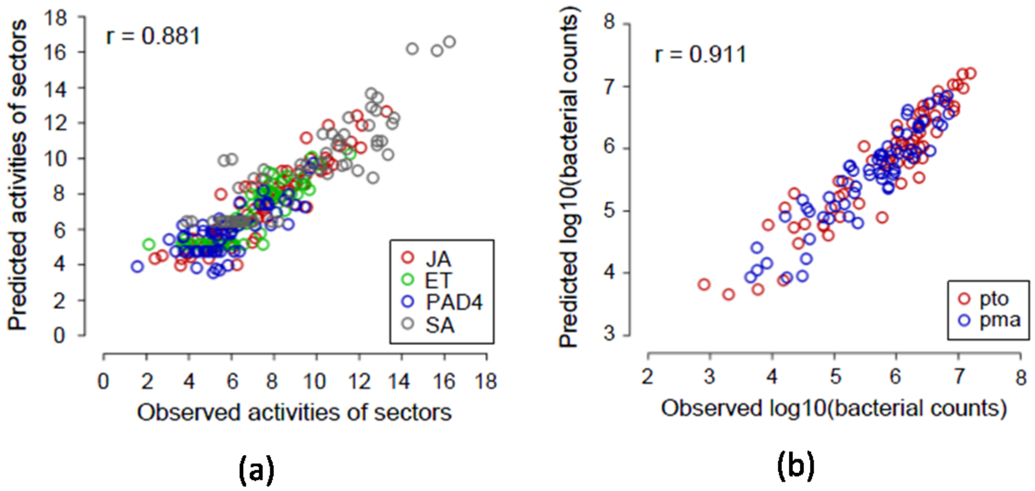


Figure 4-6 The model predictions vs. the observed data of the sector activities across the treatment:genotoype:time:sector combinations (a) and the immunity level across the treatment:genotype:strain combinations (b). The associated Pearson correlation coefficients are shown as r. The fitted values of the model and their confidence intervals are compared to the observed in Appendix I. The fitted values of final multiple regression models. Red circle, JA; green, ET; blue, PAD4; gray, SA in (a) and red, *Pto*; blue, *Pma* in (b): the separate plot for these categories are shown in Figure 4-20 and Figure 4-21.

Model predictions for each treatment:genotype:time:sector activity and each treatment:genotype:strain immunity level were made using a bagging approach similar to that described above but fitting the fixed model structure shown in Figure 4-5 using an ordinary least squares approach instead. Figure 4-6 shows high Pearson correlation coefficients (PCCs) between the predicted values and the mean observed values for the sector activities (0.881) and the $\log_{10}$(bacterial count) (0.911). These prediction analyses demonstrate that our network model faithfully captures quantitative activities of the signaling sectors and immunity levels in various conditions. To capture variant values of the sector activities across different genotypes, the incorporation of actual activity values of other sectors for predicting activity values of a sector is beneficial for increasing the prediction accuracy in the linear model. Interestingly, relatively early network responses

(3 hpt and 9 hpt) contain sufficient information to predict with high accuracy the relatively late summarized network output of the immunity level across the MAMPs, the genotypes, and the strains (the bacterial strains were inoculated at 24 hpt and counted at 72 hpt).

4.3.2 The importance of links relating to cross-talk interactions and treatment-specific effects

The previous study showed that a PTI signaling network is complex and overall signaling allocation of both flg22-PTI and elf18-PTI is similar [3]. In this section, we further test how important the links associated with the cross-communication among sectors and treatment-specific influences are to predict the observed sector activities (Table 4-4).

Table 4-4 Comparison of the original model and the model devoid of specified links

PCCs	Sector activity	Immunity level
Original model (Figure 4-5)	0.881	0.911
Model w/o inter-sector links	0.681	0.887
Model w/o MAMP-late sector links	0.752	0.921
Model w/o MAMP-late sector links but w self-interaction	0.833	0.922

When all links between signaling sectors are removed from the starting model structure, the PCCs between the predicted and observed data are reduced to 0.681 for the sector activity and 0.887 for $\log_{10}$(bacterial count), indicating that including links that model sector cross-talk helps capture the complex behavior of the signaling network. On the other hand, if we keep the inter-sector links and remove links between MAMP nodes and late sectors in the starting model, the PCC between the observed and predicted data for sector activities is 0.752. We then further add self-interactions between early and late sectors on top of the initial structure. The PCC for the sector activities with the model is now 0.833. These results justify the structure of our network model with treatment-specific links based on the predictive power the model can achieve. Since the cross-sector links already capture interactions between the four sectors, it is likely that there exist other signaling sectors, interacting with the four sectors and affecting differential activities of the four sectors in a down-stream network. The influence is successfully captured by including the treatment-specific parameters in the original model (Figure 4-5).

4.3.3 Structural stability against artificial noise

If the model structure were too complex, the structure of the final regularized model could vary with relatively small changes in the data. To test the stability of the model, random noise at varying levels was generated and added to the original dataset, the same modeling procedure was performed on the noise-added data, and the obtained models were compared to the original final model (Figure 4-7).

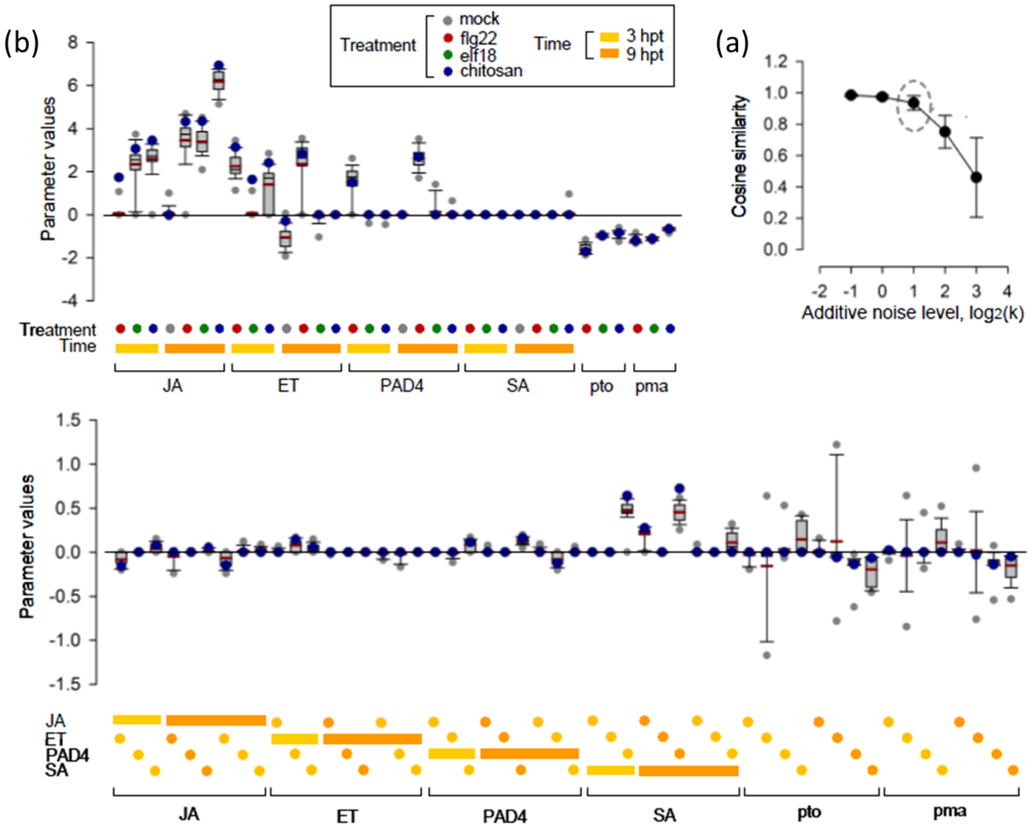


Figure 4-7 A model inference is stable against artificially added noise. (a) The parameter estimates in the models obtained with noise-added datasets are compared to the original model with the original dataset. Even when twice more artificial noise than the biological noise of the

original data was added, the parameter estimates did not change much ($\log_2 k$ = 1, dashed circle). (b) The distributions of the parameter values when $\log_2 k$ = 1 are shown by boxes-and-whiskers for the treatment-specific (upper panel) and the sector-specific (lower panel) parameters. Blue dot, parameter estimate in the original model. See key at top for the color-codes for the treatment and the time.

For example, even when the standard deviation (SD) of the added noise is twice as large as the SD of the residuals associated with the original model ($\log_2(k)$ = 1), the average similarity measure between two sets of estimated parameters, one from the original model and another from the model with noisy data, is 0.96 (Figure 4-7(a)) and the patterns of parameter values from two different models are highly similar over the different conditions (Figure 4-7(b)). Nearly identical sets of parameters are selected by Lasso with the artificial datasets having random noise. It suggests that our regression model reliably captures biological signals and it is free from overfitting issues.

4.3.4 Structural invariance for cross-communications

Another possible explanation for different levels of immunity conferred by different MAMP treatments is that the signaling network structure changes according to which MAMP treatment is used. To explore this possibility, we built separate models for different individual MAMP treatments (treatment-specific models) to compare them with the invariant model (the original model, which has the invariant links from the signaling sectors across four treatments, Figure 4-5). To build each of treatment-specific models, the data specific to one MAMP treatment and the mock data were used to train the multiple regression models similarly to the approach used for the original model (Figure 4-8). The sector activities and immunity levels were predicted with each of treatment-specific models in a similar way to one for the original model, and the predictions were compared with the observed means with the same treatment by PCC. For the invariant model, PCCs were calculated between the predictions and the observed means for particular MAMP treatments, separately (Table 4-5).

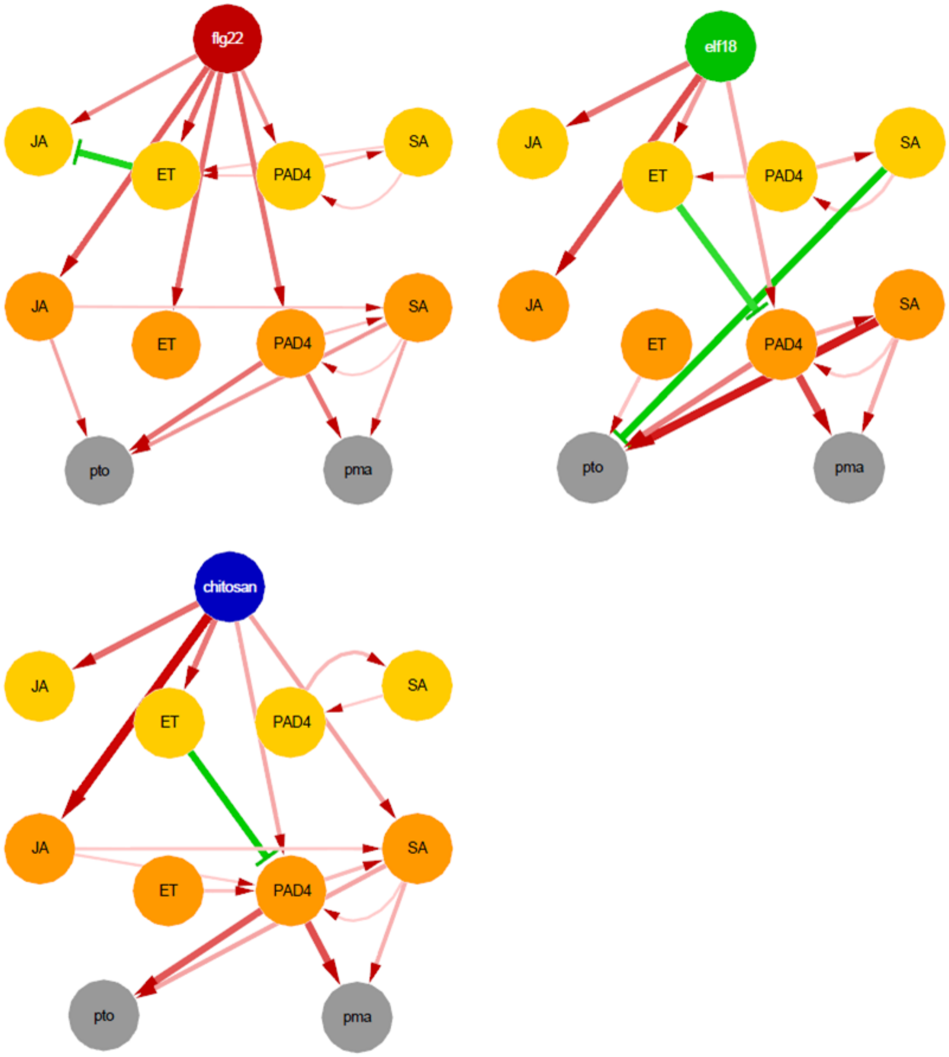


Figure 4-8 Treatment-specific models. The representations are the same as in Figure 4-5.

The predictions of the sector activities and the immunity levels by the invariant model were as good or better than those by the treatment-specific models (Table 4-5), and the invariant model was less complex (the total numbers of parameters were 45 and 82, respectively). These results support that the invariant model is the better representation of the signaling network and that differential network responses to different MAMP treatments are likely achieved by different signal input patterns, not by major rewiring in the way the network sectors interact with each other.

Table 4-5 Comparison of the invariant (original) model and the treatment-specific models.

MAMPs	The structure-invariant model		Treatment-specific model	
	Sector activities	Immunity levels	Sector activities	Immunity levels
flg22	0.89	0.93	0.88	0.93
elf18	0.85	0.91	0.84	0.93
chitosan	0.89	0.88	0.88	0.91

4.4 Interpretation of the network model and new hypotheses about the plant immune signaling network

Section 4.3 provided concrete evidence that our final model is stable in terms of the estimation of its parameters and structures and is highly predictable for dynamics of sectors and their responses to perturbations. Having confirmed good predictive performance of the model, we now discuss several ways in which we interpreted it to gain insight about the plant immune signaling network. The model reflects known cross-talk and reveals previously unappreciated connections between signaling sectors, some of which were confirmed in independent experiments. We also predict the sector activities and the immunity levels with the model relatively well when the data for a particular MAMP are held out, suggesting that the model would be able to reasonably predict the sector activities and the immunity levels after treatment with other MAMPs. The model also highlights specific network motifs responsible for the robustness and tunability of the plant immune system, which are considered important properties of the network to withstand attacks from diverse and fast-evolving pathogens. The detailed descriptions are shown in the following four sections.

4.4.1 Confirmation of hidden regulatory mechanisms by independent hormone measurements

If our modeling approaches and assumptions behind them are correct, and the experimental data quantitatively support the model, our final network model should reveal known biological mechanisms among hormone pathways. Moreover, we expect that the model can provide some novel hypotheses. Both of these directions are discussed in more detail below.

4.4.1.1 The roles of the PAD4 and SA sectors in immunity against the P. syringae strains.

A positive feedback loop structure consisting of the PAD4 and SA sectors, which is well established [121], was correctly captured by the model (bidirectional red links between the PAD4 and the SA nodes at both 3 and 9 htp in Figure 4-5). SA signaling is a positive regulator of immunity against biotrophic and hemibiotrophic pathogens, such as *P. syringe* [57]; our model concurred (red links from the SA 9-hpt to the pto and pma nodes). Although SA-independent function of PAD4 has been described [122], PAD4 has often been characterized as a positive regulator of SA signaling [123]. The high orders of network perturbation used in this study revealed a direct (i.e., not via SA) and strong contribution of PAD4 to immunity (thick red links from the PAD4 9-hpt to the pto and pma nodes). EDS1, which forms a heterodimer with PAD4, is likely involved in this direct immune contribution of PAD4 [124]. The SA sector activation by the treatments is indirect (Figure 4-5), which may be the basis of delayed activation of the SA signaling during flg22-PTI [119], possibly limiting the negative impacts of misfired immunity [104].

4.4.1.2 The JA sector activates the SA sector

Although it is often thought that JA signaling inhibits SA signaling [123], the JA sector activated the SA sector in our model (the red link from the JA to SA nodes at 9 hpt in Figure 4-5). We tested this model prediction by directly measuring the SA level at 0 and 9 hpt with flg22 in 16 combinatorial genotypes and a flg22-receptor mutant, *fls2*. Figure 4-9 shows that the change in the $\log_2$(SA level) at 9 hpt compared to that at 0 hpt in all the genotypes. The SA increase was dependent on flg22 treatment since the increase in *fls2* (white bar) was not different from the *sid2*-containing mutants (black bars). In the *pad4*–containing genotypes, the SA increase completely depended on the wild-type allele of *DDE2*, indicating that the JA sector is required for the SA increase (gray and mottled bars under "*pad4* background"). The PAD4 sector was required for SA increase in *dde2*-containing genotypes (compare left 2 bars to right 2 bars among 4 mottled bars). Since the SA increases were lower in the *DDE2/PAD4*-containing genotypes than in the corresponding *dde2/PAD4*–containing genotypes (gray and mottled bars under "*PAD4* background"), the JA and PAD4 sectors are compensatory in activation of the SA sector. Thus, this positive JA sector effect on the SA sector became evident only under high order perturbation of the system.

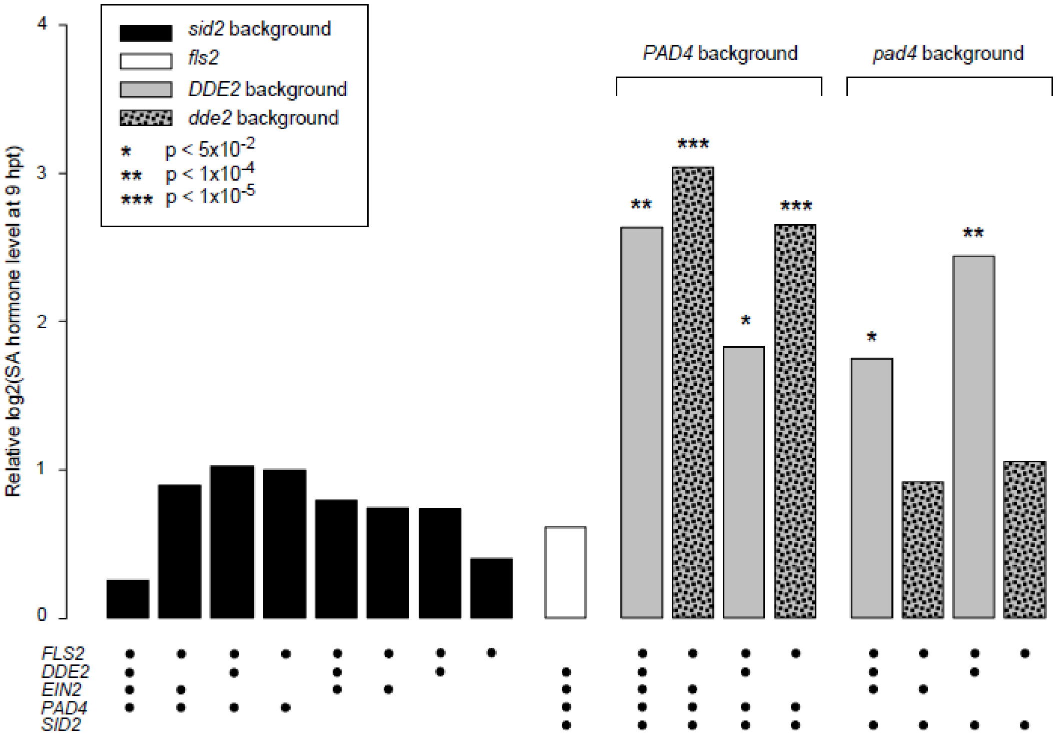


Figure 4-9 The SA upregulation after flg22 treatment required the JA sector or the PAD4 sector in a compensatory manner. The means of three biological replicates for the $\log_2$-transformed SA level increase from 0 to 9 hpt with flg22 in the indicated genotypes (black dot and blank for the wild-type and mutant alleles, respectively) are shown as bars. Significant differences from the aggregated mean of the *sid2*-containing genotypes (black bars) are indicated by asterisks.

4.4.1.3 Inhibition of the JA sector by the ET sector is important for robustness of the PTI network output

Our model predicts that the ET sector is the sole source of inhibitory effects in the immune signaling network (Figure 4-5). In particular, the model highlights inhibition of the JA sector by the ET sector. We tested this prediction by directly measuring the JA level at 9 hpt with flg22 in 16 combinatorial genotypes and *fls2* (Figure 4-10). The JA levels in *ein2*-containing genotypes were always higher than the corresponding *EIN2*-

containing genotypes (paired mottled and gray bars), indicating that the ET sector inhibits the JA sector regardless of the states of the other sectors.

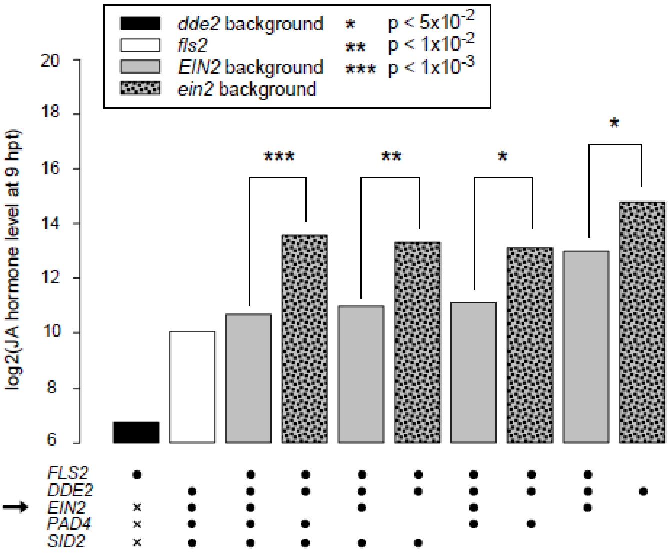


Figure 4-10 The JA level was lowered by the ET sector activity at 9 hpt with flg22. The bars show the means of the $\log_2$-transformed JA levels in three biological replicates for the indicated genotypes as in Figure 4-9. The *dde2*-containing genotypes were aggregated to one as many measurements were below detection (black bar; x for the genotype represents either wild-type or mutant alleles). Asterisks indicate the significances in the comparisons between *EIN2* (gray bars) and *ein2* (mottled bars) of the same genetic backgrounds.

4.4.2 Network properties of plant immune signaling network predicted from our model: network robustness and tunability

4.4.2.1 Network robustness from differential fragilities of the four sectors

One signaling sector inhibiting another can produce robust network output. When the first sector is compromised by, for example, a pathogen effector, the second sector is released from inhibition and backs up the function of the first sector for the network output, which we refer to as a “sector-switching” network [125]. We tested whether the ET sector inhibition of the JA sector contributes to network robustness. We measured how much loss of a sector in question affects network fragility, which we defined as the impact the further removal of another (secondary) sector has on the network output, called differential fragility. For example, the impacts of removing the PAD4 sector (a secondary sector) when the JA sector is in question was calculated for each treatment:strain combination as $|m_{dde2/pad4} - m_{dde2}|$ for the *dde2* background and $|m_{pad4} - m_{wt}|$ for the *DDE2* background, where m_k is the mean immunity level of genotype k. The differential fragility of the JA sector for the treatment:strain:PAD4 combinations was calculated by $|m_{dde2/pad4} - m_{dde2}| - |m_{pad4} - m_{wt}|$. If a differential fragility is positive or negative, removal of the sector in question decreases or increases robustness, respectively; fragility is the opposite of robustness. The fragility values for each treatment:strain:secondary_sector combination in the absence vs. the presence of each sector in question are plotted in Figure 4-22.

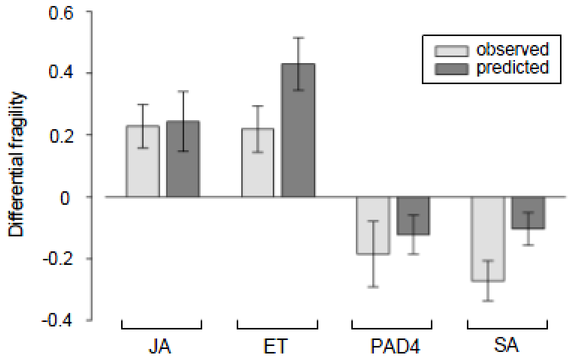


Figure 4-11 The JA and ET sectors are salient factors in network robustness. The differential fragility, that is, how unconstrained the output of the remaining network structure is when a given primary sector is present or absent was calculated for predicted (light gray bars) and observed (dark gray bars) immunity levels. A positive differential fragility means that removal of the indicated sector decreases network robustness. Each bar shows the mean and standard error across the appropriate treatment:strain:secondary_sector combinations. We also measured the differential fragility with the predicted bacterial counts which provided the similar results with the observed data. The fragility values for each sector are plotted in Figure 4-22.

Figure 4-11 shows the mean and standard error of the differential fragility values. The differential fragilities were calculated from either the observed data or model predictions (light and dark gray bars), demonstrating consistent results between them. Loss of the JA or ET sectors increased network fragility (decreased robustness) while loss of the PAD4 or SA sectors decreased fragility, indicating both the JA and ET sectors are important for network robustness. The regulatory relationship between the JA and ET sectors requires both the inhibition source (ET) as well as its target (JA); removal of either sector eliminates the ET to JA link. Thus, we conclude that this inhibitory link is important for network robustness. This observation suggests that the intact network internally represses signaling sector(s), keeping the network output suppressed relative to its maximum possible activation. This suppression could be interpreted as a tradeoff of gaining network robustness, but plants may not benefit from higher levels of PTI. Non-pathogenic microbes also present MAMPs, and if plants respond to non-pathogenic microbes with a strong immune response, it would cost plant fitness. It is conceivable that the level of inhibition by the ET sector on the JA sector has been selected to probabilistically optimize plant fitness in two ways: limiting response to imperfect information and conferring network robustness against potential perturbations by pathogen effectors.

