# IMPACT OF NETWORK PROPERTIES ON EVOLUTION OF THE PLANT IMMUNE NETWORK

A THESIS SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL OF THE UNIVERSITY OF MINNESOTA

BY

Mridu Middha

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE

Dr. Fumiaki Katagiri, Advisor

July, 2012

## Acknowledgement

I would like to acknowledge and express heartfelt gratitude to my advisor, Dr. Fumiaki Katagiri for his exceptional support, valuable advice, constant guidance and encouragement throughout our association. For serving on my committee and for their stimulating discussions and suggestions during the course of my project I would like to extend many thanks to Dr. Claudia Neuhauser and Dr. Chad Myers.

I would also like to extend my thanks to Dr. Peter Morrell and Yungil Kim for their assistance and collaboration on my research project.

This work was supported by “Hatch/Minnesota Agricultural Experimental Station funds”.

# Dedication

To my husband, Sumit.

## Abstract

This study investigates how network evolution is affected by the underlying properties of the network. The plant immune signaling network is known to be robust against network perturbations. We hypothesize that genes of this robust immune network tend to be under neutral selection because deleterious mutations in such genes do not strongly affect the immune phenotype. I examined whether remnants of the hypothesized tendency of evolution can be detected in the currently existing natural population of the model plant *Arabidopsis thaliana.*

The genome sequences of 30 *A. thaliana* accessions with diverse geographical and environmental origins were obtained from the 1001 Genomes Project (1001genomes.org) for analysis. Using the TAIR8 annotation of the *A. thaliana* reference accession genome Col-0 a dataset of ~27,000 protein-coding genes for all accessions was generated. With such population genomic data it is feasible to study whether a group of genes are under selection different from another group, such as the entire genome.

Component genes of the plant immune signaling network were identified in a relatively unbiased manner by mining AraNet, a functional gene network model of *A. thaliana* (functionalnet.org). Population genetic summary statistics of the core network component genes and of all the genes in the genome were compared. The Tajima's D value distribution for all the genes in the genome had a single mode in a negative Tajima's D value, which is suggestive of purifying selection. The Tajima's D value distribution for

the core network component genes showed that this set of network component genes was significantly enriched with genes that have Tajima's D values near zero.

This suggests that immune network genes are enriched with genes with reduced levels of purifying selection compared with the genome average, which supports our hypothesis.

Table of Contents

List of Tables

List of Figures

## General Introduction

Since plants are sessile and lack an efficient circulation system and mobile cells, plants rely on the immunity of each cell. An important plant immune mechanism is inducible defense, in which immunity is triggered based on recognition of pathogen attack (Jones and Dangl, 2006; Tsuda et al., 2009). One mode of this plant immune system, the pattern-triggered-immunity (PTI) uses pattern-recognition receptors (PRRs) located on the cell surface to recognize conserved microbial-/pathogen-associated molecular patterns (MAMPs/PAMPs) (Zipfel, 2009). Pathogens well adapted to a host plant species deliver effectors, which are often proteinaceous, into the plant cell to interfere with PTI signaling to overcome PTI. The corresponding R proteins directly or indirectly recognize some effectors, which are typically the nucleotide-binding leucine-rich repeat (NB-LRR) proteins, and this recognition triggers another mode of plant immunity, effector-triggered immunity (ETI) (Jones and Dangl, 2001). The ETI response, which uses signaling machineries overlapping with those for PTI, results in strong disease resistance and includes localized programmed cell death referred as hypersensitive response (Jones and Dangl, 2006; Block et al., 2008).

As discussed in a recent review, it appears that while both modes of plant immune network share common signaling machinery, this machinery is used differently in PTI and ETI (Katagiri and Tsuda, 2010). It is proposed that when the plant perceives attack from a pathogen, the difference in the strength, amplitude or timing of the signals coming into the network and how these signals are further processed in the network determine the mode of immunity, such as PTI and ETI (Katagiri and Tsuda, 2010).

While pathogens that are well adapted to a host plant can interfere with PTI signaling to overcome PTI, pathogens typically overcome ETI by evading recognition from R proteins and not by interfering with ETI signaling (Tsuda and Katagiri, 2010). These observations suggest that the ETI signaling network is difficult for pathogens to interfere with. When mutants for ETI phenotypes were screened in plants, most mutations were found in genes encoding R proteins or the proteins required for the function of R proteins but very few mutations were found in ETI signaling components (Tsuda and Katagiri, 2010). These observations suggest that the ETI signaling network for ETI is highly robust against network perturbations caused by mutations in the network component genes.