4.4.2.2 Network tunability in different treatments

To systematically examine the mechanistic behavior of network components triggered by different MAMP inputs, we generated activity maps that enabled us to quantitatively visualize the variation in activity levels of components and signaling flows among them in specific genotypes and/or treatments. To do so, we first made predictions of node activities for all 60 conditions (four treatments x fifteen combinatorial genotypes, excluding the quadruple deletion). In each activity map, the values and sizes of the nodes, with the exception of MAMP nodes, are proportional to their predicted activity level (either predicted activity values for four sectors or bacterial counts). For the two immunity nodes, we simply subtracted the constants ($\beta_{f,pto}, \beta_{e,pto}, \beta_{c,pto}, \beta_{f,pma}, \beta_{e,pto}, \beta_{c,pto}$) associated directly with each MAMP input such that our activity maps explained the contributions to immunity only related to the four sectors. The value and width of each edge were calculated by multiplying its estimated parameter by the predicted activity level of any source node so that it essentially reflects the amount of information flow through the corresponding point of the network (Figure 4-12 and Appendix II. All activity maps for plant PTI network).

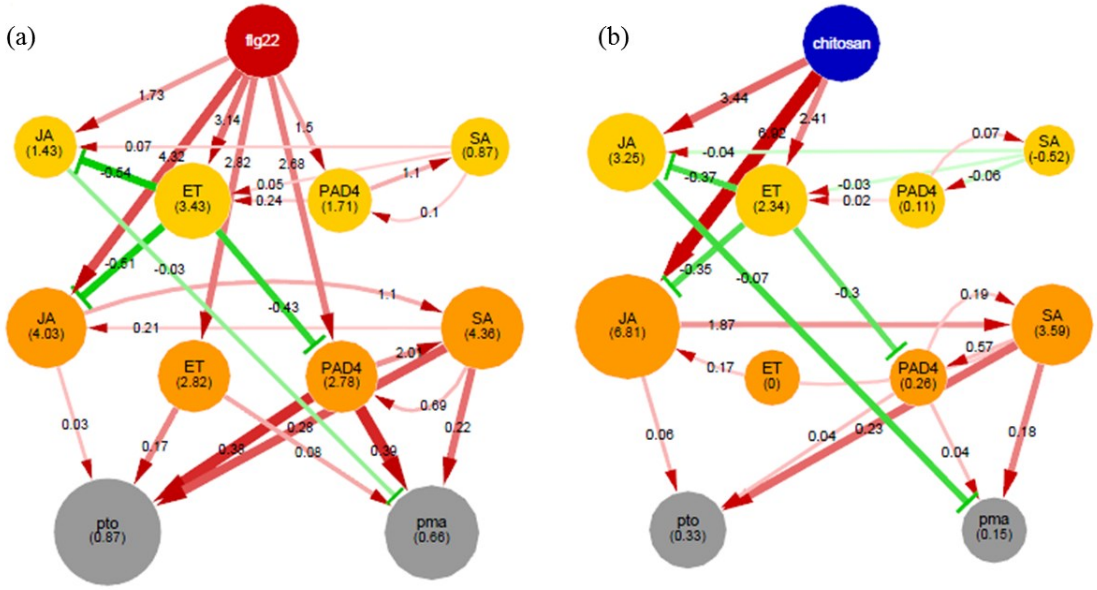


Figure 4-12 Treatments of wild type with flg22 (a) and chitosan (b) have different signal input patterns, which results in distinct network responses. These "network activity maps" make visual the model-predicted relative sector activity and immunity level by the respective node sizes, and the relative amount of signal flow by the width and color intensity of the link: link values were calculated relative to mock treatment for each MAMP treatment The range of immunity levels was scaled from 0 to 1, and the links targeting the immunity nodes are scaled according to the scaled immunity level. The predicted relative sector activity and the relative signal flow values are indicated in the node and along the link, respectively. Otherwise, the representation is the same as the model with parameter values in Figure 4-5. The maps for all treatment:genotype combinations are shown in Appendix II. All activity maps for plant PTI network.

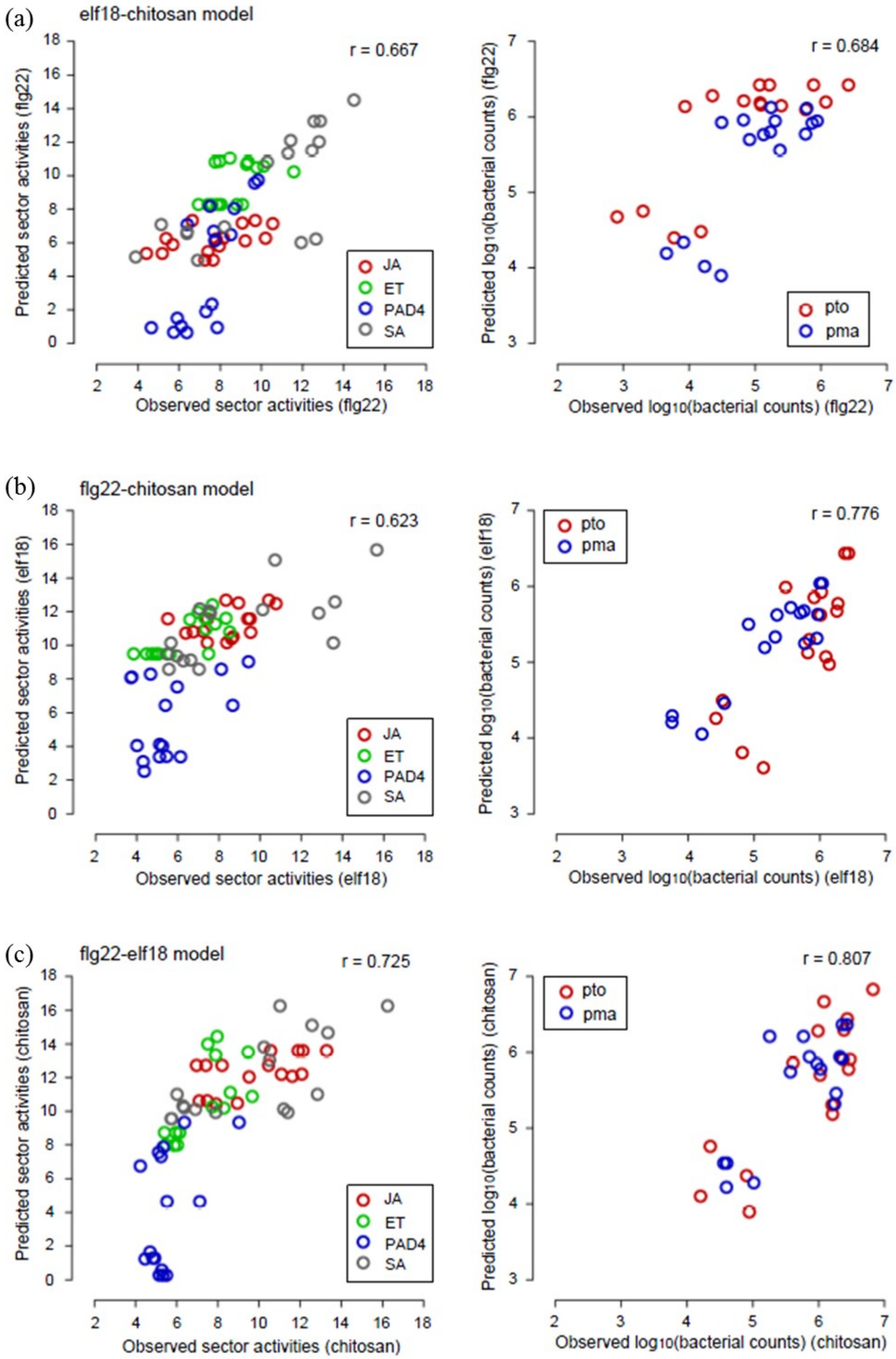
Given activity maps reflecting each genotype:treatment, the final activity maps were derived by subtracting the node value in one treatment from the value for the same node in another treatment in the same genotype. We then multiplied the values for sources with the corresponding parameter to calculate the edge value for each differential map. Given any genotype, all node values in the differential activity maps (Figure 4-12 and Appendix II. All activity maps for plant PTI network) describe the deviation of the activity level of any component under the specific MAMP treatment from that of the same component in mock, and all edge values represent the deviation of the information flow under the MAMP treatment from the mock case. Note that values assigned to both nodes and edges in the differential activity maps can be negative. For visualization purposes, we compensated for differences in absolute scale between the two types of measurements (sector activity values and log-transformed bacterial counts) by scaling the edge values associated with immunity nodes such that their mean and the mean of any edge values in each part of the network were the same.

The network activity maps revealed how different MAMP treatments lead to very different responses of the same network (Figure 4-12): the network response is modulated by a combination of multiple input points and the patterns of intensity and timing on these inputs. For example, flg22 is a MAMP derived from Gram-negative bacteria, such as *P. syringae* [126]. flg22 treatment activated the JA, ET, and PAD4 sectors at similar levels, which resulted in strong immunity against *P. syringae*, as the PAD4 sector activation at 9 hpt strongly contributes to immunity. In contrast, chitin is a MAMP derived from fungi [118], among which many are necrotrophs. Chitosan treatment predominantly activated the JA sector, which controls responses effective against necrotrophs. Thus, the network input patterns associated with flg22 and chitin appear to be selected to optimize the network response for immunity against respective pathogens.

Treatment with elf18, a bacterial MAMP, exhibited an input pattern that is more similar to that with chitosan than with flg22 (Appendix II. All activity maps for plant PTI network). This might indicate that the JA sector strongly contributes to the immunity levels against some bacterial pathogens. There are necrotrophic bacterial pathogens, such as *Pectobacterium* species [127]. Another possibility is that the input pattern from EFR, the elf18 receptor, has not yet been adapted well as EFR seems to have evolved relatively recently [128], [129]. Yet another possibility is that EFR may also recognize other MAMPs. The chitin receptor CERK1 is also a co-receptor of the bacterial MAMP peptidoglycan [130], which raises the possibility that a PRR may be involved in recognition of multiple unrelated MAMPs. If this is the case, the overall distribution of microbes that are recognized by EFR would shape the EFR-mediated network response. It will be useful to classify a broader set of PRRs according to their input patterns to the signaling network, as it was performed in this study, to gain insight in the spectrum of microbes recognized by the PRRs in nature. Furthermore, if a single PRR is indeed involved in recognition of MAMPs originated from different types of microbes, its signal input pattern into the invariant network could be locally adapted as the probabilities of encountering different types of pathogens likely vary in different environments. If this is the case, substantial variation among ecotypes of a same plant species in the signal input pattern after treatment with particular MAMPs is expected.

4.4.3 Prediction of sector activities and immunity levels for the held-out treatment

In the previous section, we propose that an invariant network is tuned by different input patterns. If indeed the hypothesis is correct, we should be able to predict the network response to an arbitrary MAMP as long as we know the input signal pattern. To test this, the model parameters were fit to subsets of the observations in which the data from one of the three MAMPs were withheld. Using only the sector activities of the wild-type genotype after treatment with the withheld MAMP, the input signal pattern was inferred, and the sector activities and the immunity levels of the other genotypes after treatment with the withheld MAMP were then predicted (see Section 4.8.5). Predicted immunity and sector activities were compared with the actual observations for each withheld MAMP treatment (Figure 4-13). The PCCs between the model predictions and observed data ranged from 0.623 to 0.725 for the sector activities and from 0.684 to 0.807 for the immunity level. This level of predictive power supports the hypothesis and suggests that the invariant model captures the core network parameters well. Furthermore, it indicates that the same invariant model will likely be able to predict the sector activities and the immunity levels across the genotypes for any other MAMPs given only sector activity measurements after MAMP treatment of the wild-type genotype.



(a) elf18-chitosan model

(b) flg22-chitosan model

(c) flg22-elf18 model

Figure 4-13 Prediction accuracy for the held-out MAMP treatment.

4.4.4 A network motif in the immune system

In the final model shown in Figure 4-5, there is an interesting pattern that emerges in plant immune signaling network. The pattern is presumably a basic architecture, responsible for the network properties, robustness and tunability of plant immune system. In this section, we extract the local pattern and analyze its properties in more detail.

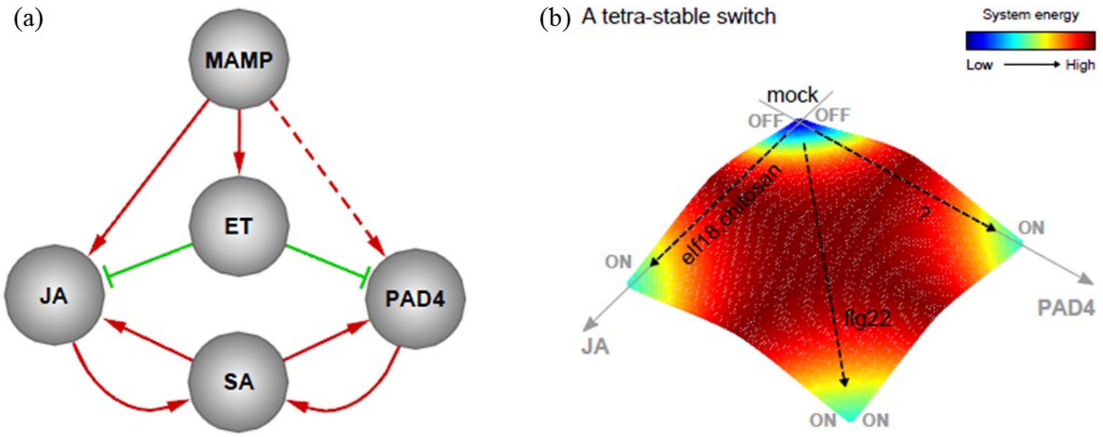


Figure 4-14 A static summary of the model suggests a JA-PAD4 dichotomy and tetra-stable network states. (a) A static summary of signal flows in the sector activity model shown in Figure 4-5. (b) A schematic representation of tetra-stable states of the signaling network. The “only JA on”, “both on”, “only PAD4 on”, and “neither on” states are speculated to correspond to lower energy states of the network.

Figure 4-14(a) depicts a static summary of major regulatory relationships in the model shown in Figure 4-5. This summary network suggests that the JA and PAD4 sectors play symmetric roles in the sense that they are both inhibited by the ET sector and are involved in positive feedback loops with the SA sector. The observation that the JA and PAD4 sectors are compensatory in activation of the SA sector (Figure 4-9) is consistent with this symmetry. A characteristic property of a positive feedback loop is its ability to act as a bi-stable switch [131]. It is conceivable that both the JA-SA and PAD4-SA feedback loops function as bi-stable switches with the activation thresholds adjusted by inhibition from the ET sector. More specifically, the JA-SA and PAD4-SA switches can be turned on only when a signal that is sufficiently strong to overcome the ET sector inhibition is directly fed into the JA and PAD4 sectors, respectively. For example, the chitosan input pattern, in which a strong signal is fed into the JA sector and a moderate signal is fed into the ET sector, may assure that the PAD4 sector activity is kept off while the SA sector is activated (Figure 4-12(b)).

With inhibition by the ET sector and the positive feedback loops with the SA sector (Figure 4-14(a)), the JA and PAD4 sectors may be pushed toward one of four rather qualitative states, “only JA on”, “only PAD4 on”, “both on”, or “neither on,” i.e., a tetra-stable switch (Figure 4-14(b)). The case with mock treatment represents the “neither on” state, the cases with chitosan and elf18 represent the “only JA on” state, and the case with flg22 represents the “both on” state (Figure 4-12 and Appendix II. All activity maps for plant PTI network). The “only PAD4 on” state might be caused by treatment with MAMPs not included in our study. An alternative, tantalizing possibility is that the “only PAD4 on” state might be related to effector-triggered immunity (ETI). ETI triggered by the *P. syringae* effector AvrRpt2 (AvrRpt2-ETI) is largely dependent on the signaling network defined by the four signaling sectors, and the AvrRpt2-ETI level is largely intact in *dde2*/*ein2*/*sid2* (i.e., only *PAD4* is wild type) [3]. These observations suggest that a different state of the same invariant signaling network obtained by modeling PTI signaling may explain AvrRpt2-ETI signaling and that the response mediated by the PAD4 sector is crucial for AvrRpt2-ETI.

A dichotomic view that the responses mediated by the JA sector and the SA sector are important for immunity against necrotrophic pathogens and immunity against biotrophic and hemibiotrophic pathogens, respectively, is generally accepted [57]. However, the symmetric roles of the JA and PAD4 sectors in the network raise a question: is it the SA sector *per se* that is important for immunity against biotrophs and hemibiotrophs or is the main role of the SA sector to activate the PAD4 sector which mediates more effective responses against these pathogens? Our observation that the (SA-signaling-independent) PAD4 sector’s effect on *Pto* and *Pma* immunity is stronger than that of the SA sector (Figure 4-5) is consistent with the latter. The high orders of network perturbations utilized in this study will allow rigorous testing of these hypotheses.

4.5 Guidelines for modeling a complex network with perturbation data

As explained in Chapter 1, in systems biology, there have been many studies to model biological systems by incorporating experimental perturbation data with computational approaches. Genetic perturbations provide causal relationships between cellular components and chemical perturbations enable the analysis of behaviors of network components under specific environmental conditions. Recently, these two experimental perturbations have been combined to increase the ability to capture more accurate system dynamics and the biological mechanisms controlling them. Nevertheless, biological prior knowledge, used for the experimental design, is still highly limited and considering all possible cases in which a biological system differentially behaves is highly intractable and nearly impossible. These challenges in modeling a complex biological system require more systematic analysis to explore what exactly is needed to characterize the biological system and how much they affect the modeling performance.

By combining dynamic activity values of the four sectors and immunity levels against two bacterial strains with exhaustive experimental perturbations, we successfully revealed the network structures with cross-talk and connect the relating sectors and the immune response in different conditions. Given the modeled system, in this section, we examine the relative impact of several modeling factors on the accurate modeling of a complex system. In particular, we first fit the regression models to partial data with limited orders of genetic perturbations and illustrate the advantage of having high order genetic perturbations regarding predictive power and structural accuracy. We then extend this analysis by modeling the signaling network with multiple constrained data in terms of the scope of the targeted sectors, the order of genetic perturbations, and the number of MAMP treatments considered. These systematic analyses provide design principles and modeling guidelines for increasing the ability to reconstitute a complex system with high accuracy. The detailed descriptions of each analysis are shown in the following two sections.

4.5.1 Models based on the data with limited orders of genetic perturbations

Previously, many studies used genetic perturbation approaches to infer causal influences of selected cellular components by measuring the resulting phenotypes from the deletion of plausible targets. These approaches necessarily require much attention to the selection of the targets of the genetic perturbation due to the large size of the search space and our incomplete knowledge of the system. To overcome this challenge, in population genomics, researchers connect natural variations of combinatorial genetic mutations and various phenotypic measurements to increase their ability to capture more accurate network structure [8]. However, it is unclear how much the order of combinatorial genetic perturbations affects the modeling accuracy. Thus, we systematically analyzed the effects of the limited orders of genetic perturbation on the modeled system with our combinatorially complete datasets associated with the four sectors.

To evaluate the benefit of higher order perturbations in terms of our ability to derive an accurate and stable model, we fit multiple regression models using the same initial model structure to three different datasets covering limited orders of genetic perturbations: (i) the "single-double" dataset consisting of the data for single mutant, and double mutant genotypes, (ii) the "up-to-single" dataset consisting of the data for wild-type and single mutant genotypes, (iii) the "up-to-double" dataset consisting of the data for wild-type, single mutant, and double mutant genotypes, and (iv) the full dataset for comparisons (the original model, Figure 4-5).

Since we adjusted the expression level values of the marker for each sector by subtracting the expression level values of the same genotypes except the mutant allele for the sector of interest (see details in Section 3.6.1.3), there were only eight effective genotypes to be considered for predictions out of the possible 16 possible combinatorial genotypes for each sector while there were zeros assigned for the rest of genotypes (eight null mutations). For the three genotype-constrained models (GCMs) with differentially limited orders of genotypes, different datasets for activity values of each sector with the genotype-constraint were used to estimate the parameters associated with the sector of interest (Table 4-6). For comparison, we also added the original model fit to the data for all 16 combinatorial genotypes of the four sectors as (iv).

Table 4-6 Effective genotypes for each set of different datasets for GCMs.

Genotype		Sectors			
		JA	ET	PAD4	SA
wild-type		(ii, iii, iv)	(ii, iii, iv)	(ii, iii, iv)	(ii, iii, iv)
single	*dde2*	0	(i, ii, iii, iv)	(i, ii, iii, iv)	(i, ii, iii, iv)
	ein2	(i, ii, iii, iv)	0	(i, ii, iii, iv)	(i, ii, iii, iv)
	pad4	(i, ii, iii, iv)	(i, ii, iii, iv)	0	(i, ii, iii, iv)
	sid2	(i, ii, iii, iv)	(i, ii, iii, iv)	(i, ii, iii, iv)	0
double	*dde2/ein2*	0	0	(i, iii, iv)	(i, iii, iv)
	dde2/pad4	0	(i, iii, iv)	0	(i, iii, iv)
	dde2/sid2	0	(i, iii, iv)	(i, iii, iv)	0
	ein2/pad4	(i, iii, iv)	0	0	(i, iii, iv)
	ein2/sid2	(i, iii, iv)	0	(i, iii, iv)	0
	pad4/sid2	(i, iii, iv)	(i, iii, iv)	0	0
triple	*dde2/ein2/pad4*	0	0	0	(iv)
	dde2/ein2/sid2	0	0	(iv)	0
	dde2/pad4/sid2	0	(iv)	0	0
	ein2/pad4/sid2	(iv)	0	0	0

The multiple regression models were fit for GCMs by resampling a subset of data from the corresponding combinatorial treatment:genotype datasets (four treatments x corresponding effective genotypes) with replacement as training sets. In each GCM, among 100 models with different λs, we chose the structures of the GCM with largest λ and the same number of non-zero parameters as in the original model. Given these structures, we then used an ordinary least square approach with the entire set of genotype-constrained data for estimating the parameters. To assess the prediction accuracies of GCMs, we also used bagging by repeatedly training regression models, having only significant parameters from Lasso regularization, with sampled data from the corresponding genotype-constrained data and evaluating their performance on the held-out data (see Section 4.3.1 for details). For comparison of the prediction accuracies between the original model and each of GCMs given each training dataset, we also applied the bagging approach to retrain the original model with the dataset and calculate PCCs (left sides: Figure 4-15). Negative $\log_{10}$-transformed p-values of the PCCs were also calculated to show the significance of the predictive powers by considering different numbers of examples in the datasets. We also tested the GCMs to predict a strict held-out dataset for only triple mutation genotypes, which was not used for modeling. In this test, we used two different sets of estimated non-zero parameters from each of the GCMs and the original model fit to the same datasets to predict values for triple mutations.

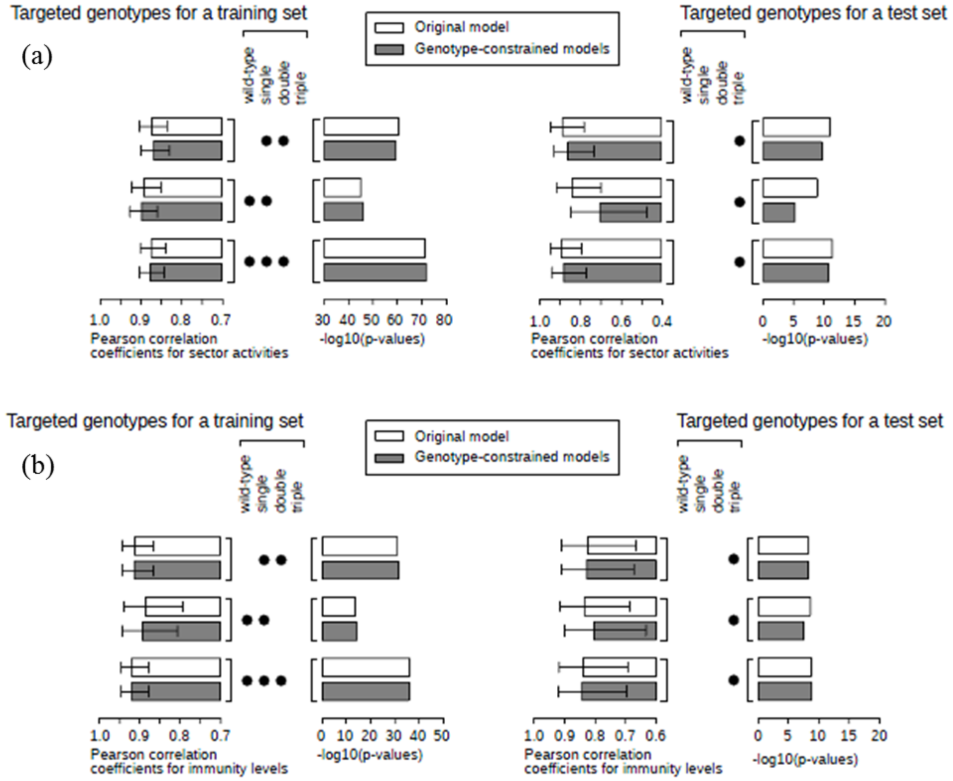


Figure 4-15 The genotype-constrained models obtained with only a low order of sector perturbations have poor predictive power compared to models derived from data collected from higher order perturbations. The PCCs between the model predictions and observed data for the triple mutants are shown. (a) PCCs for sector activities (b) PCCs for immunity levels

To measure the structural similarity in inter-sector links between any pairs of the models, we used a Jaccard index, defined as

$$J_{m,n}\left(\boldsymbol{\beta}_m^p, \boldsymbol{\beta}_n^p\right) = \frac{|\boldsymbol{\beta}_m^p \cap \boldsymbol{\beta}_n^p|}{|\boldsymbol{\beta}_m^p \cup \boldsymbol{\beta}_n^p|}. \qquad \text{Eq. 4-8}$$

where $m, n \in \{i, ii, iii, iv\}$, $p \in \{\text{treatment-specific, sector-specific}\}$, and $\boldsymbol{\beta}^{p}_{m(n)}$ is a binary column vector for p parameters associated with the sectors from $m(n)$ model. Note that models consisted of different numbers of parameters for either treatment-specific or sector-specific parameters (numbers near the heatmap in Figure 4-16) although we set the total number of parameters for each of the sector-specific models the same across the models.