The robustness of the ETI signaling network in *Arabidopsis thaliana* was directly demonstrated by genetically combining impairments of major network sectors (Tsuda et al., 2009). The impairment of each of four major sectors, salicylic acid, ethylene, jasmonic acid, and PAD4-dependent signaling sectors, has a relatively weak effect on the phenotype of ETI triggered by AvrRpt2, an effector of the bacterial pathogen *Pseudomonas syrinage*. Combining any two of the sector impairments still had limited effects on the AvrRpt2-ETI phenotype. However, combining all four sector impairments resulted in loss of most AvrRpt2-ETI. Thus, the ETI signaling network is highly robust against network perturbations.

The genes that compose such a robust network are likely to be relieved from strong selection (i.e., relatively neutral selection) during evolution of the network because deleterious mutations in many of the network genes are expected to have limited effects on the phenotype controlled by the network. I hypothesized that the set of the genes that

compose the robust plant immune signaling network is enriched with the genes under relatively neutral selection. I tested this hypothesis by examining whether remnants of relatively neutral evolution among the immune signaling network genes can be detected in the currently existing natural population of the model plant *A. thaliana*. To this goal, I compared the population genetic summary statistics between the immune signaling network genes and all genes in the genome. This was achieved by: (Chapter 1) compiling the coding sequences of all protein-coding genes in 30 diverse accessions of *A. thaliana* from the publicly available *A. thaliana* population genomic data; (Chapter 2) defining the immune signaling network genes in a relatively unbiased manner by mining a publicly available *A. thaliana* functional gene network model; (Chapter 3) and calculating population genetic summery statistics for every protein-coding gene and comparing those statistics between the immune signaling network genes and the entire genome. The work described in Chapter 2 was mostly performed by my collaborator, Yungil Kim.

I discovered that the distribution of one of the summary statistics, Tajima's D, for the immune signaling network genes is consistent with our hypothesis: genes with Tajima's D ~ 0 are significantly enriched among the immune signaling network genes compared to the genomic average. Furthermore, my approach that combines population genomics and a functional gene network model establishes a new means to study network properties and their evolution.

# Chapter 1 Compilation of Arabidopsis population genomics data

## Introduction

The hypothesis for my study is that *Arabidopsis thaliana* immune signaling network genes are enriched with genes that are relieved from strong selection compared with the rest of the genome. A common approach to detect selection that has acted on a gene in a population is to calculate population genetic summary statistics, such as Tajima's D and the derived site frequency spectrum. Using a representative data set of *A. thaliana* population calculation of population genetic summary statistics for all the genes in a population can define the genomic distributions of the statistics. This distribution of the statistics for immune signaling network genes can be compared with the genomic distributions for all genes to analyze the differences in selection from those of the genomic average.

With the advancement in sequencing technology and the launch of the 1001 Genomes Project studying genome sequences of multiple individuals in *A. thaliana* species has been made feasible. Thirty accessions were selected from the existing pool of released sequences from 1001 Genomes Project. These sequences were selected to represent the natural population of *A. thaliana* and to detect trends of selection in protein-coding genes. In this chapter multiple sequences of all thirty accessions for each of the protein-coding genes were compiled and aligned for calculating population genetic summary statistics in Chapter 3.

## *Arabidopsis thaliana*

*Arabidopsis thaliana* is a small flowering plant and a member of the mustard (Brassicaceae) family that is widely used as a model organism in plant biology. Although it has no economic value, it has many advantages in genetics, genomics, and molecular biology research (Weigel and Mott, 2009). It is highly inbred and has a diploid genome of 157Mb per haploid. The genome is distributed over five chromosomes with 27,206 protein-coding genes as defined in the most recent TAIR10 annotation release. The plant is easy to grow even in restricted space with a fast generation time of about 6 weeks from germination to seed maturation.

*A. thaliana* was the first plant to have its entire genome sequenced making it the third eukaryotic genome sequenced (The Arabidopsis Genome Initiative, 2000). The Arabidopsis Information Resource (TAIR) collects and makes a wide variety of information publicly available, by curating and maintaining a database of genetic, genomic, and molecular biology information regarding *A. thaliana*. Such information available from TAIR includes complete genome sequence of the reference accession Columbia-0 (Col-0), annotated with the gene model, gene product information, the related literature, and the data underlying the database. Furthermore, TAIR also provides through their website a range of tools, including externally available ones, for data analysis and visualization (Lamesch et al., 2010).