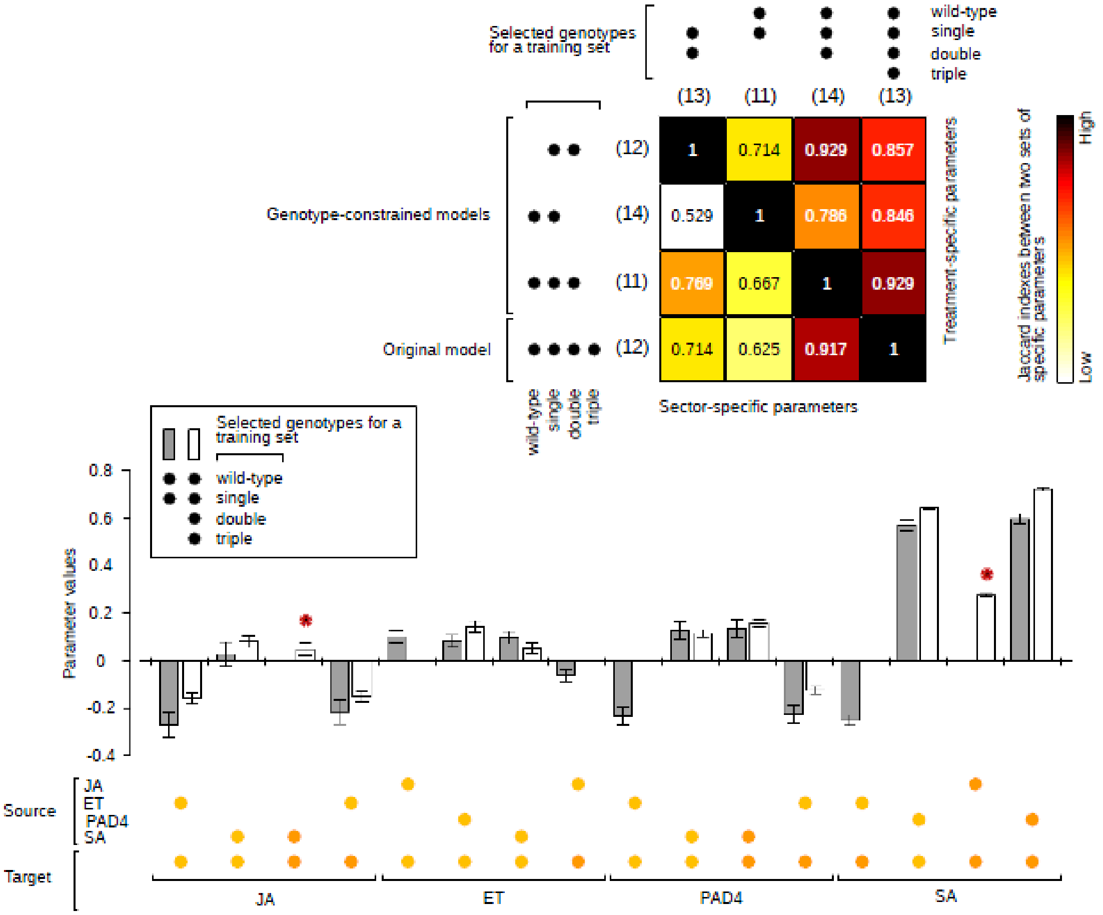


Figure 4-16 The genotype-constrained models. The genotypes used in modeling are indicated by black dots. The heatmap shows the Jacccard index values (Eq. 4-8) for the sector-specific (bottom-left half) and the treatment-specific (top-right half) parameters. The bar plot shows a comparison of the parameter estimates between the "up to single" model and the original model. Two signaling interactions captured only by the original model are indicated by red asterisks right above the bars.

Figure 4-16 and Figure 4-15 show that higher order perturbation data significantly contributed to the predictive power and accurate structure of the model. A partial dataset covering low-order genotypes contained limited information for the activities of network components such that the corresponding GCM could not reliably reconstitute the system. Furthermore, due to the compensatory roles of some of network components, high-order genetic perturbation data is critical for uncovering hidden mechanisms. For example, in the plant immunity field, the inhibition of SA hormone sector for JA sector was generally observed by many studies. However, surprisingly, our model suggested an opposite mechanism, the cross-activation between the JA and the SA sector, which was confirmed by direct SA hormone measurements (Figure 4-9). The analysis on GCMs also supported this activation by showing that the activation process could be revealed only by our full dataset with high-order perturbation data (see the cross-activation between JA and SA with red asterisks in Figure 4-16).

In terms of prediction accuracy, the network structures of the GCMs with the limited dataset covering partial perturbation depths were less accurate than that of the original model with a complete dataset. Figure 4-15 show that the network structure from the model with the wild-type and single mutation data did not have the ability to recover the information which high-order genetic perturbation data contained. Note that both the original model and each of the genotype-constrained models were trained with the same set of genotype-constrained data for estimating different sets (but same number) of parameters selected by Lasso regularization. In conclusion, high-order genetic perturbations including additional combinatorial mutations of key network components are beneficial for uncovering accurate regulatory mechanisms.

4.5.2 Models based on limited orders of perturbations and different numbers of targeted sectors

In previous modeling approaches, experimental setups related to the scope and target of measurements were generally determined by biology-experts. Despite the power of expert-driven designs, the accuracy of the obtained model is still questionable since prior knowledge is highly fragmentary in most biological applications. In Section 4.5.1, we showed that both the prediction accuracy and structural accuracy are two key factors to evaluate the fidelity of the final model. Here, we generalize the analysis on constrained models to investigate the impact of either the number of targeted components or the order of genetic perturbations on the performance of models by simultaneously considering the number of conditional perturbations. To explore the effects of these factors on modeling a complex system having an invariant structure over different MAMPs, we again fix the number of parameters required for the corresponding initial structures based on them from the original model (Figure 4-5), applied Lasso regressions with each training set incorporating differential constraints in order to choose best structure, and then estimated these selected non-zero parameters by using ordinary least square. These modeling approaches for multiple constrained models (MCMs) were the same as them for GCMs (Section 4.5.1) except for different training sets covering targeted genotypes for selected sectors of interest under specified MAMPs.

Given each MCM, we calculate Pearson correlation coefficients and their negative log10-transformed p-values by applying bagging approaches in the training set to predict the activity values of targeted sectors and the immunity levels (Appendix III. The Effects of different factors in modeling). To examine overall effects on prediction accuracies of the different factors, we computed

$$\frac{1}{N}\sum_{m=1}^{N} - \log_{10}\left(\text{p-value}_{m}^{PCC}\right). \qquad \text{Eq. 4-9}$$

where *N* is the total number of MCMs trained with the multiple constrained data covering the targeted genotypes of the selected sectors under the same MAMPs, and p-value_{m}^{PCC} is the calculated p-value of the Pearson correlation coefficient between observed and predicted data for either sector activities (heatmaps at the top, Figure 4-18(a)) or immunity levels (heatmap at the bottom, Figure 4-18(a)). For the MCMs with genotype-constraints, we also tested the predictions of both the sector activities and the immunity levels for triple mutations, a completely unused set for training the MCMs, and calculated their Pearson correlation coefficients and –log10(p-values) (Appendix III. The Effects of different factors in modeling). For the structural consistency of a set of MCMs fit to the training data for the m constraint including the orders of perturbations and the numbers of targeted MAMPs, we also calculated Jaccard index (Figure 4-18(b)), defined as

$$J_m\left(\boldsymbol{\beta}_m^{p\,\prime}, \boldsymbol{\beta}_{orig}^{p\,\prime}\right) = \frac{\left|\boldsymbol{\beta}_m^{p\,\prime} \cap \boldsymbol{\beta}_{orig}^{p\,\prime}\right|}{\left|\boldsymbol{\beta}_m^{p\,\prime} \cup \boldsymbol{\beta}_{orig}^{p\,\prime}\right|} \qquad \text{Eq. 4-10}$$

where $p \in \{\text{treatment-specific, sector-specific}\}$, $\boldsymbol{\beta}_m^{p\,\prime} = \left[\boldsymbol{\beta}_1^{pT}, \dots, \boldsymbol{\beta}_N^{p\,T}\right]$ consisting of column vectors of aligned p parameters from N MCMs with the m constraints, and $\boldsymbol{\beta}_{orig}^{p\,\prime} = \left[\boldsymbol{\beta}_{orig}^{p\;\;T}, \dots, \boldsymbol{\beta}_{orig}^{p\;\;T}\right]$ consisting of column vectors of aligned p parameters from the original model. Jaccard indexes of each of MCMs for sector-specific and treatment-specific parameters are shown in Appendix III. The Effects of different factors in modeling.

We investigate the effects of different numbers of network components modeling by training the regression model with limited data covering a smaller number of network components considered. In this analysis, we selected only a subset of data including both sector activities of the target components and log10-transformed bacterial counts under the small set of combinatorial genetic perturbations covering only the selected sectors for perturbations. For example, Figure 4-17 shows two models (JA-PAD4-SA and JA-SA models) by incorporating all the measurements associated with this small set of targeted network sectors. The significance of PCCs for the two constrained models was lower than that for the original model. In terms of prediction accuracy of sectors activities, -log$_{10}P$ for JA-PAD4-SA and JA-SA models are 32.07 and 10.95, respectively (83.58 for the original model). Regarding the immunity levels, -log$_{10}P$ for JA-PAD4-SA and JA-SA models are 23.90 and 13.37, respectively (49.73 for the original model). In addition, the structural similarity is another major factor influenced by modeling scope. The Jaccard indices of the sector-specific parameters from either JA-PAD4-SA or JA-SA model compared to the original model are 0.400 and 0.250, respectively. The same trend is observed with the treatment-specific parameters (Appendix III. The Effects of different factors in modeling). Strikingly, we observe inhibitory regulation from SA (3 hpt) to JA (9 hpt), which has been observed from other studies. Relating to the original model, the inhibitory regulation is actually an indirect effect from SA to JA via the ET sector. The models without the activity measurements of the ET sector capture only the resulting variations of JA sector activity in the case of the presence or absence of SA sector. This suggests that the activation of SA sector to JA sector is an actual (but hidden) mechanism that requires higher order perturbations to capture with a model.

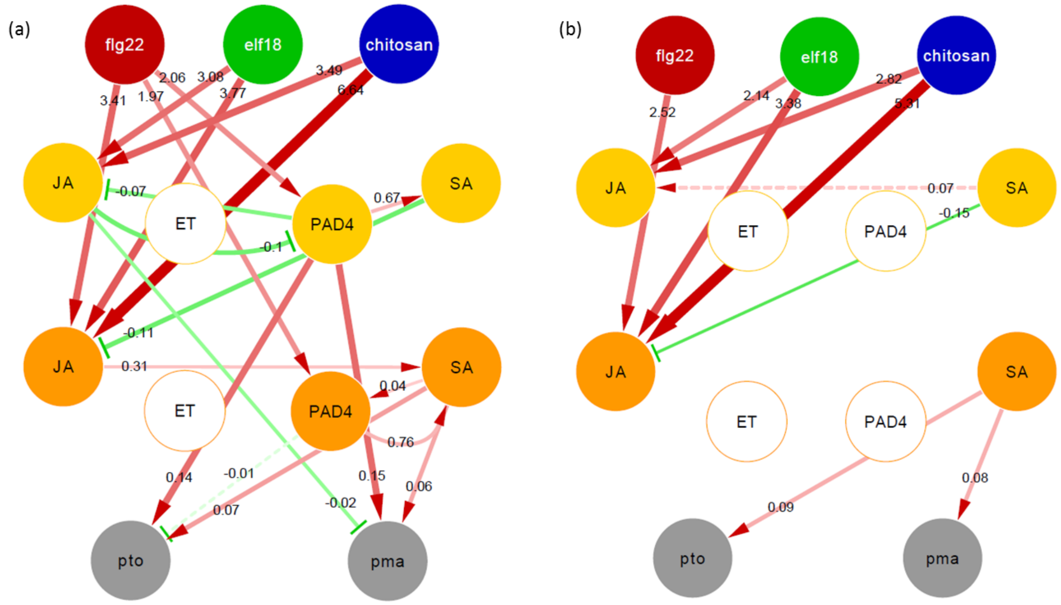


Figure 4-17 Examples of network models with limited numbers of sectors considered. (a) JA-PAD4-SA model. (b) JA-SA model. The representations are the same as in Figure 4-5. Any dashed arrows represented insignificant directional regulatory interactions. All network models with limited numbers of sectors are in Appendix IV. Network models with limited numbers of sectors considered.

Based on the number of MAMP treatments, we additionally constrained the models to examine the effects of the different numbers of both network components and MAMP treatments on the performance. As shown in the models described above, the significance of the PCC (-log$_{10}P$) between the observed and the predicted data (both sector activities and immunity levels) over different combinatorial cases decreased when a smaller number of sectors were included in the model (three columns from the right in each heatmap, Figure 4-18(a)). Likewise, the structures of the models with fewer sectors were substantially different from the original model as shown by the Jaccard Index (three columns from the right in each heatmap, Figure 4-18(b)). These results indicate that inclusion of all four sectors is important for high predictive power of the model and suggest that this inclusion also captured mechanistic relationships among the sectors more accurately when a simple model, such as linear regression, was used to explain the relationships. Admittedly, the model for limited number of network components could also reconstitute the target system capturing similar biological mechanisms and predict the activities of specified target components of interest. However, the constrained models contained only limited information of cross-communications between sectors when derived from the limited training set. Many interactions among other network components (as intermediate regulators), not considered for modeling, could be indirect or hidden. To increase the resolution of the network and bridge gaps between our prior knowledge, it was necessary to add some putative components into both network perturbation libraries and activity measurements.

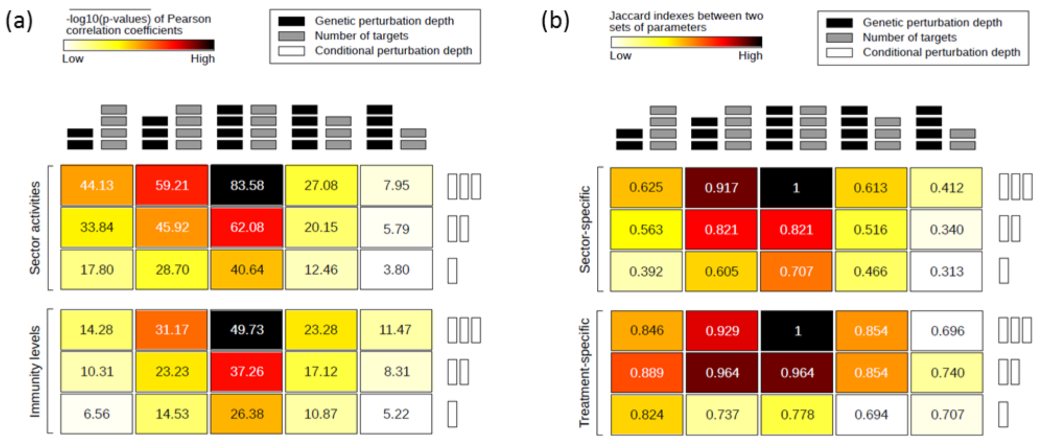


Figure 4-18 The impacts of three modeling factors including orders of genetic perturbations, depths of conditional perturbations, and numbers of targeted network components on the model performance. (a) The significance of the PCCs (-log$_{10}P$) between the predictions of different models and the observed data. Darker colors represent higher significances. The limited models were made with fewer signaling sectors (4, 3, and 2 sectors for the three right columns) or with datasets with lower orders of perturbations (up to triple, double, and single mutants for the three left columns) and with fewer MAMP treatments (3, 2, and 1 MAMP treatments, plus mock treatment, for the three rows). The average significance values for the limited models of the same

class (Eq. 4-9) are shown. The middle cell in the top row represents the original model. (b) Structural consistency across different constrained models and the original model. The Jaccard Indices between the original model and the limited models as in (a) are shown (Eq. 4-10). The significance of the PCCs including actual PCCs and the Jaccard Indices for each case of the limited models are shown in Appendix III. The Effects of different factors in modeling.

We further generalize the analysis on the effect of orders of genetic perturbations by additionally considering the number of MAMP treatments. When four-sector models were made based on datasets lacking higher orders of perturbations (Section 4.5.1), the predictive power and the structure similarity to the original model decreased and these differences between the original model and the constrained model were more significant if we considered the effects from the genetic perturbation with other factors (Figure 4-18). We similarly demonstrated the importance of having multiple MAMP inputs in the predictive power and the structure of the model (Rows of the cells in Figure 4-18).

The results from the analyses of the impacts of three modeling factors suggest that a more complete set of network components included in the model with the corresponding perturbation data could increase the ability to obtain a more realistic model with a high precisions and accurate structures. If someone models a complex system with limited efforts on biological experiments, these results would recommend that maximizing the number of the network components for measurements in the complete set of pairwise combinatorial genetic perturbations may be the first option to accomplish modeling the target network with reasonably high prediction accuracy and structural reliability. The network structure can be further refined with the additional higher order genetic perturbations designed by testable hypotheses from the initial model.

4.6 Comparison of two models from different modeling approaches: Bayesian networks vs. linear regression models

Bayesian networks build on probabilistic graphical models representing a set of random variables and their conditional dependencies via a directed acyclic graph. On the other hand, multiple regression models use linear predictor functions with multiple explanatory variables and unknown model parameters. They are highly capable with distinct merits and thus extensively used for constructing biological systems or pathways in systems biology. In chapter 3 and 4, we applied the two modeling approaches for building dynamic immune signaling network in plant and successfully reconstituted the mechanistic map of the four signaling sectors and corresponding immune responses in various conditions. Despite different modeling frameworks, these two models captured biological signals and mechanisms in similar ways. In this section, we summarize the differences in the results from the two methods and provide common insights in plant immune system based on the comparison of the two models.

4.6.1 Methodological distinction between two models

The two modeling methods have their own innate characteristics. First of all, the explanatory variables in these two models have different types. Both discrete (input and sector nodes) and continuous (output nodes) variables are incorporated in the Bayesian network approach while only continuous variables are included in the multiple regression models. To unbiasedly estimate conditional dependency between nodes in Bayesian networks, we set the number of instances from datasets across various combinatorial conditions the same by sampling them based on z-scores. All generated data were simply used in multiple regression models. Second, the Bayesian networks focused on induced sector activities and immunity levels from the three MAMPs by using only MAMP-associated data. Multiple regression models, on the other hand, were fit to not only MAMP but mock datasets and the final obtained model displayed induced network behaviors from MAMP treatments relative to them in mock. Third, due to the Bayesian network's Markov assumption, we restrict to one genotype (wild-type) for the observed activity of the four sectors at 3 hpt. On the contrary, our multiple regression formulations incorporate all combinatorial genotypes into the network model so that they are able to capture interactions between the four sectors even at the early time-point.

More importantly, different background model structures were configured in the networks on the basis of different modeling assumptions. Although both modeling approaches focus on common signaling mechanisms with the cross-communication of the four sectors regardless of system inputs, they are equipped with their own ability for capturing network behaviors differentially. First of all, Bayesian networks build on Markov properties, meaning that all the downstream nodes are only influenced by directly connected upstream nodes. For example, all signaling activities at later stages and corresponding immune responses are determined by activities of direct upstream sector nodes regardless of input signals. On the other hand, in multiple regression models, there exist additional treatment-specific parameters related to both early and late sectors. Thus, our regression-based models are more flexible for capturing MAMP-specific activities. Second, Bayesian networks genuinely support non-linearity, which is common in complex biological systems. As shown in Section 0, the comparison of outcomes given several different combinatorial states of direct upstream nodes allowed us to reveal hidden mechanisms such as compensatory roles between two sectors. To capture these interactions between sectors in linear regression formulations, we used actual activity values of other sectors as explanatory variables to predict activity values of a sector. Third, our regression model structures allow us to capture interactions within same layers by cyclic connections, which cannot be accomplished by the Bayesian network model. Fourth, self-interactions within sectors are displayed in Bayesian network models whereas, in the regression-based model, they are partially described by the difference between estimated treatment-specific parameters at 3 and 9 hpt.

Additionally, there are differences in evaluation approaches and the model complexity of the two modeling approaches. First of all, we measured prediction accuracy with cross-validation methods, which were a natural choice for our Bayesian network setting since the accurate inference of high dimensional parameters in the last two layers requires the maximal coverage of combinatorial states in its direct upstream. In multiple regression models, we instead used bootstrap aggregation approaches for measuring the prediction accuracy of the test data by random sampling. In terms of structural stability, although the selected model structures with the two modeling approaches were highly confident, the model selection criteria are different from each other. The statistical significance of all estimated parameters in our regression models was supported by the measured confidence intervals. On the contrary, only cross-talk edges could be evaluated by statistical tests for their confidence in the Bayesian network models. Regarding the complexity of modeling approaches, the number of parameters in Bayesian network models is much higher than the number in multiple regression models. For example, the number of estimated parameters in the final models from Bayesian networks and multiple regressions is 104 and 45, respectively. Most of estimated parameters in Bayesian network models are required for capturing non-linear activities.

4.6.2 Common biological insights from two models

Although the two models have many differences as described above, both models suggest similar biological mechanisms controlling the immune signaling networks (Figure 4-19).

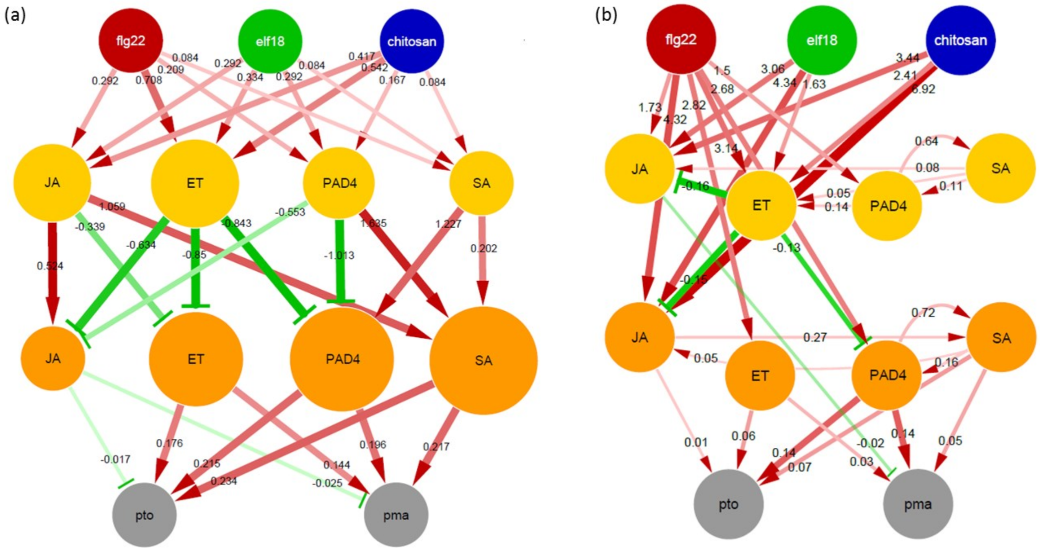


Figure 4-19 Two obtained models for plant PTI signaling networks based on (a) Bayesian network approach and (b) multiple regression model formulation

During signaling transduction, JA and ET sectors are highly activated at early stage. On the contrary, PAD4 and SA sectors are involved in many cross-communications at later stages and they directly contribute to immunity in the various combinatorial conditions. In the cross-talk layer, ET sector is the main controller for inhibiting the activity levels of both JA and PAD4 sectors. Moreover, JA and PAD4 sectors interact with SA sector, resulting in synergistic activation of the SA sector. Based on the final model from multiple regression formulation, we explained the interactions with a treatment-specific switch enabled by two bi-stable cross-activations. Interestingly, we interpreted the interaction between JA and PAD4 sectors as a compensatory effect. Because cross-talk relationships between sectors in Bayesian networks were calculated from the marginal probabilities of the sectors regardless of MAMP treatments, it is likely that the compensatory relationship between JA and PAD4 sector for activating SA sector is generally true. If there is a pathogenic attack, input patterns of which are similar to flg22 treatment, specific mechanism such as synergistic activations of SA sector can be deployed in PTI network (see Section 4.4.4 in details). The ET sector may play a key role in controlling these treatment-specific effects as a principal inhibitor. To sustain relevant immunity levels in plant in response to various attacks, PAD4 and SA sectors are major contributors to the elevated PTI levels. JA and ET sectors are more involved in indirect regulation and related network properties such that they rather indirectly contribute to the immune response.

4.7 Conclusion

Plant inducible immunity is distinct from an evolutionary perspective since pathogens not only initiate the signaling event but also rapidly evolve to attack the complex signaling network and compromise plant immunity. Multiple hormone sectors play critical roles in controlling the immunity against diverse bacteria through their interactions in a complex signaling network. In this study, we modeled the dynamics among four signaling sectors (3 hormones, jasmonic acid, ethylene, and salicylic acid, and PAD4) in the network during pattern-triggered immunity (PTI). We constructed *Arabidopsis* mutants in which all combinations of genes required for four signaling sectors were disrupted (16 combinatorial genotypes, including wild type). In each of 16 genotypes, both the expression levels of reporter genes for the signaling sector activities at two time points and the levels of PTI against two strains of bacterial pathogens were measured after treatment with MAMPs, flg22, elf18, chitosan, or mock.

Given these data, we built computational models including four signaling sectors using Bayesian networks and multiple regression, both of which have distinct merits. Although the two obtained models were built on different modeling assumptions and background structures, they captured similar biological mechanisms and network properties. In particular, our regression-based model predicted both dynamic sector activities and systematic immune responses to genetic perturbation well. The model reflects known cross-talks and reveals previously unappreciated connections between signaling sectors, some of which were confirmed in independent experiments. We also predicted the sector activities and the immunity levels with the model relatively well when the data for a particular MAMP were held out, suggesting that the model would be able to reasonably predict the sector activities and the immunity levels after treatment with other MAMPs. The model also highlights specific network motifs responsible for the robustness and tunability of the plant immune system, which are considered important properties of the network to withstand attacks from diverse and fast-evolving pathogens. More broadly, our study provides several guidelines regarding the data and computational approaches necessary for modeling a complex system from combinatorial mutant analysis.

Our immune signaling network model is a useful resource for providing valuable biological insights. Starting from the basic skeleton of plant PTI system with the four signaling sectors, we need to systematically analyze the selected marker genes for further refinement of the model. Moreover, we expand the model spatially and temporally. As shown in Figure 4-5, treatment-specific effects are dominant factors for controlling differential activities of the sectors in signal transduction. It is highly likely that there exist other related hormone sectors such as ABA affecting differential immune responses in diverse contexts. Additional sectors need to be incorporated into our network model in order to obtain higher-resolution maps and uncover hidden mechanisms. We are also aware that there are potentially more interactions downstream of the sectors. For example, SA signaling inhibits some downstream transcriptional responses requiring activation of both the JA and ET sectors [132]. By measuring the activities of a marker gene that reports SA signaling inhibition of signaling requiring both the JA and ET sector activation, we can understand dynamic behaviors of the immune signaling network more detail. More importantly, the network motif discovered here (a tetra-stable switch), which is presumably responsible for the key network properties in plant immune system, deserves further detailed analyses with computational simulation and follow-up experiments.

4.8 Supporting methods

4.8.1 Evaluation of predictive power

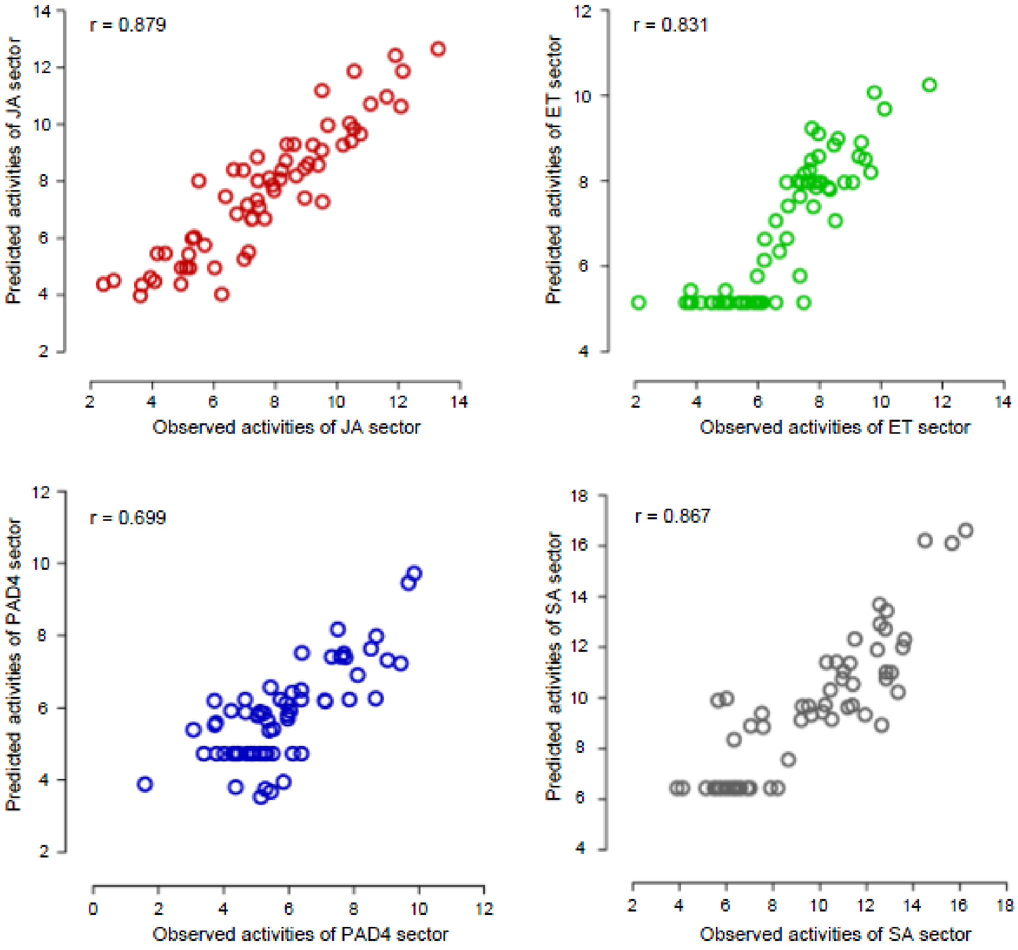


Figure 4-20 The prediction accuracy represented for each target sector.

To assess the predictive power of the network model, we also used a Bagging approach by repeatedly training the model of the selected structure (Figure 4-5) with sampled data and evaluating their performance by prediction of the out-of-bag (held-out) data. In each bootstrapping step, the treatment:genotype combinations were randomly sampled with replacement, and the data corresponding to the sampled treatment:genotype combinations were used as a training dataset. The models that were fit to the training dataset were used to predict the held-out data. This process was repeated 1000 times, and the final prediction of the sector activity value for each of 256 sector:treatment:genotype:time combinations (4 sectors x 4 treatments x 8 genotypes x 2 time points) or the bacterial count for each of 128 treatment:genotype:strain combinations (4 treatments x 16 genotypes x 2 strains) was obtained by taking the median over all predicted values for the instance. The final predicted values obtained from the Bagging procedure were then compared with the averages of the corresponding observed values by calculating a PCC between them (overall prediction accuracy of the sector activities and the immunity levels in Figure 4-6 (a) and (b), respectively, sector-specific and strain-specific prediction accuracies in Figure 4-20 and Figure 4-21, respectively).

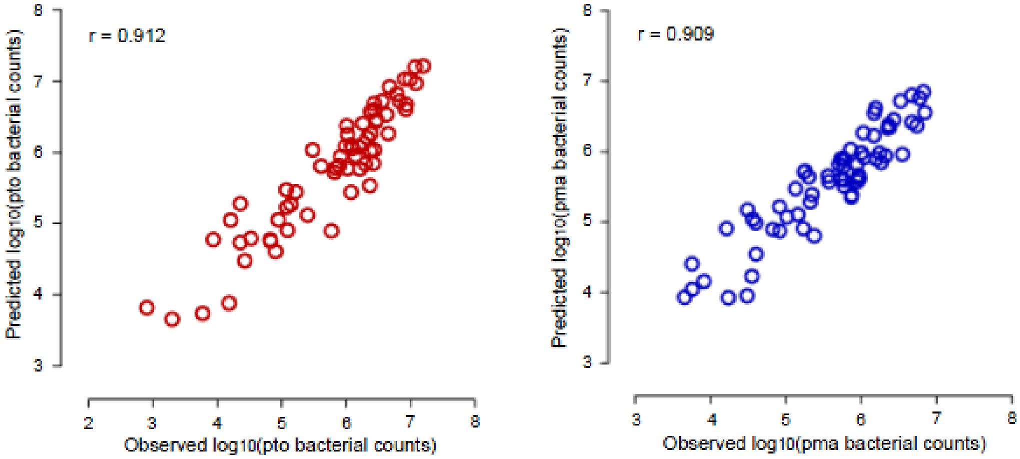


Figure 4-21 The prediction accuracy represented for each target strain.

4.8.2 A network model with different initial structures

To show the impact of cross-talk information among four sectors and treatment-specific effects on capturing the dynamic activity levels of the sectors, we first reformulated the multiple regression starting model structure by removing or adding corresponding parameters for the links or interest. For example, if we want to see the effects of cross-sector information on the activity levels of JA sector, all the terms associated with cross-talk parameters (β_{ET_3,JA_3}, β_{PAD4_3,JA_3}, β_{SA_3,JA_3}, β_{ET_3,JA_9}, β_{PAD4_3,JA_9}, β_{SA_3,JA_9}, β_{ET_9,JA_9}, β_{PAD4_9,JA_9}, β_{SA_9,JA_9}) are removed from Eq. 4-2. If we want to see the impact of the interactions between treatment nodes and a late sector node on the activity levels of JA sector instead, all the terms relating to the parameters for the interactions between treatment nodes and a late JA sector (β_{m,JA_9}, β_{f,JA_9}, β_{e,JA_9}, β_{c,JA_9}) are removed from Eq. 4-2. We further add the self-interaction terms (β_{JA_3,JA_9}) into the previous regression formulation to distinguish between within-sector and treatment-specific effects on the activity levels of JA sector. Given this as the starting model structures, the multiple regression modeling approach similar to that for the original model is applied. The PCCs between observed data and predicted data for the sector activities and immunity levels are calculated (Table 4-4).

4.8.3 Network model with noise-added data

The standard deviation of the observed data was defined as the standard deviation of the residuals when a linear model with the sector:genotype:treatment:time combinations as the fixed effect was fit to the sector activity data or when a linear model with the genotype:treatment:strain interactions as the fixed effect was fit to the immunity level data. In each different level of artificial noise, we generated 100 different sets of Gaussian noise having a zero mean and a standard deviation *k* times that of the observed data (*k* = $2^{-1}, 2^0, 2^1, 2^2, 2^3$) and added the noise to the observed data. The noise for the sector activity data and the immunity level data were generated separately since they had different standard deviations. With each set of the 100 different noise-added datasets, the similar modeling approach to the original model was applied. The structural stability of the original model (Figure 4-5) was evaluated by the cosine similarity (uncentered PCC) of the parameter estimates between each of the 100 models and the original model obtained with the data without additional noise (Figure 4-7(a)). In one case (*k* = 2), the distributions of the parameter estimates across the 100 models were compared with the parameter estimates in the original model (Figure 4-7(b)).

4.8.4 SA and JA measurements and analysis

Plants of 16 combinatorial mutants and an *fls2* mutant were used. Three well-expanded leaves per plant were infiltrated with 1 μM flg22, and the infiltrated leaves were harvested at 9 hpt and flash frozen. At the time of flg22 treatment, leaves of untreated plants were harvested for 0 hpt samples. Leaves from four plants of a same genotype were pooled for one biological sample. Three biological replicates were made from independent experiments. The frozen tissue was macerated to powder and freeze-dried. Extraction and determination of SA and JA from Arabidopsis were performed with an UPLC-MS/MS (AQITY UPLC™ System/Quattro Premier XE; Waters) with an ODS column (AQUITY UPLC BEH C18, 1.7 µm, 2.1 × 100 mm, Waters) [133]. The detailed conditions of UPLC-MS/MS are described [134].

A mixed-effects linear model with the genotype:time interaction as the fixed effect and the experiments as a random effect was fit to the SA level values that were $\log_2$-transformed for data normalization. The mean estimates of the difference between 0 hpt and 9 hpt were derived from this model and used in Figure 4-9. The same type of the model was fit to the same dataset except that the *sid2*-containing genotypes were aggregated to one genotype. This second model was used to derive the mean estimates and their standard errors to compare the difference between the genotypes in the difference between 0 and 9 hpt (difference of difference) by two-tailed *t*-test.

For the JA level, only 9-hpt data were used as many 0-hpt JA levels were below detection. Many JA level values from the genotypes containing *dde2* at 9 hpt were also below detection, and all these genotypes were aggregated into one genotype for this reason. Thus, it is very likely that the mean estimate for the aggregated *dde2*-containing genotype is overestimated. A mixed-effect linear model with the genotype as the fixed effect and the experiments as a random effect was fit to the JA level values that were $\log_2$-transformed for data normalization. The mean estimates shown in Figure 4-10 and the mean estimates and their standard errors of the difference between the genotypes were derived from the model. The latter were used to compare the difference between the genotypes by two-tailed *t*-test.

4.8.5 Predicting the sector activities and the immunity levels of a held-out treatment

If the parameter values for the links from the sectors (sector-specific parameter values) are invariant across the treatments, the sector-specific parameter values that are estimated with some MAMP treatments should be able to be used to predict the sector activities and the immunity levels after treatment with a new MAMP across the genotypes, given the parameter values for the links from the new MAMP treatment (new MAMP-specific parameter values). This notion of predictability of the sector activities and the immunity levels with new MAMP treatment was tested by: holding out the data from one of the MAMP treatments in the network modeling for estimation of the sector-specific parameter values; estimating the MAMP-specific parameter values for the held-out MAMP using the wild-type and quadruple mutant data with the held-out MAMP treatment; and predicting the sector activity and immunity level values in the rest of the genotypes with the held-out MAMP treatment.

First, the data from one of the three MAMP treatments were held out from the full dataset (i.e., the resulting dataset contains the data from two MAMP treatments and mock treatment), and the multiple regression models were fit to the dataset with one MAMP treatment held-out, using the approach similar to that used for the original model. The first step yielded the estimates for the sector-specific parameters. Second, multiple regression models in which the values for the sector-specific parameters were fixed to the estimates obtained in the first step were fit using least square to the sector activity data of the wild-type genotype and the immunity level data of the wild-type and quadruple mutant genotypes with the held-out MAMP treatment. Only the parameter estimates with $p < 0.05$ were considered, and the others were set to zero. The second step yielded the estimates for the new MAMP-specific parameters. Third, with the parameter values from the first and second steps, the following least square solution by using the estimated parameters were obtained to predict the sector activity values, $\hat{\boldsymbol{x}}$:

$$\hat{\boldsymbol{x}} = (\boldsymbol{A}^T \cdot \boldsymbol{A})^{-1} \cdot \boldsymbol{A}^T \cdot \boldsymbol{b} \qquad \text{Eq. 4-11}$$

where

$$\hat{\boldsymbol{x}} = \left[\widehat{JA}_3, \widehat{ET}_3, \widehat{PAD4}_3, \widehat{SA}_3, \widehat{JA}_9, \ \widehat{ET}_9, \widehat{PAD4}_9, \widehat{SA}_9\right]^T,$$

$$\boldsymbol{A} = \begin{bmatrix} -1 & \beta_{ET_3,JA_3} & \beta_{PAD4_3,JA_3} & \beta_{SA_3,JA_3} & 0 & 0 & 0 & 0 \\ \beta_{JA_3,ET_3} & -1 & \beta_{PAD4_3,ET_3} & \beta_{SA_3,ET_3} & 0 & 0 & 0 & 0 \\ \beta_{JA_3,PAD4_3} & \beta_{ET_3,PAD4_3} & -1 & \beta_{SA_3,PAD4_3} & 0 & 0 & 0 & 0 \\ \beta_{JA_3,SA_3} & \beta_{ET_3,SA_3} & \beta_{PAD4_3,SA_3} & -1 & 0 & 0 & 0 & 0 \\ 0 & \beta_{ET_3,JA_9} & \beta_{PAD4_3,JA_9} & \beta_{SA_3,JA_9} & -1 & \beta_{ET_9,JA_9} & \beta_{PAD4_9,JA_9} & \beta_{SA_9,JA_9} \\ \beta_{JA_3,ET_9} & 0 & \beta_{PAD4_3,ET_9} & \beta_{SA_3,ET_9} & \beta_{JA_9,ET_9} & -1 & \beta_{PAD4_9,ET_9} & \beta_{SA_9,ET_9} \\ \beta_{JA_3,PAD4_9} & \beta_{ET_3,PAD4_9} & 0 & \beta_{SA_3,PAD4_9} & \beta_{JA_9,PAD4_9} & \beta_{ET_9,PAD4_9} & -1 & \beta_{SA_9,PAD4_9} \\ \beta_{JA_3,SA_9} & \beta_{ET_3,SA_9} & \beta_{PAD4_3,SA_9} & 0 & \beta_{JA_9,SA_9} & \beta_{ET_9,SA_9} & \beta_{PAD4_9,SA_9} & -1 \end{bmatrix},$$

$$\begin{aligned} \boldsymbol{b} = [&-\beta_0^{JA} - \beta_{MAMP,JA_3},\ -\beta_0^{ET} - \beta_{MAMP,ET_3},\ -\beta_0^{PAD4} - \beta_{MAMP,PAD4_3},\ -\beta_0^{SA} \\ &- \beta_{MAMP,SA_3},\ -\beta_0^{JA} - \beta_{m,JA_9} - \beta_{MAMP,JA_9},\ -\beta_0^{ET} - \beta_{m,ET_9} \\ &- \beta_{MAMP,ET_9},\ -\beta_0^{PAD4} - \beta_{m,PAD4_9} - \beta_{MAMP,PAD4_9},\ -\beta_0^{SA} - \beta_{m,SA_9} \\ &- \beta_{MAMP,SA_9}]^T. \end{aligned}$$

After solving the equations, all obtained, predicted sector activities were adjusted by adding the difference between maximum values of observed and predicted activities of the sector in that the values of sector activities should be positive. To predict the log-transformed bacterial counts, we calculated the following linear combinations to obtain the predicted immunity level values:

$$\widehat{\boldsymbol{y}} = \begin{bmatrix} \hat{y}^{pto} \\ \hat{y}^{pma} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}^{pto} \\ \boldsymbol{\beta}^{pma} \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 1 \\ \widehat{\boldsymbol{x}} \end{bmatrix} \qquad \text{Eq. 4-12}$$

where

$$\begin{aligned} \boldsymbol{\beta}^{pto} = [&\beta_0^{pto},\ \beta_{MAMP,pto}, \beta_{JA_3,pto},\ \beta_{ET_3,pto}, \beta_{PAD4_3,pto}, \beta_{SA_3,pto}, \\ &\beta_{JA_9,pto},\ \beta_{ET_9,pto}, \beta_{PAD4_9,pto}, \beta_{SA_9,pto}], \end{aligned}$$

$$\begin{aligned} \boldsymbol{\beta}^{pma} = [&\beta_0^{pma},\ \beta_{MAMP,pma}, \beta_{JA_3,pma},\ \beta_{ET_3,pma}, \beta_{PAD4_3,pma}, \beta_{SA_3,pma}, \\ &\beta_{JA_9,pma},\ \beta_{ET_9,pma}, \beta_{PAD4_9,pma}, \beta_{SA_9,pma}], \end{aligned}$$

and $\widehat{\boldsymbol{x}}$ is a column vector with the adjusted, predicted values of the sector activities from the above. PCCs were calculated between the observed and predicted values of the sector activities and of the bacterial counts separately for each of three MAMP-held out cases (flg22, elf18, or chitosan held-out) (Figure 4-13).

4.8.6 Differential fragilities of the sectors

We defined differential fragility as the impact of removal of the signaling sector in question on the fragility, which is the phenotypic difference between the presence and absence of the secondary signaling sector. To test the effect of the ET sector on the network outcome, we first defined two fragility measurements depending on the sector genotype as follows. The fragility of *i* sector in *EIN2* (*x*-axis) and *ein2* (*y*-axis) in Figure 4-22(b) is

$$x_{i,j}^{ET} = \left|m_{gene_i,t}^{s} - m_{wildtype,t}^{s}\right|, y_{i,j}^{ET} = \left|m_{gene_i/ein2,t}^{s} - m_{ein2,t}^{s}\right| \qquad \text{Eq. 4-13}$$

where $m_{gene_i,t}^{s}$ is a mean of observed bacterial count of *s* strain (either *pto* or *pma*) with the *i* sector deletion ($gene_i \in \{dde2, pad4, sid2\}$) and treatment *t.*

The same fragility metric such as the effect of ET sector was used to examine the effect of JA, PAD4, and SA sectors like

$$x_{i,j}^{JA} = \left|m_{gene_i,t}^{s} - m_{wildtype,t}^{s}\right|, y_{i,j}^{JA} = \left|m_{gene_i/dde2,t}^{s} - m_{dde2,t}^{s}\right|, \text{ where } gene_i \in \{ein2, pad4, sid2\} \qquad \text{Eq. 4-14}$$

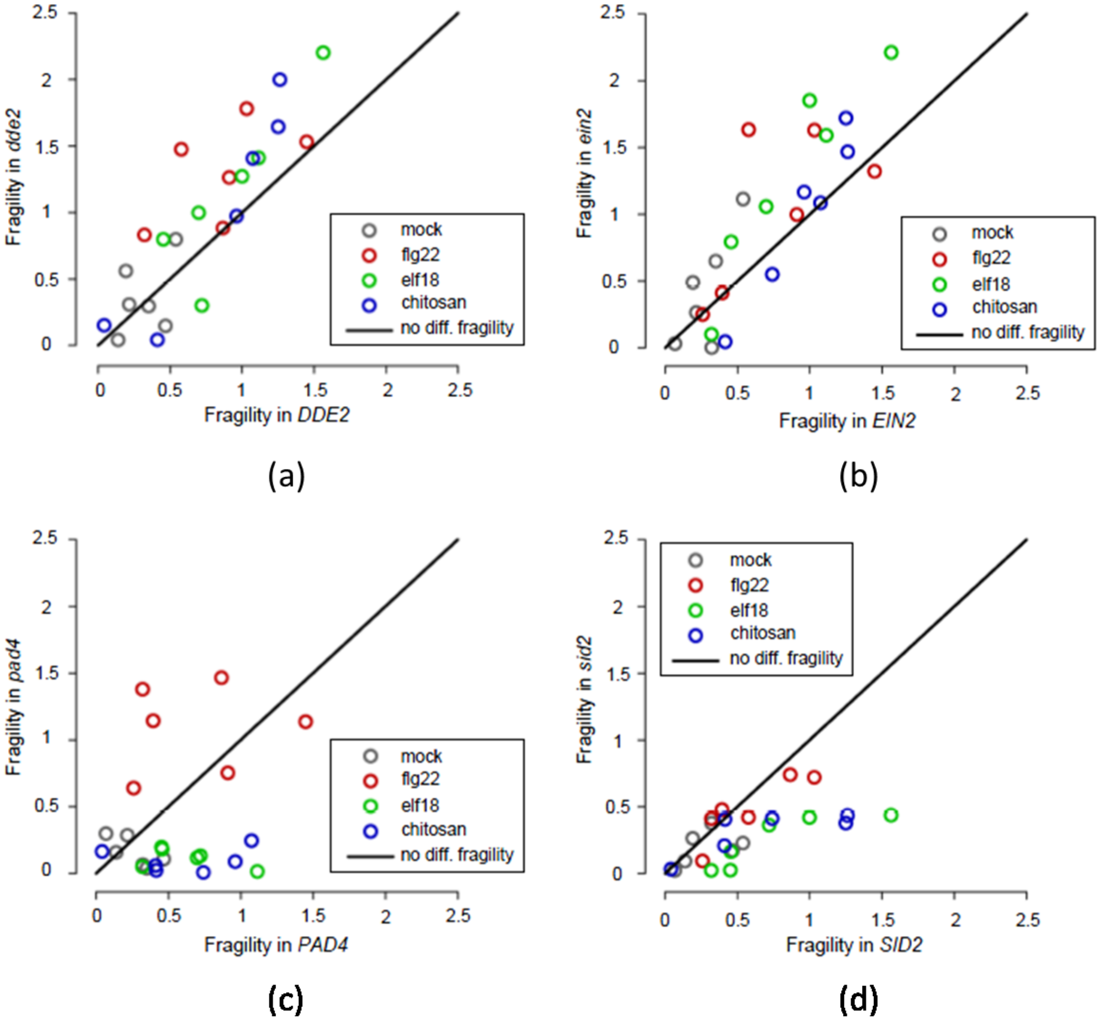
$$x_{i,j}^{PAD4} = \left|m_{gene_i,t}^{s} - m_{wildtype,t}^{s}\right|, y_{i,j}^{PAD4} = \left|m_{gene_i/pad4,t}^{s} - m_{pad4,t}^{s}\right|, \text{ where } gene_i \in \{dde2, ein2, sid2\} \qquad \text{Eq. 4-15}$$

$$x_{i,j}^{SA} = \left|m_{gene_i,t}^{s} - m_{wildtype,t}^{s}\right|, y_{i,j}^{SA} = \left|m_{gene_i/sid2,t}^{s} - m_{sid2,t}^{s}\right|, \text{ where } gene_i \in \{dde2, ein2, pad4\}. \qquad \text{Eq. 4-16}$$

The differential fragility of the *s* sector is then calculated as the average difference between two fragility measures,

$$\frac{1}{N_c N_s}\sum_{j=1}^{N_c}\sum_{i=1}^{N_s} y_{i,j}^{s} - x_{i,j}^{s} \qquad \text{Eq. 4-17}$$

where N_c is the total number of the strain:treatment combination and N_s is the total number of the other sectors. If a differential fragility is positive or negative, removal of the sector in question decreases or increases robustness, respectively; fragility is the opposite of robustness. The fragility values for each treatment:strain:secondary_sector combination in the absence vs. the presence of each sector in question are plotted in Figure 4-22. The mean and standard error of the differential fragility of each sector across the treatments, the strains, and the secondary sectors were calculated separately for the observed data and the model predictions and are shown in Figure 4-11.



(a)

(b)

(c)

(d)

Figure 4-22 Fragilities compared in the presence and the absence of each signaling sector. (a)-(d) the JA, ET, PAD4, and SA sectors, respectively. Dots above the $y = x$ line represent the cases where the fragility increases upon loss of the sector. Any case close to y = x (solid lines) has no differential fragility.

5 Conclusion

Recent progress in high-throughput bio-technology, particularly in experimental perturbations, has led to much attention to the field of systems biology. Great efforts to understand the general principles of complex biological systems have motivated the careful design of large-scale experiments to collect massive amounts of data as well as the development of the computational approaches to deal with these data. As described in this thesis, perturbation studies, while generally sharing the same approach of perturbation and measurement followed by modeling, vary dramatically in their scale and focus. For example, genome-wide studies can measure coarse phenotypes across millions of mutants while more focused approaches may collect high-resolution phenotypes for a small set of candidate genes.

In this thesis, we approached two modeling problems at opposite ends of this spectrum. In the yeast setting (Chapter 2), the generation of high-throughput genetic interaction data covering an entire yeast genome enabled the elucidation of general biological principles for the functional organization of the genome. Given the large set of genetic interactions, we designed relevant statistical approaches to re-examine previous pathway models and proposed new guidelines relating genetic and physical interactions from the unbiased data. We showed that positive genetic interactions were highly enriched between functionally distant complex-pairs, contrary to previous observations. In the model plant setting (Chapter 3 and 4), measurements of the activity levels of four signaling sectors and immunity levels for all combinatorial genetic perturbations associated with the four sectors allowed us to systematically investigate modeling approaches capturing both the dynamics and immune response in the plant. The final network model based on two different modeling approaches successfully predicted both gene expression states and immunity levels for held-out mutant data. Given our model, we uncovered hidden mechanisms between the sectors and proposed new hypotheses based on a basic building block, called a tetra-stable switch. More broadly, our study established several guidelines for experimental and computational factors involved in modeling complex biological systems.