## 1001 Genomes Project

*A. thaliana* is considered to have naturally spread from Central Asia to Eurasia and is currently found throughout the Northern Hemisphere in temperate regions. It was introduced into North America during historic times.

Naturally occurring *A. thaliana* accessions are genetically and phenotypically diverse. A number of nonsense and missense mutations underlie phenotypic variation, in addition to the alleles that alter gene expressions levels (Weigel and Mott, 2009). With the advancement of fast and affordable sequencing technologies it has become feasible to sequence and compare multiple genome sequences of the same species to investigate the genome variations within the species.

The 1001 Genomes Project, launched in 2008, is a collaborative effort to determine the sequence variation in whole-genome sequences of 1001 accessions of *A. thaliana*. The goal of the project is to generate genome sequences of 1001 *A. thaliana* accessions using the next generation sequencing technologies for the purpose of studying genotypic differences that underlie the phenotypic diversity across multiple genomes of the species. The 1001 Genomes Project sequences a large number of genomes at 8x coverage using different technologies, primarily Illumina's Genome Analyzer and Applied Biosystems' SOLiD. With this depth of coverage, they also aim to identify less frequent (allele frequency less than 1%) haplotypes (Weigel and Mott, 2009).

As the pilot of 1001 Genomes Project, 80 complete genomes of *A. thaliana* accessions were selected from 8 regions across Eurasia and were sequenced with paired end Illumina short reads. With such population genomic data, it is feasible to ask whether a group of genes are under selection pressures different from the rest of the genome.

## TAIR genome release

The TAIR goes through automated gene updates by employing TIGR PASA annotation pipeline (Haas et al 2003) for new releases. After the Arabidopsis cDNAs and ESTs are trimmed for contaminating vector sequences and poly-A tails are removed, the cDNAs are aligned to the genome via BLAT. Strict validation criteria are applied, including the minimum ORF size, maximum number of UTR exons, etc. and then validated alignments are compared to the pre-existing gene models. Manual annotators review the suggested automated updates while the curators can utilize these as evidence to support updates to gene structures (Lamesch et al., 2010). Annotation of new genomes is largely based on the annotation of existing complete genomes.

The first 80 *A. thaliana* accessions that were sequenced in the 1001 Genomes Project were aligned to the reference sequence Col-0 of TAIR8 genome release. This study uses accessions from the first dataset that was released, based on the TAIR8 gene model. The TAIR8 has 27,025 protein-coding genes distributed over five chromosomes.

## Methods

### Selecting thirty genomes

Data for eighty *A. thaliana* accessions that are aligned to the TAIR8 assembly of the reference sequence Col-0 were downloaded from 1001 Genomes Data Center (1001genomes.org). The downloaded *A. thaliana* genome data matrix has base calls of all eighty genomes, in which the nucleotide positions are numbered according to the nucleotide position in the reference Col-0 genome.

Twenty random sequence regions of 0.5Mb length from each genome sequence were selected from this data matrix of *A. thaliana*. For each accession definitive calls of either nucleotide, blank or D for every base was calculated to determine overall sequence quality. For each accession, the average and standard deviation of the definitive calls among the randomly selected 20 samples were calculated and sequence quality was compared among all these accessions from the matrix. The accessions with high average values and low standard deviation values were determined to have high quality sequences.

To make the data selection geographically and environmentally diverse the table of selected accessions was referred. Considering the habitat and location of *A. thaliana* accessions which had good quality sequences, 29 accessions were further subjectively selected.

High quality selected twenty-nine accession sequences from the matrix format were translated to genome-data of BED file format (Table 1.1) with name of the chromosome, start position of the nucleotide and stop position and the nucleotide call at

every position in thirty accessions (including Col-0). BED format provides a flexible way to define data and the first three columns are required, chromosome, start position and stop position and other columns are optional, like name, score, strand etc.