The data space of genetic interactions will continue to expand in three different aspects. In a single organism such as *S. cerevisiae*, genetic interaction data can be generated under different conditions, in higher than 2nd order mutations, and by measuring different phenotypes other than colony-based growth rates as a proxy of fitness. Based on systematic analyses of the data with relevant statistical approaches, the multi-dimensional genetic map will provide a genetic basis connecting genotype to phenotype and a functional blueprint of cellular components in the biological system. Moreover, in other organisms, there will be more efforts on generating genetic interaction data, which has been relatively limited beyond yeast to date. Given these data, conducting cross-species comparisons could provide useful information on functional conservation of the interactions between related gene pairs, which could be a powerful tool for understanding biological systems in higher eukaryotes such as human.

Relative to our work on the plant immune system, more in-depth studies on network motifs including the tetra-stable switch, proposed as a basic architecture facilitating the network properties of robustness and tunability, are required for testing our hypotheses and establishing the specific role of this intriguing network structure. We anticipate that further study of this motif will allow for better understanding of general principles of how the PTI signaling network controls the balance between the plant's growth and optimal immune responses under diverse environmental conditions. In general, our modeling of the PTI immune network is just a starting point. Further analyses must be done with higher resolution dynamic measurements on a larger set of network components to established a more detailed, mechanistic model. A similar strategy can be followed to study effector-triggered immunity (ETI) network model, another major branch of plant immunity which we have not addressed here. A comprehensive, high-resolution of both models of plant immunity could have enormous impact, for example, in engineering plants to increase crop yields or increase disease resistance.

Bibliography

[1] R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, F. Ewen, A. R. Kerlavage, C. J. Bult, J. Tomb, B. A. Dougherty, J. M. Merrick, K. Mckenney, G. Sutton, W. Fitzhugh, C. Fields, D. Jeannie, J. Scott, R. Shirley, L. Liu, A. Glodek, J. M. Kelley, F. Janice, C. A. Phillips, T. Spriggs, E. Hedblom, and M. D. Cotton, "Whole-Genome Random Sequencing and Assembly of Haemophilus Influenzae Rd Published by : American Association for the Advancement of Science Stable URL : http://www.jstor.org/stable/2887657," vol. 269, 1995.

[2] E. a. Winzeler, "Functional Characterization of the S. cerevisiae Genome by Gene Deletion and Parallel Analysis," *Science*, vol. 285, no. 5429, pp. 901–906, Aug. 1999.

[3] K. Tsuda, M. Sato, T. Stoddard, J. Glazebrook, and F. Katagiri, "Network properties of robust immunity in plants.," *PLoS genetics*, vol. 5, no. 12, p. e1000772, Dec. 2009.

[4] A. P. Capaldi, T. Kaplan, Y. Liu, N. Habib, A. Regev, N. Friedman, and E. K. O'Shea, "Structure and function of a transcriptional network activated by the MAPK Hog1.," *Nature genetics*, vol. 40, no. 11, pp. 1300–6, Nov. 2008.

[5] E. Yeger-Lotem, L. Riva, L. J. Su, A. D. Gitler, A. G. Cashikar, O. D. King, P. K. Auluck, M. L. Geddie, J. S. Valastyan, D. R. Karger, S. Lindquist, and E. Fraenkel, "Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity.," *Nature genetics*, vol. 41, no. 3, pp. 316–23, Mar. 2009.

[6] N. Bing and I. Hoeschele, "Genetical genomics analysis of a yeast segregant population for transcription network inference.," *Genetics*, vol. 170, no. 2, pp. 533–42, Jun. 2005.

[7] D. J. Gaffney, J.-B. Veyrieras, J. F. Degner, R. Pique-Regi, A. a Pai, G. E. Crawford, M. Stephens, Y. Gilad, and J. K. Pritchard, "Dissecting the regulatory architecture of gene expression QTLs.," *Genome biology*, vol. 13, no. 1, p. R7, Jan. 2012.

[8] Y.-A. Kim, S. Wuchty, and T. M. Przytycka, "Identifying causal genes and dysregulated pathways in complex diseases.," *PLoS computational biology*, vol. 7, no. 3, p. e1001095, Mar. 2011.

[9] B. G. M. Church, "Genomes for All," *Scientific American*, vol. 294, 2006.

[10] M. L. Metzker, “Sequencing technologies - the next generation.,” *Nature reviews. Genetics*, vol. 11, no. 1, pp. 31–46, Jan. 2010.

[11] H. Y. K. Lam, M. J. Clark, R. Chen, R. Chen, G. Natsoulis, M. O’Huallachain, F. E. Dewey, L. Habegger, E. a Ashley, M. B. Gerstein, A. J. Butte, H. P. Ji, and M. Snyder, “Performance comparison of whole-genome sequencing platforms,” *Nature Biotechnology*, vol. 30, no. 1, pp. 78–82, Dec. 2011.

[12] A. Hodgkinson and A. Eyre-Walker, “Variation in the mutation rate across mammalian genomes.,” *Nature reviews. Genetics*, vol. 12, no. 11, pp. 1–11, Oct. 2011.

[13] D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, J. M. Boutell, J. Bryant, R. J. Carter, R. Keira Cheetham, A. J. Cox, D. J. Ellis, M. R. Flatbush, N. a Gormley, S. J. Humphray, L. J. Irving, M. S. Karbelashvili, S. M. Kirk, H. Li, X. Liu, K. S. Maisinger, L. J. Murray, B. Obradovic, T. Ost, M. L. Parkinson, M. R. Pratt, I. M. J. Rasolonjatovo, M. T. Reed, R. Rigatti, C. Rodighiero, M. T. Ross, A. Sabot, S. V. Sankar, A. Scally, G. P. Schroth, M. E. Smith, V. P. Smith, A. Spiridou, P. E. Torrance, S. S. Tzonev, E. H. Vermaas, K. Walter, X. Wu, L. Zhang, M. D. Alam, C. Anastasi, I. C. Aniebo, D. M. D. Bailey, I. R. Bancarz, S. Banerjee, S. G. Barbour, P. a Baybayan, V. a Benoit, K. F. Benson, C. Bevis, P. J. Black, A. Boodhun, J. S. Brennan, J. a Bridgham, R. C. Brown, A. a Brown, D. H. Buermann, A. a Bundu, J. C. Burrows, N. P. Carter, N. Castillo, M. Chiara E Catenazzi, S. Chang, R. Neil Cooley, N. R. Crake, O. O. Dada, K. D. Diakoumakos, B. Dominguez-Fernandez, D. J. Earnshaw, U. C. Egbujor, D. W. Elmore, S. S. Etchin, M. R. Ewan, M. Fedurco, L. J. Fraser, K. V. Fuentes Fajardo, W. Scott Furey, D. George, K. J. Gietzen, C. P. Goddard, G. S. Golda, P. a Granieri, D. E. Green, D. L. Gustafson, N. F. Hansen, K. Harnish, C. D. Haudenschild, N. I. Heyer, M. M. Hims, J. T. Ho, A. M. Horgan, K. Hoschler, S. Hurwitz, D. V. Ivanov, M. Q. Johnson, T. James, T. a Huw Jones, G.-D. Kang, T. H. Kerelska, A. D. Kersey, I. Khrebtukova, A. P. Kindwall, Z. Kingsbury, P. I. Kokko-Gonzales, A. Kumar, M. a Laurent, C. T. Lawley, S. E. Lee, X. Lee, A. K. Liao, J. a Loch, M. Lok, S. Luo, R. M. Mammen, J. W. Martin, P. G. McCauley, P. McNitt, P. Mehta, K. W. Moon, J. W. Mullens, T. Newington, Z. Ning, B. Ling Ng, S. M. Novo, M. J. O’Neill, M. a Osborne, A. Osnowski, O. Ostadan, L. L. Paraschos, L. Pickering, A. C. Pike, A. C. Pike, D. Chris Pinkard, D. P. Pliskin, J. Podhasky, V. J. Quijano, C. Raczy, V. H. Rae, S. R. Rawlings, A. Chiva Rodriguez, P. M. Roe, J. Rogers, M. C. Rogert Bacigalupo, N. Romanov, A. Romieu, R. K. Roth, N. J. Rourke, S. T. Ruediger, E. Rusman, R. M. Sanches-Kuiper, M. R. Schenker, J. M. Seoane, R. J. Shaw, M. K. Shiver, S. W. Short, N. L. Sizto, J. P. Sluis, M. a Smith, J. Ernest Sohna Sohna, E. J. Spence, K. Stevens, N. Sutton, L. Szajkowski, C. L. Tregidgo, G. Turcatti, S. Vandevondele, Y. Verhovsky, S. M. Virk, S. Wakelin, G. C. Walcott, J. Wang, G. J. Worsley, J. Yan,

L. Yau, M. Zuerlein, J. Rogers, J. C. Mullikin, M. E. Hurles, N. J. McCooke, J. S. West, F. L. Oaks, P. L. Lundberg, D. Klenerman, R. Durbin, and A. J. Smith, "Accurate whole human genome sequencing using reversible terminator chemistry.," *Nature*, vol. 456, no. 7218, pp. 53–9, Nov. 2008.

[14] G. Liti, D. M. Carter, A. M. Moses, J. Warringer, L. Parts, S. a James, R. P. Davey, I. N. Roberts, A. Burt, V. Koufopanou, I. J. Tsai, C. M. Bergman, D. Bensasson, M. J. T. O'Kelly, A. van Oudenaarden, D. B. H. Barton, E. Bailes, A. N. Nguyen, M. Jones, M. a Quail, I. Goodhead, S. Sims, F. Smith, A. Blomberg, R. Durbin, and E. J. Louis, "Population genomics of domestic and wild yeasts.," *Nature*, vol. 458, no. 7236, pp. 337–41, Mar. 2009.

[15] R. E. Mills, K. Walter, C. Stewart, R. E. Handsaker, K. Chen, C. Alkan, A. Abyzov, S. C. Yoon, K. Ye, R. K. Cheetham, A. Chinwalla, D. F. Conrad, Y. Fu, F. Grubert, I. Hajirasouliha, F. Hormozdiari, L. M. Iakoucheva, Z. Iqbal, S. Kang, J. M. Kidd, M. K. Konkel, J. Korn, E. Khurana, D. Kural, H. Y. K. Lam, J. Leng, R. Li, Y. Li, C.-Y. Lin, R. Luo, X. J. Mu, J. Nemesh, H. E. Peckham, T. Rausch, A. Scally, X. Shi, M. P. Stromberg, A. M. Stütz, A. E. Urban, J. A. Walker, J. Wu, Y. Zhang, Z. D. Zhang, M. A. Batzer, L. Ding, G. T. Marth, G. McVean, J. Sebat, M. Snyder, J. Wang, K. Ye, E. E. Eichler, M. B. Gerstein, M. E. Hurles, C. Lee, S. A. McCarroll, and J. O. Korbel, "Mapping copy number variation by population-scale genome sequencing.," *Nature*, vol. 470, no. 7332, pp. 59–65, Feb. 2011.

[16] M. Stoneking and J. Krause, "Learning about human population history from ancient and modern genomes," *Nature Reviews Genetics*, vol. 12, no. 9, pp. 603–614, Aug. 2011.

[17] "A map of human genome variation from population-scale sequencing," *Nature*, vol. 473, no. 7348, pp. 544–544, May 2011.

[18] T. International and H. Consortium, "A haplotype map of the human genome.," *Nature*, vol. 437, no. 7063, pp. 1299–320, Oct. 2005.

[19] W. W. Soon, M. Hariharan, and M. P. Snyder, "High-throughput sequencing for biology and medicine.," *Molecular systems biology*, vol. 9, no. 640, p. 640, Jan. 2013.

[20] J. Shendure and E. Lieberman Aiden, "The expanding scope of DNA sequencing.," *Nature biotechnology*, vol. 30, no. 11, pp. 1084–94, Nov. 2012.

[21] D. R. Rhodes and A. M. Chinnaiyan, "Integrative analysis of the cancer transcriptome.," *Nature genetics*, vol. 37 Suppl, no. June, pp. S31–7, Jun. 2005.

[22] A. D. Ewing, T. J. Ballinger, D. Earl, C. C. Harris, L. Ding, R. K. Wilson, and D. Haussler, “Retrotransposition of gene transcripts leads to structural variation in mammalian genomes.,” *Genome biology*, vol. 14, no. 3, p. R22, Mar. 2013.

[23] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad, “RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays.,” *Genome research*, vol. 18, no. 9, pp. 1509–17, Sep. 2008.

[24] B. T. Wilhelm, S. Marguerat, I. Goodhead, and J. Bähler, “Defining transcribed regions using RNA-seq.,” *Nature protocols*, vol. 5, no. 2, pp. 255–66, Jan. 2010.

[25] E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, and C. B. Burge, “Alternative isoform regulation in human tissue transcriptomes.,” *Nature*, vol. 456, no. 7221, pp. 470–6, Nov. 2008.

[26] Y. Katz, E. T. Wang, E. M. Airoldi, and C. B. Burge, “Analysis and design of RNA sequencing experiments for identifying isoform regulation.,” *Nature methods*, vol. 7, no. 12, pp. 1009–15, Dec. 2010.

[27] F. Ozsolak and P. M. Milos, “RNA sequencing: advances, challenges and opportunities.,” *Nature reviews. Genetics*, vol. 12, no. 2, pp. 87–98, Feb. 2011.

[28] C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn, and L. Pachter, “Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks,” *Nature Protocols*, vol. 7, no. 3, pp. 562–578, Mar. 2012.

[29] J. K. Kim and J. C. Marioni, “Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data.,” *Genome biology*, vol. 14, no. 1, p. R7, Jan. 2013.

[30] C. Trapnell, D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn, and L. Pachter, “Differential analysis of gene regulation at transcript resolution with RNA-seq,” *Nature Biotechnology*, vol. 31, no. 1, pp. 46–53, Dec. 2012.

[31] R. Aebersold and M. Mann, “Mass spectrometry-based proteomics.,” *Nature*, vol. 422, no. 6928, pp. 198–207, Mar. 2003.

[32] J. V. Olsen, B. Blagoev, F. Gnad, B. Macek, C. Kumar, P. Mortensen, and M. Mann, “Global, in vivo, and site-specific phosphorylation dynamics in signaling networks.,” *Cell*, vol. 127, no. 3, pp. 635–48, Nov. 2006.

[33] E. Phizicky, P. I. H. Bastiaens, H. Zhu, M. Snyder, and S. Fields, “Protein analysis on a proteomic scale.,” *Nature*, vol. 422, no. 6928, pp. 208–15, Mar. 2003.

[34] A. R. Joyce and B. Ø. Palsson, “The model organism as a system: integrating ‘omics’ data sets.,” *Nature reviews. Molecular cell biology*, vol. 7, no. 3, pp. 198–210, Mar. 2006.

[35] B. a Shoemaker and A. R. Panchenko, “Deciphering protein-protein interactions. Part I. Experimental techniques and databases.,” *PLoS computational biology*, vol. 3, no. 3, p. e42, Mar. 2007.

[36] H. Yu, P. Braun, M. a Yildirim, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, N. Li, N. Simonis, T. Hao, J.-F. Rual, A. Dricot, A. Vazquez, R. R. Murray, C. Simon, L. Tardivo, S. Tam, N. Svrzikapa, C. Fan, A.-S. de Smet, A. Motyl, M. E. Hudson, J. Park, X. Xin, M. E. Cusick, T. Moore, C. Boone, M. Snyder, F. P. Roth, A.-L. Barabási, J. Tavernier, D. E. Hill, and M. Vidal, “High-quality binary protein interaction map of the yeast interactome network.,” *Science (New York, N.Y.)*, vol. 322, no. 5898, pp. 104–10, Oct. 2008.

[37] N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A. P. Tikuisis, T. Punna, J. M. Peregrín-Alvarez, M. Shales, X. Zhang, M. Davey, M. D. Robinson, A. Paccanaro, J. E. Bray, A. Sheung, B. Beattie, D. P. Richards, V. Canadien, A. Lalev, F. Mena, P. Wong, A. Starostine, M. M. Canete, J. Vlasblom, S. Wu, C. Orsi, S. R. Collins, S. Chandran, R. Haw, J. J. Rilstone, K. Gandi, N. J. Thompson, G. Musso, P. St Onge, S. Ghanny, M. H. Y. Lam, G. Butland, A. M. Altaf-Ul, S. Kanaya, A. Shilatifard, E. O’Shea, J. S. Weissman, C. J. Ingles, T. R. Hughes, J. Parkinson, M. Gerstein, S. J. Wodak, A. Emili, and J. F. Greenblatt, “Global landscape of protein complexes in the yeast Saccharomyces cerevisiae.,” *Nature*, vol. 440, no. 7084, pp. 637–43, Mar. 2006.

[38] K. G. Guruharsha, J.-F. Rual, B. Zhai, J. Mintseris, P. Vaidya, N. Vaidya, C. Beekman, C. Wong, D. Y. Rhee, O. Cenaj, E. McKillip, S. Shah, M. Stapleton, K. H. Wan, C. Yu, B. Parsa, J. W. Carlson, X. Chen, B. Kapadia, K. VijayRaghavan, S. P. Gygi, S. E. Celniker, R. a Obar, and S. Artavanis-Tsakonas, “A protein complex network of Drosophila melanogaster.,” *Cell*, vol. 147, no. 3, pp. 690–703, Oct. 2011.

[39] S. Li, C. M. Armstrong, N. Bertin, H. Ge, S. Milstein, M. Boxem, P.-O. Vidalain, J.-D. J. Han, A. Chesneau, T. Hao, D. S. Goldberg, N. Li, M. Martinez, J.-F. Rual, P. Lamesch, L. Xu, M. Tewari, S. L. Wong, L. V. Zhang, G. F. Berriz, L. Jacotot, P. Vaglio, J. Reboul, T. Hirozane-Kishikawa, Q. Li, H. W. Gabel, A. Elewa, B. Baumgartner, D. J. Rose, H. Yu, S. Bosak, R. Sequerra, A. Fraser, S. E. Mango, W. M. Saxton, S. Strome, S. Van Den Heuvel, F. Piano, J. Vandenhaute, C. Sardet, M. Gerstein, L. Doucette-Stamm, K. C. Gunsalus, J. W. Harper, M. E. Cusick, F. P. Roth, D. E. Hill, and M. Vidal, “A map of the interactome network of the metazoan C. elegans.,” *Science (New York, N.Y.)*, vol. 303, no. 5657, pp. 540–3, Jan. 2004.

[40] U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlaff, C. Abraham, N. Bock, S. Kietzmann, A. Goedde, E. Toksöz, A. Droege, S. Krobitsch, B. Korn, W. Birchmeier, H. Lehrach, and E. E. Wanker, "A human protein-protein interaction network: a resource for annotating the proteome.," *Cell*, vol. 122, no. 6, pp. 957–68, Sep. 2005.

[41] S. J. Dixon, M. Costanzo, A. Baryshnikova, B. Andrews, and C. Boone, "Systematic mapping of genetic interaction networks.," *Annual review of genetics*, vol. 43, pp. 601–25, Jan. 2009.

[42] a H. Tong, M. Evangelista, a B. Parsons, H. Xu, G. D. Bader, N. Pagé, M. Robinson, S. Raghibizadeh, C. W. Hogue, H. Bussey, B. Andrews, M. Tyers, and C. Boone, "Systematic genetic analysis with ordered arrays of yeast deletion mutants.," *Science (New York, N.Y.)*, vol. 294, no. 5550, pp. 2364–8, Dec. 2001.

[43] X. Pan, P. Ye, D. S. Yuan, X. Wang, J. S. Bader, and J. D. Boeke, "A DNA integrity network in the yeast Saccharomyces cerevisiae.," *Cell*, vol. 124, no. 5, pp. 1069–81, Mar. 2006.

[44] B. Lehner, C. Crombie, J. Tischler, A. Fortunato, and A. G. Fraser, "Systematic mapping of genetic interactions in Caenorhabditis elegans identifies common modifiers of diverse signaling pathways.," *Nature genetics*, vol. 38, no. 8, pp. 896–903, Aug. 2006.

[45] A. Roguev, D. Talbot, G. L. Negri, M. Shales, G. Cagney, S. Bandyopadhyay, B. Panning, and N. J. Krogan, "Quantitative genetic-interaction mapping in mammalian cells," *Nature Methods*, Feb. 2013.

[46] L. Hartwell, "Genetics. Robust interactions.," *Science (New York, N.Y.)*, vol. 303, no. 5659, pp. 774–5, Feb. 2004.

[47] M. Costanzo, A. Baryshnikova, J. Bellay, Y. Kim, E. D. Spear, C. S. Sevier, H. Ding, J. L. Y. Koh, K. Toufighi, S. Mostafavi, J. Prinz, R. P. St Onge, B. VanderSluis, T. Makhnevych, F. J. Vizeacoumar, S. Alizadeh, S. Bahr, R. L. Brost, Y. Chen, M. Cokol, R. Deshpande, Z. Li, Z.-Y. Lin, W. Liang, M. Marback, J. Paw, B.-J. San Luis, E. Shuteriqi, A. H. Y. Tong, N. van Dyk, I. M. Wallace, J. a Whitney, M. T. Weirauch, G. Zhong, H. Zhu, W. a Houry, M. Brudno, S. Ragibizadeh, B. Papp, C. Pál, F. P. Roth, G. Giaever, C. Nislow, O. G. Troyanskaya, H. Bussey, G. D. Bader, A.-C. Gingras, Q. D. Morris, P. M. Kim, C. a Kaiser, C. L. Myers, B. J. Andrews, and C. Boone, "The genetic landscape of a cell.," *Science (New York, N.Y.)*, vol. 327, no. 5964, pp. 425–31, Jan. 2010.

[48] A. Baryshnikova, M. Costanzo, Y. Kim, H. Ding, J. Koh, K. Toufighi, J.-Y. Youn, J. Ou, B.-J. San Luis, S. Bandyopadhyay, M. Hibbs, D. Hess, A.-C. Gingras, G. D. Bader, O. G. Troyanskaya, G. W. Brown, B. Andrews, C. Boone, and C. L. Myers, "Quantitative analysis of fitness and genetic interactions in yeast on a genome scale.," *Nature methods*, vol. 7, no. 12, pp. 1017–24, Dec. 2010.

[49] A. H. Y. Tong, G. Lesage, G. D. Bader, H. Ding, H. Xu, X. Xin, J. Young, G. F. Berriz, R. L. Brost, M. Chang, Y. Chen, X. Cheng, G. Chua, H. Friesen, D. S. Goldberg, J. Haynes, C. Humphries, G. He, S. Hussein, L. Ke, N. Krogan, Z. Li, J. N. Levinson, H. Lu, P. Ménard, C. Munyana, A. B. Parsons, O. Ryan, R. Tonikian, T. Roberts, A.-M. Sdicu, J. Shapiro, B. Sheikh, B. Suter, S. L. Wong, L. V. Zhang, H. Zhu, C. G. Burd, S. Munro, C. Sander, J. Rine, J. Greenblatt, M. Peter, A. Bretscher, G. Bell, F. P. Roth, G. W. Brown, B. Andrews, H. Bussey, and C. Boone, "Global mapping of the yeast genetic interaction network.," *Science (New York, N.Y.)*, vol. 303, no. 5659, pp. 808–13, Feb. 2004.

[50] A. Typas, R. J. Nichols, D. A. Siegele, M. Shales, S. R. Collins, B. Lim, H. Braberg, N. Yamamoto, R. Takeuchi, B. L. Wanner, H. Mori, J. S. Weissman, N. J. Krogan, and C. A. Gross, "High-throughput , quantitative analyses of genetic interactions in E . coli," *Nature methods*, vol. 5, no. 9, pp. 781–787, 2008.

[51] F. Bohdana, S. Phanse, G. Butland, M. Babu, J. J. Dı, B. Gold, W. Yang, J. Li, A. G. Gagarinova, O. Pogoutse, H. Mori, B. L. Wanner, H. Lo, J. Wasniewski, C. Christopoulos, M. Ali, P. Venn, A. Safavi-naini, N. Sourour, S. Caron, J. Choi, L. Laigle, A. Nazarians-armavil, A. Deshpande, S. Joe, K. A. Datsenko, N. Yamamoto, B. J. Andrews, C. Boone, H. Ding, B. Sheikh, G. Moreno-hagelsieb, J. F. Greenblatt, and A. Emili, "eSGA : E . coli synthetic genetic array analysis," *Nature methods*, vol. 5, no. 9, pp. 789–795, 2008.

[52] R. Kelley and T. Ideker, "Systematic interpretation of genetic interactions using protein networks.," *Nature biotechnology*, vol. 23, no. 5, pp. 561–6, May 2005.

[53] I. Ulitsky and R. Shamir, "Pathway redundancy and protein essentiality revealed in the Saccharomyces cerevisiae interaction networks.," *Molecular systems biology*, vol. 3, p. 104, Jan. 2007.

[54] D. Fiedler, H. Braberg, M. Mehta, G. Chechik, G. Cagney, P. Mukherjee, A. C. Silva, M. Shales, S. R. Collins, S. van Wageningen, P. Kemmeren, F. C. P. Holstege, J. S. Weissman, M.-C. Keogh, D. Koller, K. M. Shokat, and N. J. Krogan, "Functional organization of the S. cerevisiae phosphorylation network.," *Cell*, vol. 136, no. 5, pp. 952–63, Mar. 2009.

[55] C. Boone, H. Bussey, and B. J. Andrews, "Exploring genetic interactions and networks with yeast.," *Nature reviews. Genetics*, vol. 8, no. 6, pp. 437–49, Jun. 2007.

[56] J. L. Hartman, J. L. H. Iv, B. Garvik, and L. Hartwell, "Principles for the Buffering of Genetic Variation," vol. 291, no. 1001, 2001.

[57] J. Glazebrook, "Contrasting mechanisms of defense against biotrophic and necrotrophic pathogens.," *Annual review of phytopathology*, vol. 43, pp. 205–27, Jan. 2005.

[58] J. D. G. Jones and J. L. Dangl, "The plant immune system.," *Nature*, vol. 444, no. 7117, pp. 323–9, Nov. 2006.

[59] S. T. Chisholm, G. Coaker, B. Day, and B. J. Staskawicz, "Host-microbe interactions: shaping the evolution of the plant immune response.," *Cell*, vol. 124, no. 4, pp. 803–14, Feb. 2006.