Table 1.1: Data for 30 accessions from TAIR8 in bed format

| Chromosome | Start Position | Stop Position | Name **(base calls at every nucleotide position for all 30 accessions separated by '_')** |
|---|---|---|---|
| **chr1** | 12695 | 12696 | G_G_G_G_G_G_G_G_G_G_G_G_G_G_G_G_G_G_G_G_G_G_G_G_G_G_G_G_G_G_G_G |
| **chr1** | 12696 | 12697 | A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A |
| **chr1** | 12697 | 12698 | A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A |
| **chr1** | 12698 | 12699 | A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A |

The gene models for all five chromosomes of Col-0 for TAIR8 genome release were downloaded from The Arabidopsis Information Resource (Lamesch et al., 2010). The gene models were defined for every chromosome in individual tables specifying multiple start and stop, defining exons for the protein-coding genes based on TAIR8 annotation. This annotation information was translated into TAIR8-annotation BED file format with name of the chromosome, start position of coding sequence, stop position, name of protein-coding region and strand orientation (Figure 1.2)

Table 1.2: Gene models defined in BED format for all five chromosomes for TAIR8 genome release

| Chromosome | Start Position | Stop Position | Name (protein coding region) | Score (NA) | Strand |
|---|---|---|---|---|---|
| **chr1** | 5174 | 5326 | AT1G01010.1p | 1000 | + |
| **chr1** | 5439 | 5630 | AT1G01010.1p | 1000 | + |
| **chr1** | 8571 | 8666 | AT1G01020.1p | 1000 | - |
| **chr1** | 8417 | 8464 | AT1G01020.1p | 1000 | - |

Using intersectBED tool from BEDtools library (Quinlan and Hall, 2010) protein-coding genes for the selected thirty accessions, including reference genome from the genome-data BED file were intersected with the TAIR8-annotation BED file, defined for the reference genome accession Col-0, to select nucleotides at all positions of exons of all protein-coding regions for all selected accessions (Figure 1.3).

Table 1.3: Intersected BED file with sequence data and annotation information

(Data file is shown split vertically)

| **Chromosome** | **Start Position** | **Stop Position** | **Name (protein coding region)** | **Score (NA)** |
|---|---|---|---|---|
| **chr1** | 3766 | 3767 | AT1G01010.1 | 1000 |
| **chr1** | 3767 | 3768 | AT1G01010.1 | 1000 |
| **chr1** | 3768 | 3769 | AT1G01010.1 | 1000 |
| **chr1** | 3769 | 3770 | AT1G01010.1 | 1000 |

| **Strand** | Name **(base calls at every nucleotide position for all 30 accessions separated by '_')** |
|---|---|
| + | G_G_G_G_G_G_G_G_G_G_G_G_G_G_G_G_G_G_G_G_G_G_G_G_G_G_G_G_G_G_G_G |
| + | A_A_A_A_A_A_A_A_A_A_A_A_A_A_G_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A_A |
| + | T_T_T_T_T_T_T_T_T_T_T_T_T_T_T_T_T_T_T_T_T_T_T_T_T_T_T_T_T_T_T_T |
| + | C_C_C_C_C_C_C_C_C_C_C_C_C_C_C_C_C_C_C_C_C_C_C_C_C_C_C_C_C_C_C_C |

## Multiple Sequence Alignment

Data from the intersected file with coding sequences of all thirty accessions for every protein-coding sequence were parsed. This generated multiple aligned FASTA files for each protein-coding sequences containing the sequence for all 30 accessions.

The sequences in negative orientation were reverse complemented to get all the coding regions in positive orientation.

The lengths of isoforms, which are different forms of the same protein that may be produced for the same gene by alternative splicing, were determined. By combining the exon lengths for each isoform, the longest isoform for every protein-coding region was selected.

## Quality Control

Quality control was performed to test the accuracy of parsing methods for protein-coding sequences for all accessions. Protein-coding sequences for Col-0 genome from multiple aligned FASTA files were selected and translated to peptide sequence using '*transeq'* tool from Emboss suite of tools for bioinformatics analysis (Rice et al., 2000).

TAIR8 peptide sequence for Col-0 was downloaded from The Arabidopsis Information Resource and compared with the Col-0 peptide sequence translated from our multiple aligned FASTA files. Two methods, perl script for string match and Protein BLAST (Altschul et al., 1997) to align two sequences were used for the comparison.

## Results and Discussion

It is known that increasing the number of samples analyzed does not dramatically increase the statistical power in inferring evolutionary history of a particular gene (Nielsen and Wakeley, 2001). I have studied and analyzed the data from the high quality genome sequences of 30 diverse accessions of *A. thaliana* (Figure 1.1). Increasing the number of accessions from 30 to 40 would not increase the statistical power significantly while computing time for pair-wise sequence analysis would increase by 79%. Thus, we decided to use 30 accessions. These selected 30 accessions span over 15 countries (strains from 8 Eurasian regions (Figure 1.2)) with diverse locations and have been found in a range of over 14 varied habitats (Table 1.4).


Figure 1.1: Diverse geographical accessions of *A. thaliana* selected for analysis

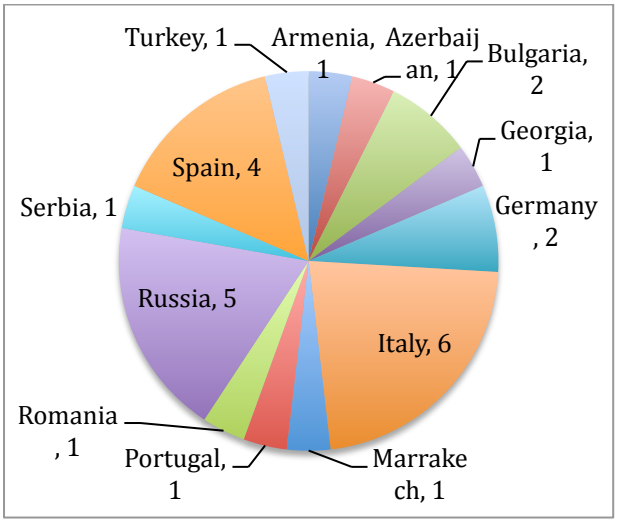


Figure 1.2: Selected accessions span over 15 countries

Table 1.4: Altitudes and habitats of some of the accessions that were selected

| Accessions | Altitude | Habitat |
|---|---|---|
| Agu-1 | 1045 | Mediterranean forest of evergreen Quercus species |
| Bak-7 | 1189 | Open roadside |
| Cdm-0 | 470 | Stone wall between grasslands and a road |
| Col-0 | 50 | NA |
| ICE134 | 565 | Steppe |
| ICE138 | 301 | Pine forest |
| Kastel-1 | 389 | Open rocky outcrop and adjacent roadside, full sun |
| Ped-0 | 1090 | Mediterranean xeric scrublands |
| Qui-0 | 700 | Roadsides in the surroundings of a village |
| Xan-1 | 37 | Partially shaded roadside and open pasture |
| Yeg-1 | 1528 | Disturbed, open area |

From a total number of 32,615 coding regions, which were parsed earlier, 27,025 protein-coding genes were selected after removing multiple isoforms for a gene. Thus, in the dataset of 27,025 protein-coding genes a single protein-coding gene was represented by a single coding sequence.

A comprehensive coding sequence dataset of 30 accessions of *A. thaliana* was compiled. The dataset contains available information regarding full-length coding sequences of 27,025 genes for twenty-nine accessions and reference sequence Col-0 based on TAIR8 annotation. Sequence annotation for this dataset was validated and checked for consistency at the protein level by comparing the peptide sequence translated from the protein-coding sequence of Col-0 in our compiled files with the TAIR8 annotated peptide sequence.

27,009 parsed and translated sequences were found identical to the downloaded TAIR8 peptide sequences using both the methods of string matching and BLAST alignment. The 16 sequences that did not match were removed from further analysis. The compiled sequence files passing the quality control were used in Chapter 3 for population genomic analysis.

The dataset of multiple aligned sequences in FASTA format of environmentally and geographically diverse sequences of *A. thaliana*, a model plant, is highly compatible or easily convertible to the file formats used in various population genetics tools. Thus, it can be readily used for population genomics studies in *A. thaliana*. With such population genomic dataset it is feasible to study the factors that might affect the genomic patterns of diversity, such as mechanistic properties of the underlying gene networks. In Chapter 3,

this dataset is used to ask whether a group of genes are under selection different from another group, such as the rest of the genome.

# Chapter 2 Identification of the Arabidopsis immune signaling network genes

## Introduction

To test whether the immune signaling network genes are robust against network perturbations, immune network component genes need to be distinguished from all the genes of the genome. For this purpose, seed genes were used as starting points to mine AraNet, a probabilistic functional gene network of *A. thaliana*. This is a novel approach of selecting in a relatively unbiased manner, which allows comparison of selection trends between the selected group of genes and the whole genome.