[60] B. Schwessinger and C. Zipfel, "News from the frontline: recent insights into PAMP-triggered immunity in plants.," *Current opinion in plant biology*, vol. 11, no. 4, pp. 389–95, Aug. 2008.

[61] F. M. Ausubel, "Are innate immune signaling pathways in plants and animals conserved?," *Nature immunology*, vol. 6, no. 10, pp. 973–9, Oct. 2005.

[62] T. Asai, G. Tena, J. Plotnikova, M. R. Willmann, W.-L. Chiu, L. Gomez-Gomez, T. Boller, F. M. Ausubel, and J. Sheen, "MAP kinase signalling cascade in Arabidopsis innate immunity.," *Nature*, vol. 415, no. 6875, pp. 977–83, Mar. 2002.

[63] C. M. J. Pieterse, A. Leon-Reyes, S. Van der Ent, and S. C. M. Van Wees, "Networking by small-molecule hormones in plant immunity.," *Nature chemical biology*, vol. 5, no. 5, pp. 308–16, May 2009.

[64] A. F. Bent and D. Mackey, "Elicitors, effectors, and R genes: the new paradigm and a lifetime supply of questions.," *Annual review of phytopathology*, vol. 45, pp. 399–436, Jan. 2007.

[65] K. E. Hammond-Kosack and J. D. Jones, "Resistance gene-dependent plant defense responses.," *The Plant cell*, vol. 8, no. 10, pp. 1773–91, Oct. 1996.

[66] R. Tenhaken, a Levine, L. F. Brisson, R. a Dixon, and C. Lamb, "Function of the oxidative burst in hypersensitive disease resistance.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 92, no. 10, pp. 4158–63, May 1995.

[67] M. A. Torres, J. L. Dangl, and J. D. G. Jones, “Arabidopsis gp91 phox homologues AtrbohD and AtrbohF are required for accumulation of reactive oxygen intermediates in the plant defense response,” vol. 99, no. 1, 2002.

[68] R. Lamb, Christopher, Lawton, Michael, Dron, Michel, Dixon, “Signals and transduction mechanisms for activation of plant defenses against microbial attack,” *Cell*, vol. 56, no. 2, pp. 215 – 224, 1989.

[69] Y. Tao, Z. Xie, W. Chen, J. Glazebrook, H. Chang, B. Han, T. Zhu, G. Zou, and F. Katagiri, “Quantitative Nature of Arabidopsis Responses during Compatible and Incompatible Interactions with the Bacterial Pathogen Pseudomonas syringae,” vol. 15, no. February, pp. 317–330, 2003.

[70] K. Sachs, O. Perez, D. Pe’er, D. A. Lauffenburger, and G. P. Nolan, “Causal protein-signaling networks derived from multiparameter single-cell data.,” *Science (New York, N.Y.)*, vol. 308, no. 5721, pp. 523–9, Apr. 2005.

[71] A.-C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L. J. Jensen, S. Bastuck, B. Dümpelfeld, A. Edelmann, M.-A. Heurtier, V. Hoffman, C. Hoefert, K. Klein, M. Hudak, A.-M. Michon, M. Schelder, M. Schirle, M. Remor, T. Rudi, S. Hooper, A. Bauer, T. Bouwmeester, G. Casari, G. Drewes, G. Neubauer, J. M. Rick, B. Kuster, P. Bork, R. B. Russell, and G. Superti-Furga, “Proteome survey reveals modularity of the yeast cell machinery.,” *Nature*, vol. 440, no. 7084, pp. 631–6, Mar. 2006.

[72] A. Beyer, S. Bandyopadhyay, and T. Ideker, “Integrating physical and genetic maps: from genomes to interaction networks.,” *Nature reviews. Genetics*, vol. 8, no. 9, pp. 699–710, Sep. 2007.

[73] H. Yu, P. Braun, M. A. Yildirim, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, N. Li, N. Simonis, T. Hao, J.-F. Rual, A. Dricot, A. Vazquez, R. R. Murray, C. Simon, L. Tardivo, S. Tam, N. Svrzikapa, C. Fan, A.-S. de Smet, A. Motyl, M. E. Hudson, J. Park, X. Xin, M. E. Cusick, T. Moore, C. Boone, M. Snyder, F. P. Roth, A.-L. Barabási, J. Tavernier, D. E. Hill, and M. Vidal, “High-quality binary protein interaction map of the yeast interactome network.,” *Science (New York, N.Y.)*, vol. 322, no. 5898, pp. 104–10, Oct. 2008.

[74] A. J. M. Walhout, “Unraveling transcription regulatory networks by protein-DNA and protein-protein interaction mapping.,” *Genome research*, vol. 16, no. 12, pp. 1445–54, Dec. 2006.

[75] D. K. Pokholok, C. T. Harbison, S. Levine, M. Cole, N. M. Hannett, T. I. Lee, G. W. Bell, K. Walker, P. A. Rolfe, E. Herbolsheimer, J. Zeitlinger, F. Lewitter, D. K.

Gifford, and R. a Young, “Genome-wide map of nucleosome acetylation and methylation in yeast.,” *Cell*, vol. 122, no. 4, pp. 517–27, Aug. 2005.

[76] B. Ren, F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. L. Volkert, C. J. Wilson, S. P. Bell, and R. a Young, “Genome-wide location and function of DNA binding proteins.,” *Science (New York, N.Y.)*, vol. 290, no. 5500, pp. 2306–9, Dec. 2000.

[77] S. L. Ooi, D. D. Shoemaker, and J. D. Boeke, “DNA helicase gene interaction network defined using synthetic lethality analyzed by microarray.,” *Nature genetics*, vol. 35, no. 3, pp. 277–86, Nov. 2003.

[78] C. Lo, U. Korf, H. Appelhans, H. Su, A. Poustka, S. Wiemann, and D. Arlt, “Combinatorial RNAi for quantitative protein SCIENCES,” vol. 104, no. 16, pp. 6579–6584, 2007.

[79] S. S. Lee, R. Y. N. Lee, A. G. Fraser, R. S. Kamath, J. Ahringer, and G. Ruvkun, “A systematic RNAi screen identifies a critical role for mitochondria in C. elegans longevity.,” *Nature genetics*, vol. 33, no. 1, pp. 40–8, Jan. 2003.

[80] D. Segrè, A. Deluna, G. M. Church, and R. Kishony, “Modular epistasis in yeast metabolism.,” *Nature genetics*, vol. 37, no. 1, pp. 77–83, Jan. 2005.

[81] S. Bandyopadhyay, R. Kelley, N. J. Krogan, and T. Ideker, “Functional maps of protein complexes from quantitative genetic interaction data.,” *PLoS computational biology*, vol. 4, no. 4, p. e1000065, Apr. 2008.

[82] R. P. St Onge, R. Mani, J. Oh, M. Proctor, E. Fung, R. W. Davis, C. Nislow, F. P. Roth, and G. Giaever, “Systematic pathway analysis using high-resolution fitness profiling of combinatorial gene deletions.,” *Nature genetics*, vol. 39, no. 2, pp. 199–206, Feb. 2007.

[83] A. Battle, M. C. Jonikas, P. Walter, J. S. Weissman, and D. Koller, “Automated identification of pathways from quantitative genetic interaction data.,” *Molecular systems biology*, vol. 6, p. 379, Jun. 2010.

[84] R. Mani, R. P. St Onge, J. L. Hartman, G. Giaever, and F. P. Roth, “Defining genetic interaction.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 9, pp. 3461–6, Mar. 2008.

[85] M. C. Jonikas, S. R. Collins, V. Denic, E. Oh, E. M. Quan, V. Schmid, J. Weibezahn, B. Schwappach, P. Walter, J. S. Weissman, and M. Schuldiner, “Comprehensive characterization of genes required for protein folding in the

endoplasmic reticulum.," *Science (New York, N.Y.)*, vol. 323, no. 5922, pp. 1693–7, Mar. 2009.

[86] M. C. Bassik, M. Kampmann, R. J. Lebbink, S. Wang, M. Y. Hein, I. Poser, J. Weibezahn, M. a Horlbeck, S. Chen, M. Mann, A. a Hyman, E. M. Leproust, M. T. McManus, and J. S. Weissman, "A systematic mammalian genetic interaction map reveals pathways underlying ricin susceptibility.," *Cell*, vol. 152, no. 4, pp. 909–22, Feb. 2013.

[87] A. Baryshnikova, M. Costanzo, C. L. Myers, B. Andrews, and C. Boone, "Genetic Interaction Networks: Toward an Understanding of Heritability.," *Annual review of genomics and human genetics*, no. June, pp. 1–23, Jun. 2013.

[88] L. Avery and S. Wasserman, "Ordering gene function: the interpretation of epistasis in regulatory hierarchies.," *Trends in genetics : TIG*, vol. 8, no. 9, pp. 312–6, Sep. 1992.

[89] B. L. Drees, V. Thorsson, G. W. Carter, A. W. Rives, M. Z. Raymond, I. Avila-Campillo, P. Shannon, and T. Galitski, "Derivation of genetic interaction networks from quantitative phenotype data.," *Genome biology*, vol. 6, no. 4, p. R38, Jan. 2005.

[90] N. Van Driessche, J. Demsar, E. O. Booth, P. Hill, P. Juvan, B. Zupan, A. Kuspa, and G. Shaulsky, "Epistasis analysis with global transcriptional phenotypes.," *Nature genetics*, vol. 37, no. 5, pp. 471–7, May 2005.

[91] D. K. Breslow, D. M. Cameron, S. R. Collins, M. Schuldiner, J. Stewart-ornstein, H. W. Newman, S. Braun, H. D. Madhani, N. J. Krogan, and J. S. Weissman, "A comprehensive strategy enabling high-resolution functional analysis of the yeast genome," *Nature Methods*, vol. 5, no. 8, 2008.

[92] K. Tarassov, V. Messier, C. R. Landry, S. Radinovic, M. M. Serna Molina, I. Shames, Y. Malitskaya, J. Vogel, H. Bussey, and S. W. Michnick, "An in vivo map of the yeast protein interactome.," *Science (New York, N.Y.)*, vol. 320, no. 5882, pp. 1465–70, Jun. 2008.

[93] D. Ungar, T. Oka, M. Krieger, and F. M. Hughson, "Retrograde transport on the COG railway.," *Trends in cell biology*, vol. 16, no. 2, pp. 113–20, Feb. 2006.

[94] W. Xu, F. J. Smith, R. Subaran, and A. P. Mitchell, "Multivesicular Body-ESCRT Components Function in pH Response Regulation in Saccharomyces cerevisiae and Candida albicans," vol. 15, no. December, pp. 5528–5537, 2004.

[95] A. P. Mitchell, “A VAST staging area for regulatory proteins.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 20, pp. 7111–2, May 2008.

[96] M. Hayashi, T. Fukuzawa, H. Sorimachi, and T. Maeda, “Constitutive Activation of the pH-Responsive Rim101 Pathway in Yeast Mutants Defective in Late Steps of the MVB / ESCRT Pathway,” vol. 25, no. 21, pp. 9478–9490, 2005.

[97] K. Rothfels, J. C. Tanny, H. Friesen, C. Commisso, and J. Segall, “Components of the ESCRT Pathway , DFG16 , and YGR122w Are Required for Rim101 To Act as a Corepressor with Nrg1 at the Negative Regulatory Element of the DIT1 Gene of Saccharomyces cerevisiae,” vol. 25, no. 15, pp. 6772–6788, 2005.

[98] H. A. Kemp and G. F. Sprague, “Far3 and Five Interacting Proteins Prevent Premature Recovery from Pheromone Arrest in the Budding Yeast Saccharomyces cerevisiae,” vol. 23, no. 5, pp. 1750–1763, 2003.

[99] M. Goudreault, L. M. D’Ambrosio, M. J. Kean, M. J. Mullin, B. G. Larsen, A. Sanchez, S. Chaudhry, G. I. Chen, F. Sicheri, A. I. Nesvizhskii, R. Aebersold, B. Raught, and A.-C. Gingras, “A PP2A phosphatase high density interaction network identifies a novel striatin-interacting phosphatase and kinase complex linked to the cerebral cavernous malformation 3 (CCM3) protein.,” *Molecular & cellular proteomics : MCP*, vol. 8, no. 1, pp. 157–71, Jan. 2009.

[100] a H. Tong, M. Evangelista, a B. Parsons, H. Xu, G. D. Bader, N. Pagé, M. Robinson, S. Raghibizadeh, C. W. Hogue, H. Bussey, B. Andrews, M. Tyers, and C. Boone, “Systematic genetic analysis with ordered arrays of yeast deletion mutants.,” *Science (New York, N.Y.)*, vol. 294, no. 5550, pp. 2364–8, Dec. 2001.

[101] C. L. Myers, D. Robson, A. Wible, M. A. Hibbs, C. Chiriac, C. L. Theesfeld, K. Dolinski, and O. G. Troyanskaya, “Discovery of biological networks from diverse functional genomic data.,” *Genome biology*, vol. 6, no. 13, p. R114, Jan. 2005.

[102] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, “BioGRID: a general repository for interaction datasets.,” *Nucleic acids research*, vol. 34, no. Database issue, pp. D535–9, Jan. 2006.

[103] S. Pu, J. Wong, B. Turner, E. Cho, and S. J. Wodak, “Up-to-date catalogues of yeast protein complexes.,” *Nucleic acids research*, vol. 37, no. 3, pp. 825–31, Feb. 2009.

[104] F. Katagiri and K. Tsuda, “Understanding the plant immune system.,” *Molecular plant-microbe interactions : MPMI*, vol. 23, no. 12, pp. 1531–6, Dec. 2010.

[105] C. Zipfel, G. Kunze, D. Chinchilla, A. Caniard, J. D. G. Jones, T. Boller, and G. Felix, “Perception of the Bacterial PAMP EF-Tu by the Receptor EFR Restricts Agrobacterium-Mediated Transformation,” *Cell*, vol. 125, no. 4, pp. 749–760, May 2006.

[106] A. Miya, P. Albert, T. Shinya, Y. Desaki, K. Ichimura, K. Shirasu, Y. Narusaka, N. Kawakami, H. Kaku, and N. Shibuya, “CERK1, a LysM receptor kinase, is essential for chitin elicitor signaling in Arabidopsis.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 49, pp. 19613–8, Dec. 2007.

[107] D. Chinchilla, Z. Bauer, M. Regenass, T. Boller, and G. Felix, “The Arabidopsis Receptor Kinase FLS2 Binds flg22 and Determines the Specificity of Flagellin Perception,” vol. 18, no. February, pp. 465–476, 2006.

[108] M. C. Wildermuth, J. Dewdney, G. Wu, and F. M. Ausubel, “Isochorismate synthase is required to synthesize salicylic acid for plant defence.,” *Nature*, vol. 414, no. 6863, pp. 562–5, Nov. 2001.

[109] J.-H. Park, R. Halitschke, H. B. Kim, I. T. Baldwin, K. a Feldmann, and R. Feyereisen, “A knock-out mutation in allene oxide synthase results in male sterility and defective wound signal transduction in Arabidopsis due to a block in jasmonic acid biosynthesis.,” *The Plant journal : for cell and molecular biology*, vol. 31, no. 1, pp. 1–12, Jul. 2002.

[110] J. M. Alonso, “EIN2, a Bifunctional Transducer of Ethylene and Stress Responses in Arabidopsis,” *Science*, vol. 284, no. 5423, pp. 2148–2152, Jun. 1999.

[111] D. Jirage, T. L. Tootle, T. L. Reuber, L. N. Frost, B. J. Feys, J. E. Parker, F. M. Ausubel, and J. Glazebrook, “Arabidopsis thaliana PAD4 encodes a lipase-like gene that is important for salicylic acid signaling.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 23, pp. 13583–8, Nov. 1999.

[112] N. F. Daphne Koller, *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, 2009.

[113] J. Friedman, T. Hastie, and R. Tibshirani, “Journal of Statistical Software,” vol. 33, no. 1, 2010.

[114] J. Hastie, Trevor; Tibshirani, Robert; Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer, 2009.

[115] S. C. J. J. Siegel, *Nonparametric Statistics for the Behavioral Sciences*, 2nd ed. McGraw-Hill, 1988.

[116] D. Heckerman, “A Tutorial on Learning With Bayesian Networks,” *TechReport, Microsoft Research*, vol. MSR-TR-95-, p. 57, 1995.

[117] K. P. Murphy, “Dynamic Bayesian Networks: Representation, Inference and Learning,” *doctoral dissertation, University of California, Berkeley*, 2002.

[118] A. Silipo, G. Erbs, T. Shinya, J. M. Dow, M. Parrilli, R. Lanzetta, N. Shibuya, M.-A. Newman, and A. Molinaro, “Glyco-conjugates as elicitors or suppressors of plant innate immunity.,” *Glycobiology*, vol. 20, no. 4, pp. 406–19, Jan. 2010.

[119] K. Tsuda, Y. Qi, L. V. Nguyen, G. Bethke, Y. Tsuda, J. Glazebrook, and F. Katagiri, “An efficient Agrobacterium-mediated transient transformation of Arabidopsis.,” *The Plant journal : for cell and molecular biology*, vol. 69, no. 4, pp. 713–9, Feb. 2012.

[120] T. Boller and G. Felix, “A renaissance of elicitors: perception of microbe-associated molecular patterns and danger signals by pattern-recognition receptors.,” *Annual review of plant biology*, vol. 60, pp. 379–406, Jan. 2009.

[121] J. Shah, “The salicylic acid loop in plant defense,” *Current Opinion in Plant Biology*, vol. 6, no. 4, pp. 365–371, Aug. 2003.

[122] J. Glazebrook, W. Chen, B. Estes, H.-S. Chang, C. Nawrath, J.-P. Métraux, T. Zhu, and F. Katagiri, “Topology of the network integrating salicylate and jasmonate signal transduction derived from global expression phenotyping.,” *The Plant journal : for cell and molecular biology*, vol. 34, no. 2, pp. 217–28, Apr. 2003.

[123] a C. Vlot, D. A. Dempsey, and D. F. Klessig, “Salicylic Acid, a multifaceted hormone to combat disease.,” *Annual review of phytopathology*, vol. 47, pp. 177–206, Jan. 2009.

[124] M. Wiermer, B. J. Feys, and J. E. Parker, “Plant immunity: the EDS1 regulatory node.,” *Current opinion in plant biology*, vol. 8, no. 4, pp. 383–9, Aug. 2005.

[125] M. Sato, K. Tsuda, L. Wang, J. Coller, Y. Watanabe, J. Glazebrook, and F. Katagiri, “Network modeling reveals prevalent negative regulatory relationships between signaling sectors in Arabidopsis immune signaling.,” *PLoS pathogens*, vol. 6, no. 7, p. e1001011, Jan. 2010.

[126] C. Zipfel, S. Robatzek, L. Navarro, E. J. Oakeley, J. D. G. Jones, G. Felix, and T. Boller, “Bacterial disease resistance in Arabidopsis through flagellin perception.,” *Nature*, vol. 428, no. 6984, pp. 764–7, Apr. 2004.

[127] P. Singh, Y.-C. Kuo, S. Mishra, C.-H. Tsai, C.-C. Chien, C.-W. Chen, M. Desclos-Theveniau, P.-W. Chu, B. Schulze, D. Chinchilla, T. Boller, and L. Zimmerli, “The lectin receptor kinase-VI.2 is required for priming and positively regulates Arabidopsis pattern-triggered immunity.,” *The Plant cell*, vol. 24, no. 3, pp. 1256–70, Mar. 2012.

[128] Y. Saijo, N. Tintor, X. Lu, P. Rauf, K. Pajerowska-Mukhtar, H. Häweker, X. Dong, S. Robatzek, and P. Schulze-Lefert, “Receptor quality control in the endoplasmic reticulum for plant innate immunity.,” *The EMBO journal*, vol. 28, no. 21, pp. 3439–49, Nov. 2009.

[129] S. Lacombe, A. Rougon-Cardoso, E. Sherwood, N. Peeters, D. Dahlbeck, H. P. van Esse, M. Smoker, G. Rallapalli, B. P. H. J. Thomma, B. Staskawicz, J. D. G. Jones, and C. Zipfel, “Interfamily transfer of a plant pattern-recognition receptor confers broad-spectrum bacterial resistance.,” *Nature biotechnology*, vol. 28, no. 4, pp. 365–9, May 2010.

[130] R. Willmann, H. M. Lajunen, G. Erbs, M. Newman, D. Kolb, and K. Tsuda, “mediate bacterial peptidoglycan sensing and immunity to bacterial infection,” pp. 1–6, 2011.

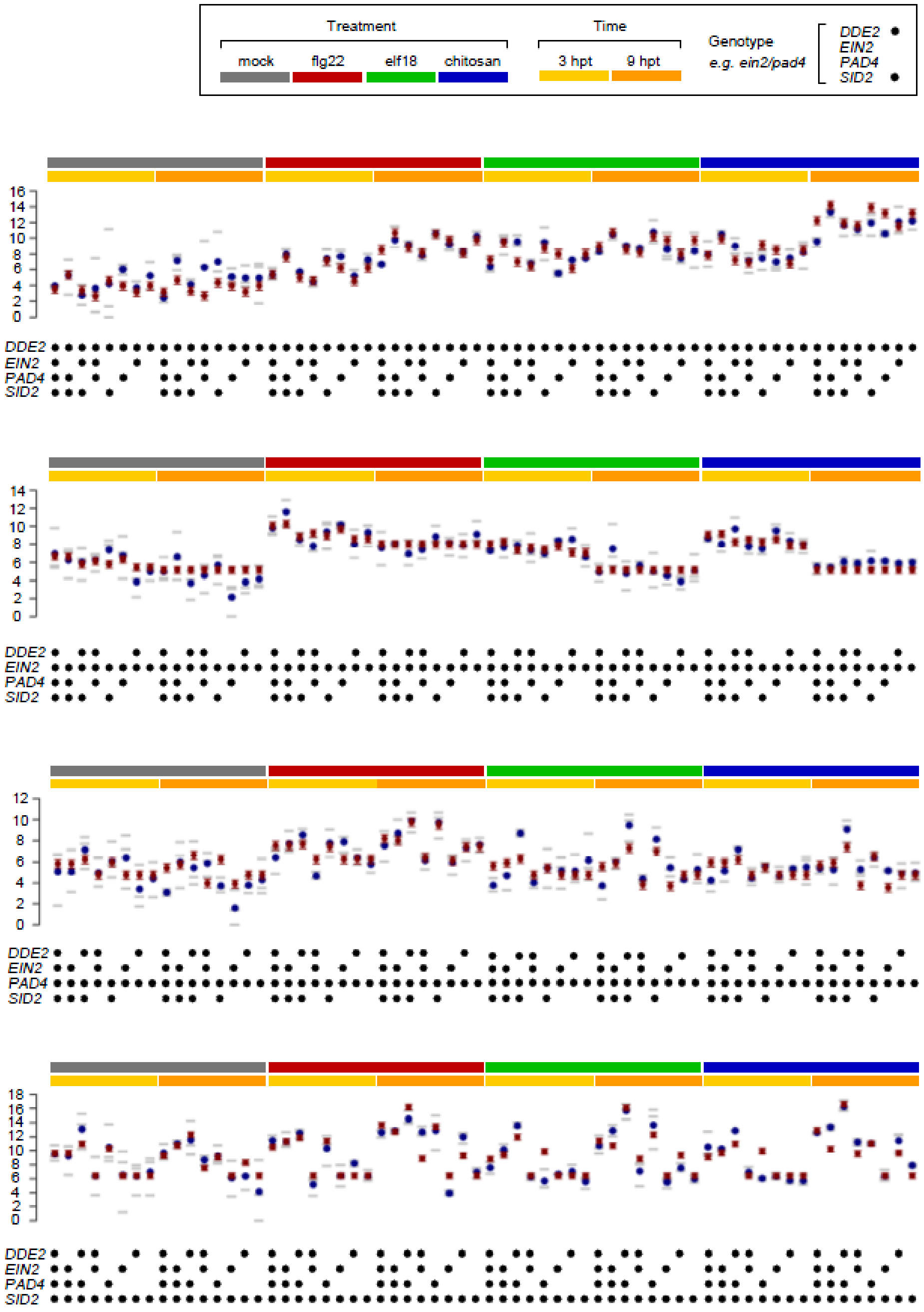
[131] U. Alon, *An Introduction to Systems Biology: Design Principles of Biological Circuits*, 1st ed. Chapman & Hall/CRC Mathematical & Computational Biology, 2006.

[132] D. Van der Does, A. Leon-Reyes, A. Koornneef, M. C. Van Verk, N. Rodenburg, L. Pauwels, A. Goossens, A. P. Körbes, J. Memelink, T. Ritsema, S. C. M. Van Wees, and C. M. J. Pieterse, “Salicylic acid suppresses jasmonic acid signaling downstream of SCFCOI1-JAZ by targeting GCC promoter motifs via transcription factor ORA59.,” *The Plant cell*, vol. 25, no. 2, pp. 744–61, Feb. 2013.