It is likely for genes to be under selection if their gene functions are already known, so it is important to identify the network component genes in an unbiased way. Once the list of component network genes has been compiled the protein-coding sequences can then be identified from the population genomics data set of 30 accessions of *A. thaliana* that has been parsed and verified in Chapter 1. Population genetic summary statistics would be calculated in Chapter 3 and the genomic distribution for all the genes and network component genes can be compared and analyzed.

## AraNet – A functional gene network

Availability of genome-scale data enables different approaches in inferring gene functions. One interesting approach is to build a network of genes, in which genes that are inferred to have similar functions are connected by edges. AraNet is such a functional gene network of *A. thaliana* (Lee et al., 2010). This network contains approximately ~73% of the total *A. thaliana* protein-coding genes. The network structure of the AraNet

was inferred based on diverse functional genomics, proteomics, and comparative genomics data sets.

Each functional linkage in AraNet among the genes is weighed by the log likelihood of the linked genes to participate in the same biological processes (Lee et al., 2010). A vital point of the AraNet functional network is that non- molecular phenotypes, such as the immune phenotypes that are of interest to our study, were not used as evidence for the inference. Therefore, AraNet can be considered as a relatively unbiased functional gene network when it is used for analysis related to non-molecular phenotypes.

## Seed genes and network component genes

The function of a gene is typically determined based on the phenotype caused by alteration of the gene activity, such as mutations and overexpression of the gene. For example, alterations of the activities of genes in the functional category of immunity likely result in immune phenotype. This means that genes with functions already assigned are likely to be selected strongly.

If genes were selected based on their known functions associated with immunity, this procedure of selecting a group of genes would result in an enriched set of genes under evolutionary selection. Based on such a biased gene selection procedure, it would not be possible to investigate the effect of the network properties on evolution of the network. Therefore, it was needed to identify immune signaling network component genes independent of their known functions.

To select immune signaling network component gene candidates in a relatively unbiased manner, AraNet was mined. It is speculated that the robustness of the network is

a property of a core part of the signaling network. Genes close to the inputs or outputs of the network, which we refer to as peripheral genes, are likely under selection. For example, genes encoding receptors that recognize pathogen-derived molecules define inputs to the immune network. Thus, the network component gene selection procedure was designed to identify network component genes that likely reside in the core part of the network.

The starting points to mine the functional gene network, referred to as seed genes, were selected based on their known regulatory functions in immunity. AraNet was then mined using these known seed genes to find the shortest path between each pair of seed genes. The component genes that lay in the shortest path for many of the traversed paths between these pairs of seed genes were considered as network component candidates. These identified component candidates gene pairs were then included as additional starting points and more immune network component gene candidates were identified similarly. Considering that the seed genes are likely to be strongly selected, the seed genes were not included in our list of identified network component genes.

## Methods

### Identifying seed genes

*A. thaliana* genes that are shown to be involved or strongly implicated in immune signaling were selected as seed genes. Eventually 82 seed genes were identified from published literature as alterations of these genes led to immune phenotypes.

### Identification of the network components in a relatively unbiased manner

In AraNet we first selected the genes that lie in the shortest path between every pair of these seed genes using Dijkstra's algorithm. To find the shortest path, the distance between two genes in AraNet was defined as follows. Considering the weight of each edge that links two gene nodes, representing the confidence level of the edge, we calculated the distance between two genes via a particular path by this formula (Opsahl et al., 2010):

$$1/w1^{\alpha} + 1/w2^{\alpha} + \ldots$$

here w1, w2, .. are the weight of the edges along the path. Seven different values of alphas were surveyed, ranging from 0 (binary edge) to 3 (strength-based) to vary the emphasis of the edge weights. The genes that were identified to be on the shortest paths were ranked based on the weighted betweenness centrality (Freeman, 1977) which represents how frequently a component lies in the shortest paths that are calculated between pairs of all genes. Thus the top ranked genes being associated with nearly all seeds in a shortest-path manner, makes them good candidates for the core network component genes.

## Results and Discussion

Figure 2.1 demonstrates the use of different values of α (α = 0, 0.5, 1, 1.5, 2, 2.5, 3) for finding the shortest paths among seed genes (Table 2.1). At α = 2 the average number of seed genes that belong to any shortest paths saturated around 55. We used this criterion to choose the α value because if the shortest paths discovered with a particular α value traverse more seed genes, such a set of shortest paths likely better represent the network defined by the seed genes used.