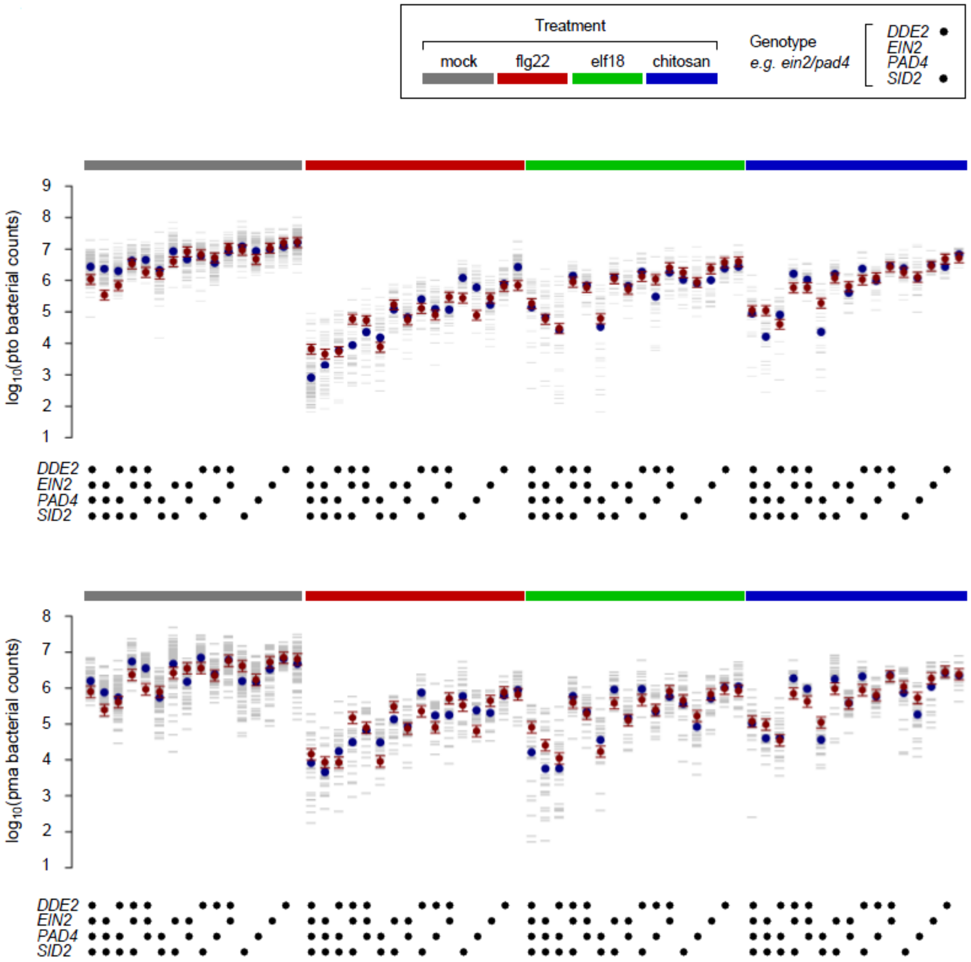
[133] M. Kojima and H. Sakakibara, “High-Throughput Phenotyping in Plants,” vol. 918, 2012.

[134] M. Kojima, T. Kamada-Nobusada, H. Komatsu, K. Takei, T. Kuroha, M. Mizutani, M. Ashikari, M. Ueguchi-Tanaka, M. Matsuoka, K. Suzuki, and H. Sakakibara, “Highly sensitive and high-throughput analysis of plant hormones using MS-probe modification and liquid chromatography-tandem mass spectrometry: an application for hormone profiling in Oryza sativa.,” *Plant & cell physiology*, vol. 50, no. 7, pp. 1201–14, Jul. 2009.

Appendix I. The fitted values of final multiple regression models in plant immune signaling network

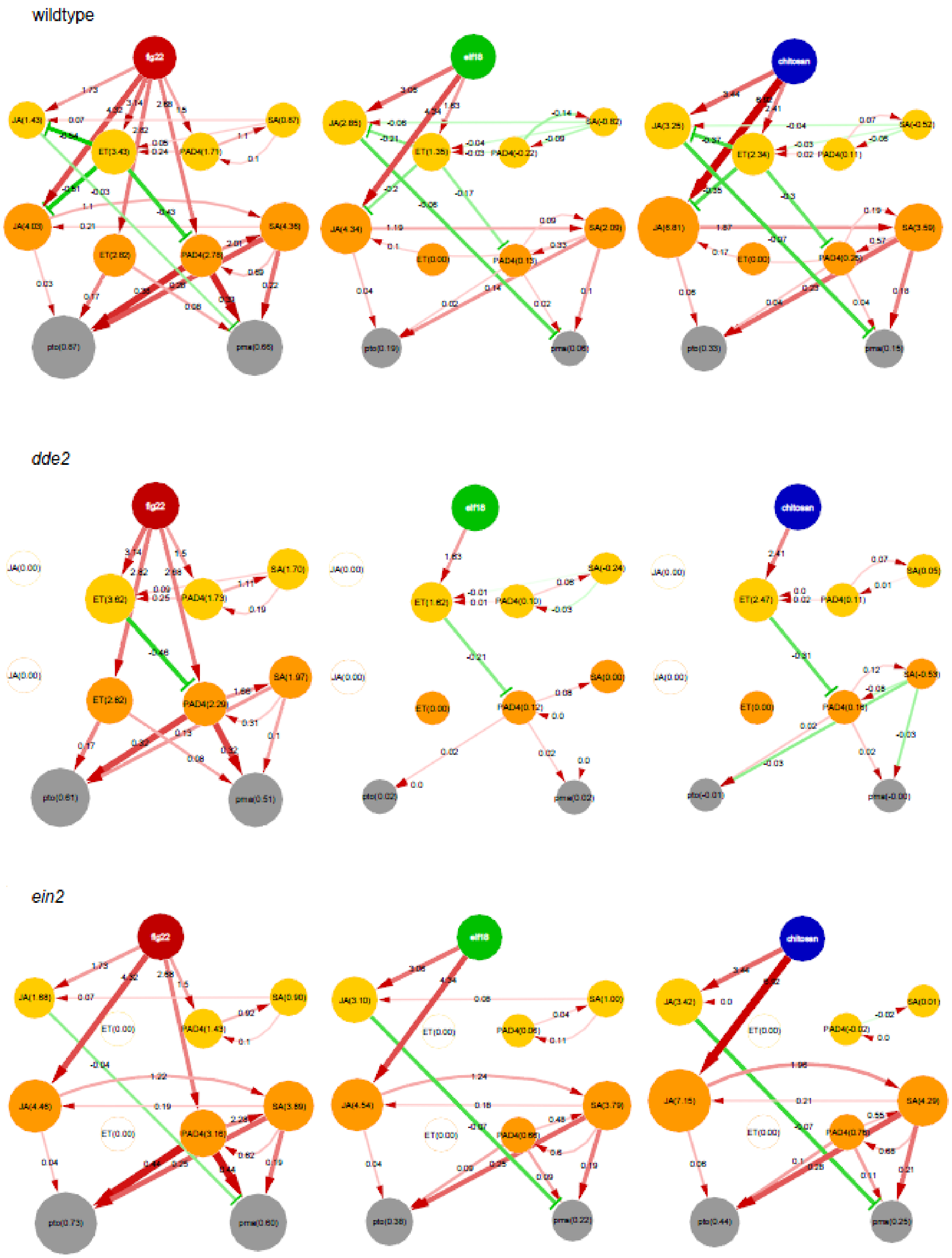


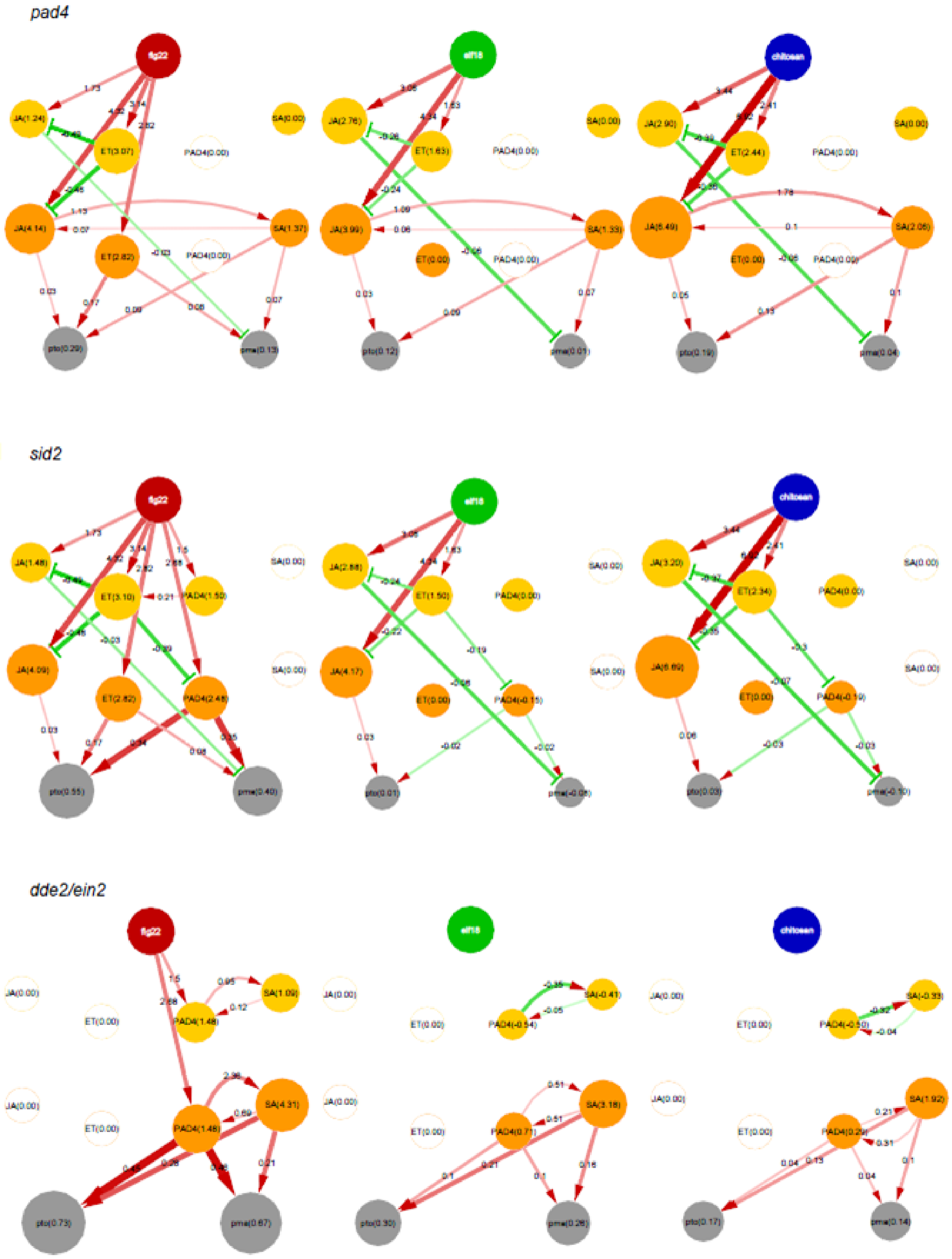
The fitted values of the final multiple regression models for the sector activities for the genotype:treatment:time:sector combinations. The wild-type alleles in the genotype are shown by black dots. The treatment and the time are color-coded as shown at the top. Each plot corresponds to each indicated sector. Gray bar, observed sector activity value; blue dot, mean of the observations; brown dot, mean estimate and its 95% confidence interval according to the final model.



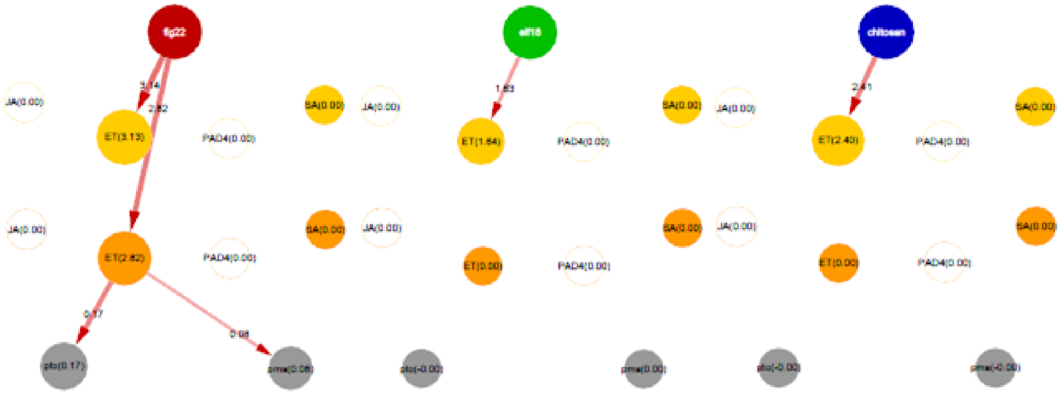
The fitted values of the final multiple regression models for the immunity levels for the genotype:treatment:strain combinations. The representations are the same as in Figure 3-10. Each plot shows $\log_{10}$-transformed bacterial counts.

Appendix II. All activity maps for plant PTI network

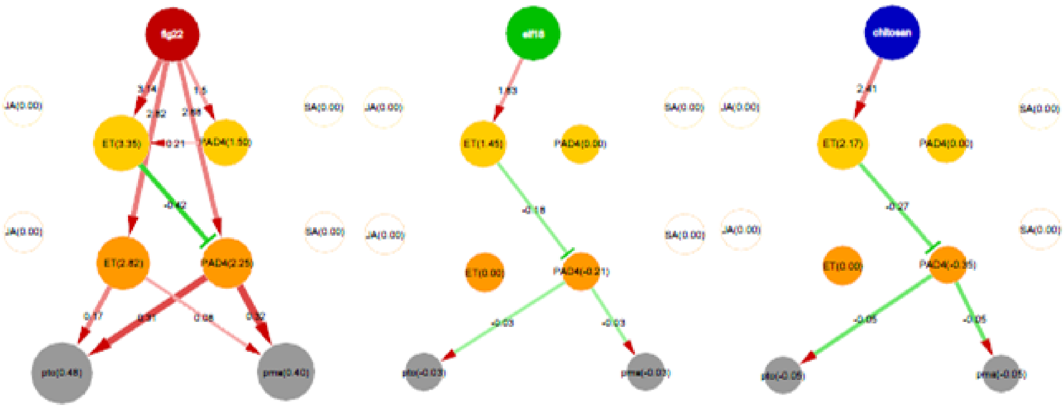



pad4
flg22
1.73
4.32
3.14
2.82
JA(1.24)
-0.49
ET(3.07)
PAD4(0.00)
SA(0.00)
-2.46
1.13
JA(4.14)
0.07
SA(1.37)
ET(2.82)
-0.03
PAD4(0.00)
0.03
0.17
0.09
0.08
0.07
pto(0.29)
pma(0.13)
elf18
3.05
4.34
1.63
JA(2.76)
-0.26
ET(1.63)
PAD4(0.00)
SA(0.00)
-0.24
1.09
JA(3.99)
0.06
SA(1.33)
ET(0.00)
-0.06
PAD4(0.00)
0.03
0.09
0.07
pto(0.12)
pma(0.01)
chitosan
3.44
6.92
2.41
JA(2.90)
-0.39
ET(2.44)
PAD4(0.00)
SA(0.00)
-0.36
1.78
JA(6.49)
0.1
SA(2.06)
ET(0.00)
-0.06
PAD4(0.00)
0.05
0.13
0.1
pto(0.19)
pma(0.04)
sid2
flg22
1.73
4.32
3.14
2.82
1.5
2.88
JA(1.48)
-0.49
ET(3.10)
0.21
PAD4(1.50)
SA(0.00)
-2.46
-0.03
-0.39
JA(4.09)
ET(2.82)
PAD4(2.48)
SA(0.00)
0.03
0.17
0.34
0.08
0.35
pto(0.55)
pma(0.40)
elf18
3.05
4.34
1.63
JA(2.88)
-0.24
ET(1.50)
PAD4(0.00)
SA(0.00)
-0.22
-0.19
JA(4.17)
-0.06
ET(0.00)
PAD4(-0.15)
SA(0.00)
0.03
-0.02
-0.02
pto(0.01)
pma(-0.08)
chitosan
3.44
6.92
2.41
JA(3.20)
-0.37
ET(2.34)
PAD4(0.00)
SA(0.00)
-0.35
-0.3
JA(6.69)
-0.07
ET(0.00)
PAD4(-0.19)
SA(0.00)
0.06
-0.03
-0.03
pto(0.03)
pma(-0.10)
dde2/ein2
flg22
1.5
2.88
0.95
SA(1.09)
JA(0.00)
ET(0.00)
0.12
PAD4(1.48)
2.98
SA(4.31)
0.69
JA(0.00)
ET(0.00)
PAD4(1.48)
0.45
0.28
0.48
0.21
pto(0.73)
pma(0.67)
elf18
JA(0.00)
-0.35
SA(-0.41)
-0.05
ET(0.00)
PAD4(-0.54)
0.51
SA(3.18)
0.51
JA(0.00)
ET(0.00)
PAD4(0.71)
0.1
0.21
0.1
0.16
pto(0.30)
pma(0.26)
chitosan
JA(0.00)
SA(-0.33)
-0.32
ET(0.00)
PAD4(-0.50)
-0.04
SA(1.92)
0.21
JA(0.00)
ET(0.00)
PAD4(0.29)
0.31
0.04
0.13
0.04
0.1
pto(0.17)
pma(0.14)

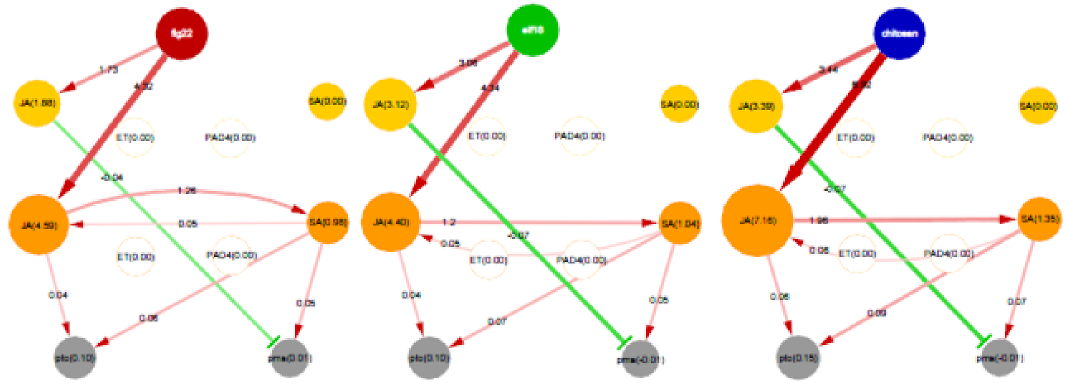
dde2/pad4



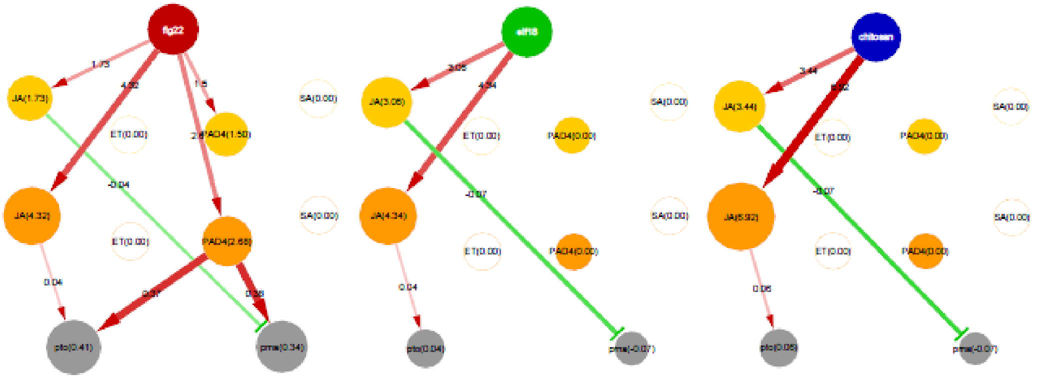
dde2/sid2



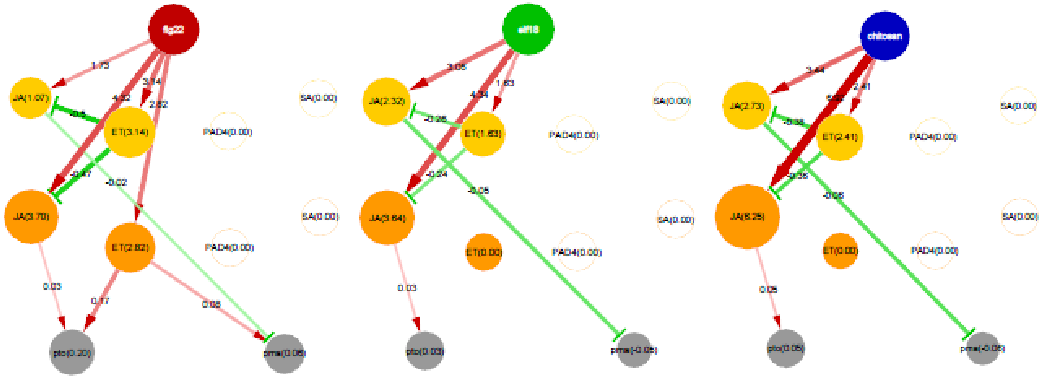
ein2/pad4



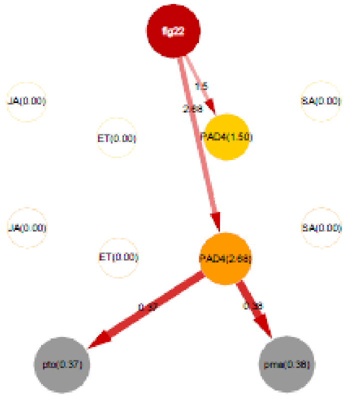
ein2/sid2

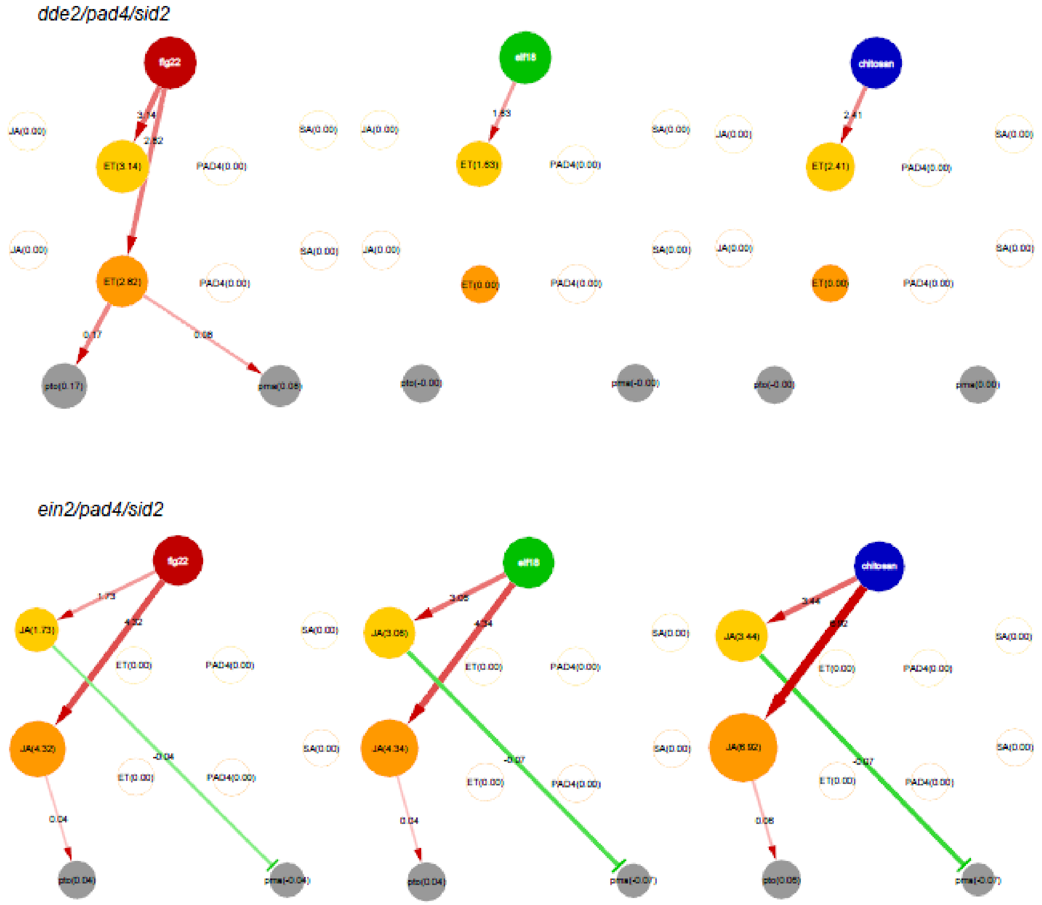


pad4/sid2

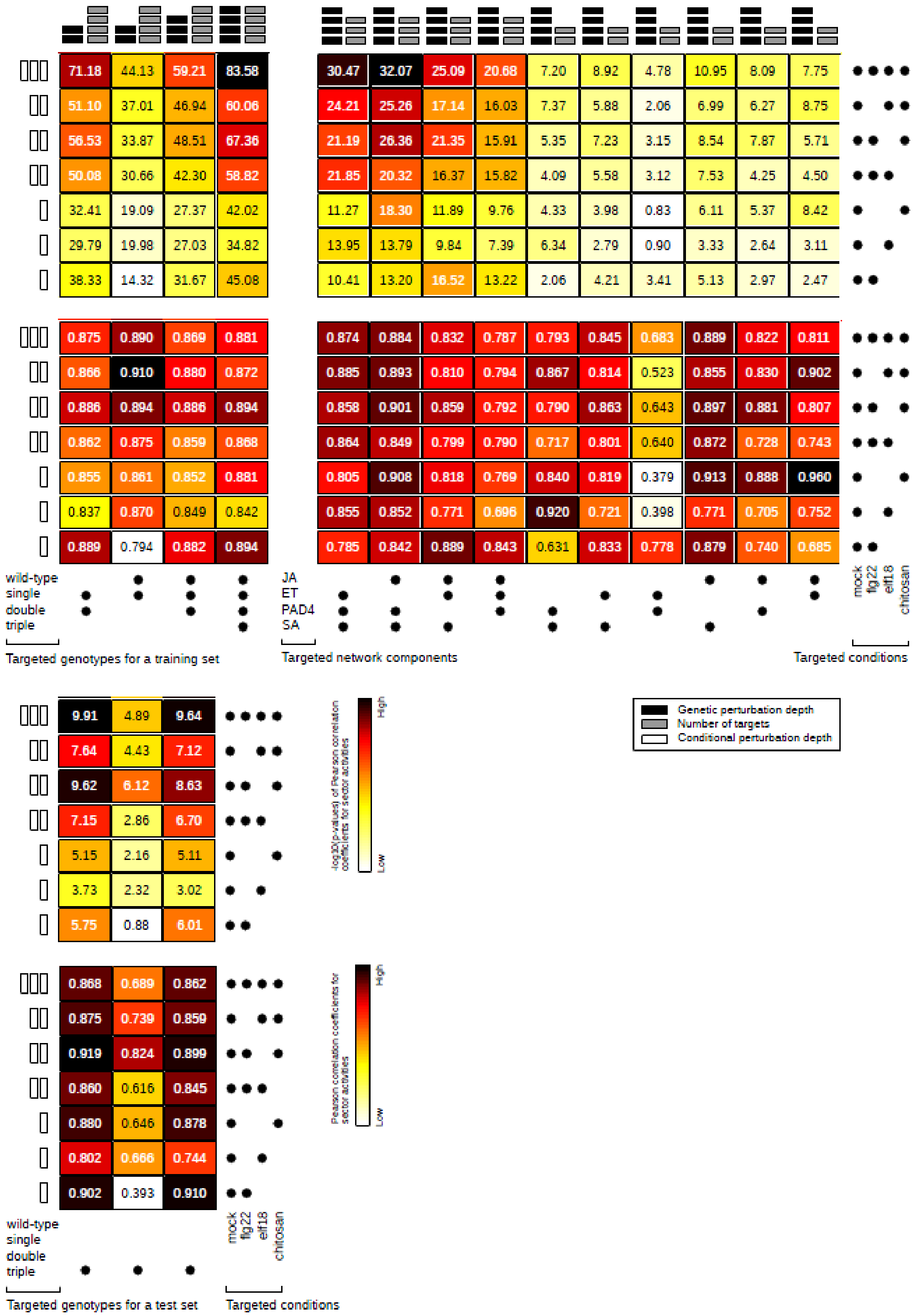


dde2/ein2/sid2




dde2/pad4/sid2
flg22
elf18
chitosan
3.14
2.82
1.83
2.41
JA(0.00)
ET(3.14)
PAD4(0.00)
SA(0.00)
JA(0.00)
ET(1.83)
PAD4(0.00)
SA(0.00)
JA(0.00)
ET(2.41)
PAD4(0.00)
SA(0.00)
JA(0.00)
ET(2.82)
PAD4(0.00)
SA(0.00)
JA(0.00)
ET(0.00)
PAD4(0.00)
SA(0.00)
JA(0.00)
ET(0.00)
PAD4(0.00)
SA(0.00)
0.17
0.08
pto(0.17)
pma(0.08)
pto(-0.00)
pma(-0.00)
pto(-0.00)
pma(0.00)
ein2/pad4/sid2
flg22
elf18
chitosan
1.73
4.32
3.05
4.34
3.44
6.92
JA(1.73)
ET(0.00)
PAD4(0.00)
SA(0.00)
JA(3.05)
ET(0.00)
PAD4(0.00)
SA(0.00)
JA(3.44)
ET(0.00)
PAD4(0.00)
SA(0.00)
JA(4.32)
ET(0.00)
-0.04
PAD4(0.00)
SA(0.00)
JA(4.34)
ET(0.00)
-0.07
PAD4(0.00)
SA(0.00)
JA(6.92)
ET(0.00)
-0.07
PAD4(0.00)
SA(0.00)
0.04
0.04
0.08
pto(0.04)
pma(-0.04)
pto(0.04)
pma(-0.07)
pto(0.08)
pma(-0.07)

Appendix III. The Effects of different factors in modeling a complex system



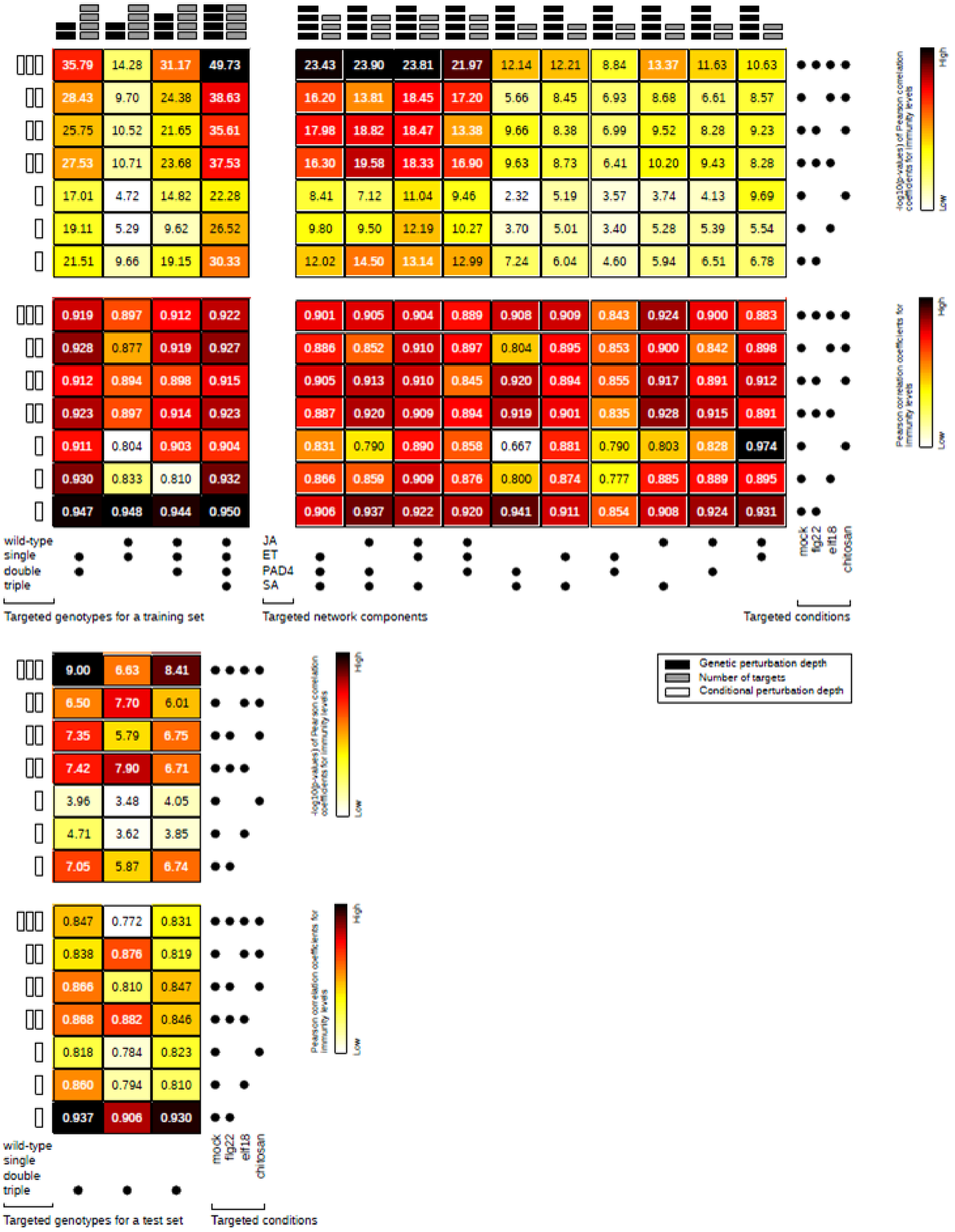
▯▯▯	71.18	44.13	59.21	83.58	30.47	32.07	25.09	20.68	7.20	8.92	4.78	10.95	8.09	7.75	●●●●
▯▯	51.10	37.01	46.94	60.06	24.21	25.26	17.14	16.03	7.37	5.88	2.06	6.99	6.27	8.75	● ●●
▯▯	56.53	33.87	48.51	67.36	21.19	26.36	21.35	15.91	5.35	7.23	3.15	8.54	7.87	5.71	●● ●
▯▯	50.08	30.66	42.30	58.82	21.85	20.32	16.37	15.82	4.09	5.58	3.12	7.53	4.25	4.50	●●●
▯	32.41	19.09	27.37	42.02	11.27	18.30	11.89	9.76	4.33	3.98	0.83	6.11	5.37	8.42	● ●
▯	29.79	19.98	27.03	34.82	13.95	13.79	9.84	7.39	6.34	2.79	0.90	3.33	2.64	3.11	● ●
▯	38.33	14.32	31.67	45.08	10.41	13.20	16.52	13.22	2.06	4.21	3.41	5.13	2.97	2.47	●●

▯▯▯	0.875	0.890	0.869	0.881	0.874	0.884	0.832	0.787	0.793	0.845	0.683	0.889	0.822	0.811	●●●●
▯▯	0.866	0.910	0.880	0.872	0.885	0.893	0.810	0.794	0.867	0.814	0.523	0.855	0.830	0.902	● ●●
▯▯	0.886	0.894	0.886	0.894	0.858	0.901	0.859	0.792	0.790	0.863	0.643	0.897	0.881	0.807	●● ●
▯▯	0.862	0.875	0.859	0.868	0.864	0.849	0.799	0.790	0.717	0.801	0.640	0.872	0.728	0.743	●●●
▯	0.855	0.861	0.852	0.881	0.805	0.908	0.818	0.769	0.840	0.819	0.379	0.913	0.888	0.960	● ●
▯	0.837	0.870	0.849	0.842	0.855	0.852	0.771	0.696	0.920	0.721	0.398	0.771	0.705	0.752	● ●
▯	0.889	0.794	0.882	0.894	0.785	0.842	0.889	0.843	0.631	0.833	0.778	0.879	0.740	0.685	●●

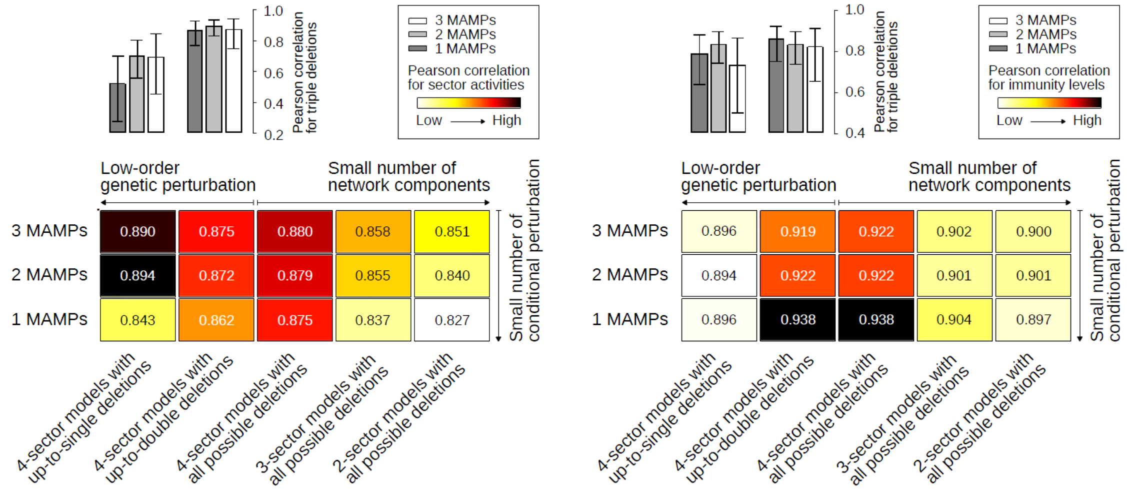
▯▯▯	9.91	4.89	9.64	●●●●
▯▯	7.64	4.43	7.12	● ●●
▯▯	9.62	6.12	8.63	●● ●
▯▯	7.15	2.86	6.70	●●●
▯	5.15	2.16	5.11	● ●
▯	3.73	2.32	3.02	● ●
▯	5.75	0.88	6.01	●●

▯▯▯	0.868	0.689	0.862	●●●●
▯▯	0.875	0.739	0.859	● ●●
▯▯	0.919	0.824	0.899	●● ●
▯▯	0.860	0.616	0.845	●●●
▯	0.880	0.646	0.878	● ●
▯	0.802	0.666	0.744	● ●
▯	0.902	0.393	0.910	●●

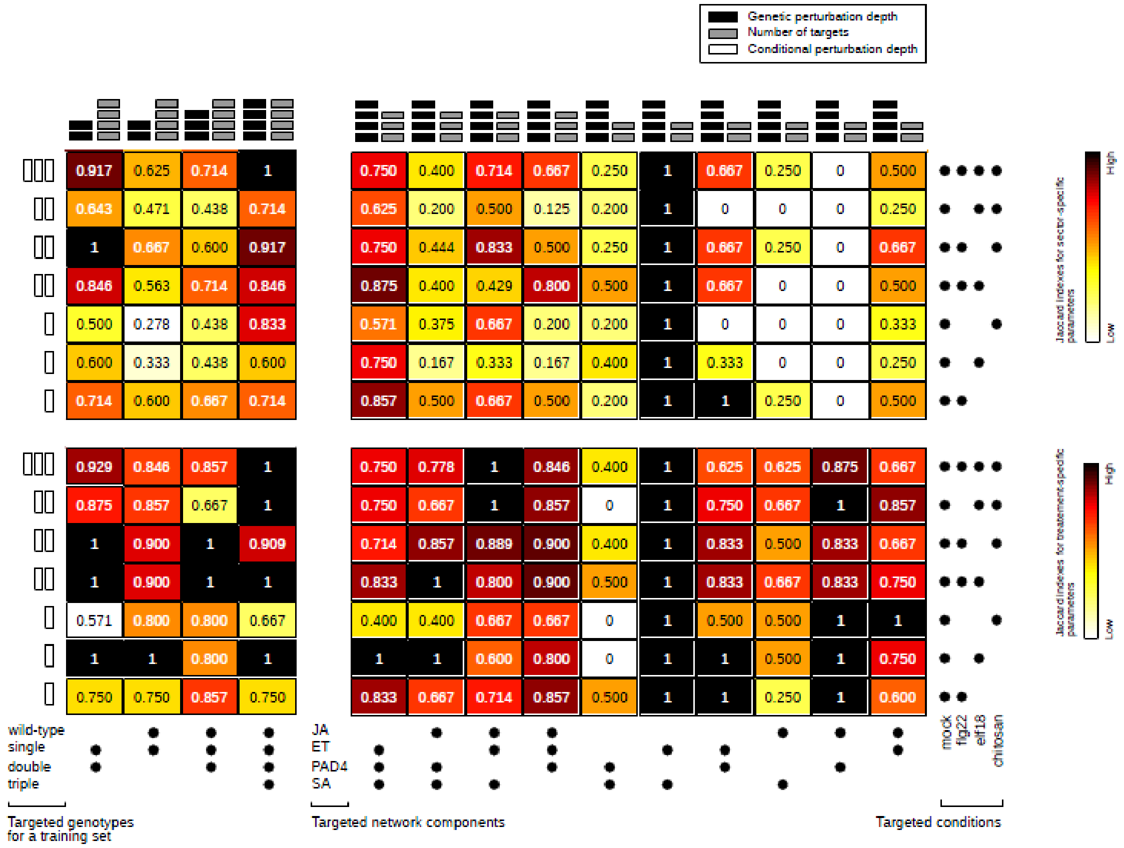
The PCC (-log$_{10}P$) and its significance for the prediction accuracies of the sector activities with constrained models. The left heatmap is for the genotype-constrained models. The genotypes used are shown at the bottom by black dots. The right heatmap is for fewer signaling sectors in the models. The sectors included are shown at the bottom by black dots. The MAMP treatments used differ in different rows of the heatamps. The MAMP treatments used are shown at the right by black dots. The bottom heatmap is for prediction of the sector activities in the triple mutant genotypes.



The PCC ($-\log_{10}P$) and its significance for the prediction accuracies of the immune levels with constrained models.



The PCCs between the predictions of different models and the observed data. Darker colors represent higher significances. The average PCC values for the constrained models of the same class (Eq. 4-9) are shown. The middle cell in the top row represents the original model. Bar plots show the PCCs for the prediction of the values in triple deletion from the constrained models with a limited order of genetic perturbations (two left columns).



0.917	0.625	0.714	1	0.750	0.400	0.714	0.667	0.250	1	0.667	0.250	0	0.500
0.643	0.471	0.438	0.714	0.625	0.200	0.500	0.125	0.200	1	0	0	0	0.250
1	0.667	0.600	0.917	0.750	0.444	0.833	0.500	0.250	1	0.667	0.250	0	0.667
0.846	0.563	0.714	0.846	0.875	0.400	0.429	0.800	0.500	1	0.667	0	0	0.500
0.500	0.278	0.438	0.833	0.571	0.375	0.667	0.200	0.200	1	0	0	0	0.333
0.600	0.333	0.438	0.600	0.750	0.167	0.333	0.167	0.400	1	0.333	0	0	0.250
0.714	0.600	0.667	0.714	0.857	0.500	0.667	0.500	0.200	1	1	0.250	0	0.500

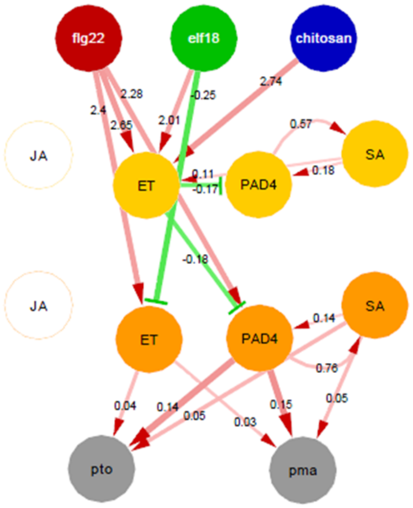
0.929	0.846	0.857	1	0.750	0.778	1	0.846	0.400	1	0.625	0.625	0.875	0.667
0.875	0.857	0.667	1	0.750	0.667	1	0.857	0	1	0.750	0.667	1	0.857
1	0.900	1	0.909	0.714	0.857	0.889	0.900	0.400	1	0.833	0.500	0.833	0.667
1	0.900	1	1	0.833	1	0.800	0.900	0.500	1	0.833	0.667	0.833	0.750
0.571	0.800	0.800	0.667	0.400	0.400	0.667	0.667	0	1	0.500	0.500	1	1
1	1	0.800	1	1	1	0.600	0.800	0	1	1	0.500	1	0.750
0.750	0.750	0.857	0.750	0.833	0.667	0.714	0.857	0.500	1	1	0.250	1	0.600

The Jaccard index values of the non-zero parameters between the constrained models and the original model.

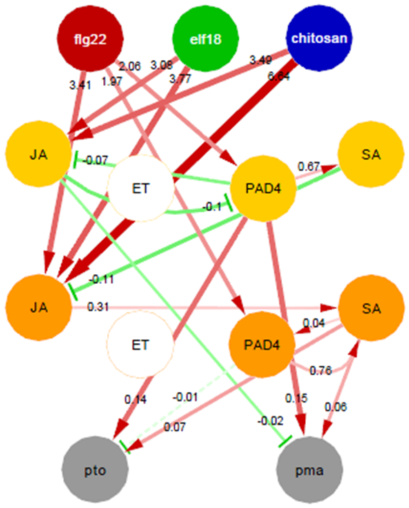
Appendix IV. Network models with limited numbers of sectors considered for plant PTI signaling network

3-sector models

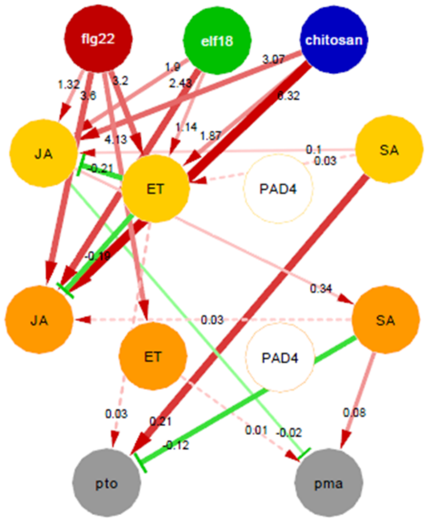
ET-PAD4-SA



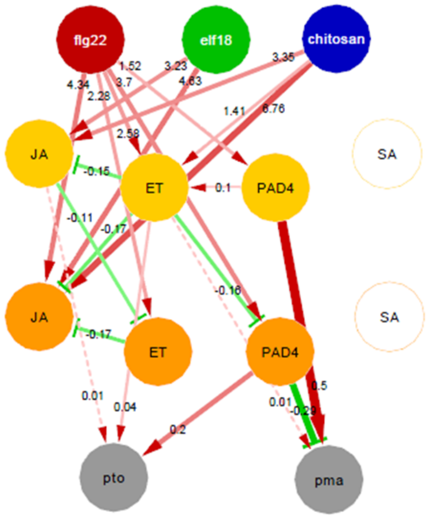
JA-PAD4-SA



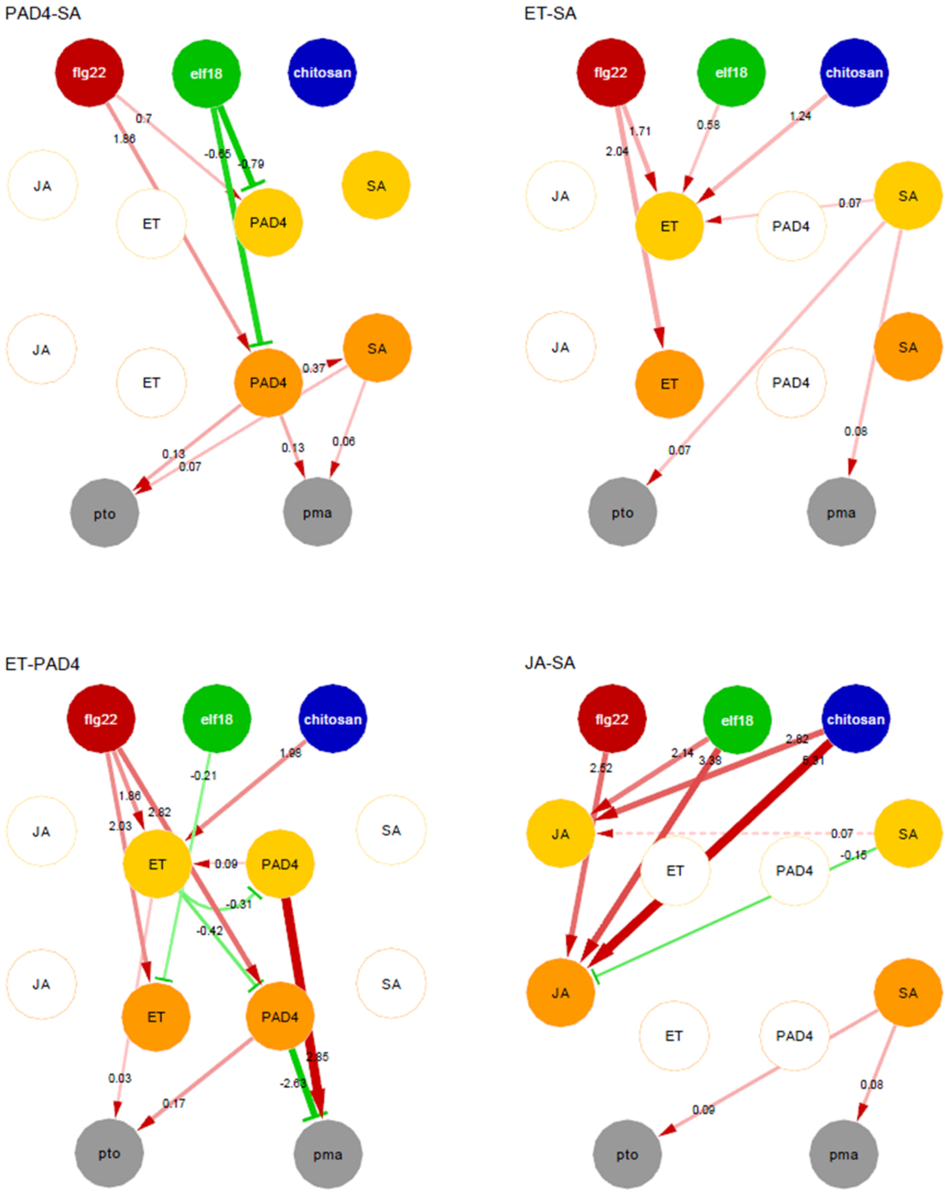
JA-ET-SA



JA-ET-PAD4



2-sector models

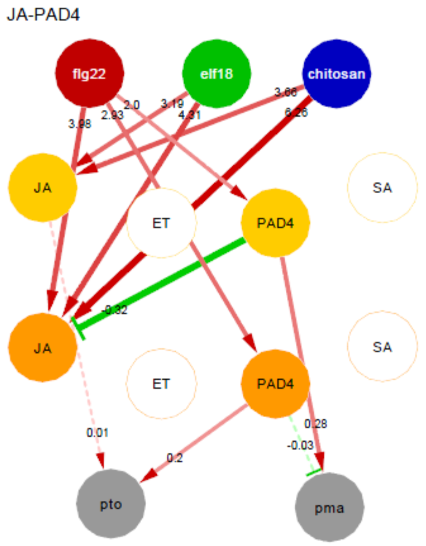


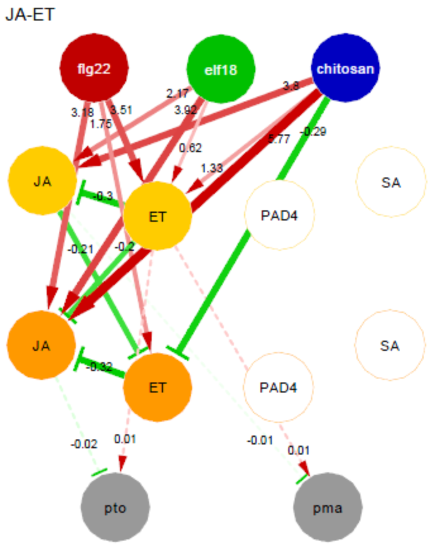
PAD4-SA

ET-SA

ET-PAD4

JA-SA


JA-PAD4
flg22
elf18
chitosan
3.98
2.93
2.0
3.19
4.31
3.66
6.26
JA
ET
PAD4
SA
-0.32
JA
ET
PAD4
SA
0.01
0.2
0.28
-0.03
pto
pma


JA-ET
flg22
elf18
chitosan
3.18
1.75
3.51
2.17
3.92
3.8
5.77
-0.29
0.62
1.33
JA
-0.3
ET
PAD4
SA
-0.21
-0.2
JA
-0.32
ET
PAD4
SA
-0.02
0.01
-0.01
0.01
pto
pma