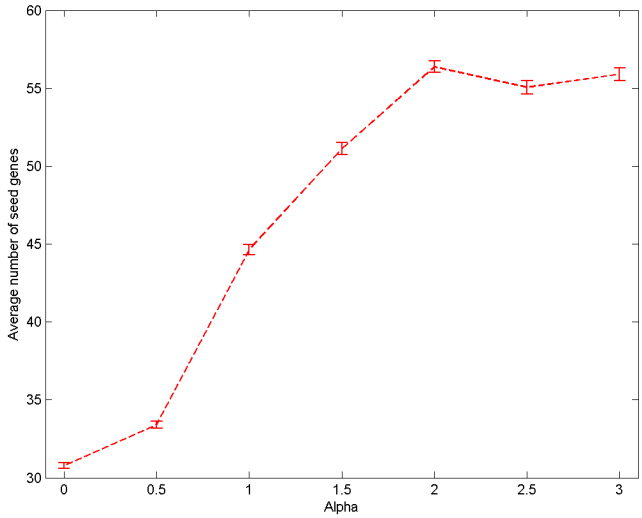


Figure 2.1: Using different α values to ascertain the optimal value of α. The x axis is values of α and the y axis is the average number of seed genes that were traversed between at least one pair of seed genes for the given value of α.

Table 2.1: Sample table of 14 seed genes with their common names

| AGI code | Common name |
|---|---|
| **At1g70700** | JAZ9 |
| **At1g71860** | PTP1 |
| **At1g77760** | NIA1 |
| **At1g80460** | NHO1/GLI1 |
| **At1g80840** | WRKY40 |
| **At2g17290** | CPK6 |
| **At2g19190** | FRK1 |
| **At3g55450** | PBL1 |
| **At3g56400** | WRKY70 |
| **At4g01370** | MPK4 |
| **At4g03110** | RBP-DR1 |
| **At4g03550** | PMR4/GSL5 |
| **At4g08500** | MEKK1 |
| **At4g09570** | CPK4 |

Five hundred and sixty-nine network component genes were identified using alpha value = 2. These five hundred and sixty-nine network component genes were listed in decreasing order of betweenness centrality. The top-ranked genes were with the highest betweenness centrality and the ones with the most central in the network. These genes were associated with most seeds in a shortest-path manner and had greater probability to be in the core part of the network.

Top 10% genes (fifty-five genes, removing 2 genes which were not found in the TAIR8 annotation from the top fifty seven) were selected to represent core immune network component genes for comparative analysis with all genes of the *A. thaliana* genome.

Studying immune signaling network by comparing properties of a group of genes to that of all genes of the genome is a novel approach. This was enabled by identification of core network component gene candidates in a relatively unbiased manner.

# Chapter 3 Comparison of population genetic summary statistic values between the immune signaling network genes and all the genes

## The Hypothesis and an approach to test the hypothesis

The hypothesis of my study is that the immune signaling network is enriched with genes that are relieved from selection compared with the rest of the genome. Detection of the remnant of such a network evolution history in the genomes of the currently existing *A. thaliana* population would provide a support to this hypothesis. A common approach to detect selection that has acted on a gene in a population is to calculate population genetic summary statistics, such as Tajima's D and the derived site frequency spectrum (Oleksyk et al., 2010).

## Introduction

Prior to the advent of sequencing technologies, whole genome sequencing data for many individuals of a species was not available for in-depth research and comparisons. Population genetic summary statistics were calculated for a small number of genes, and statistic values of the genes were compared with those calculated, based on the neutral theory model. In this way, it could be inferred whether the gene of interest is under negative or positive, or purifying or balancing selections (Oleksyk et al., 2010). As the statistical values were compared with the neutral theory model, interest of the field has been focused on the genes that are under significant selection. Such studies based on a small number of genes may not satisfy the underlying assumptions used to interpret

summary statistic values. For example, the population size and demography, in addition to selection, affect genetic variation in a population (Montesinos et al., 2009).

However, obtaining genome sequences of multiple individuals in a single species has become affordable with the declining cost of sequencing. Launch of collaborative endeavors, such as 1001 Genomes Project, makes sequence data of individual genomes publicly available. Such resources aid in extensive research like calculating population genetic summary statistics for all the genes in a population. Using this information it is realistic to argue whether the summary statistic value of a particular gene or the summary statistic distribution of a particular gene group is significantly different from those of the genomic average. It is reasonable to assume that the genomic average value represents non-gene specific effects, such as the population demography and average level of the selection effect and that comparisons with the genomic average should highlight the specific selection on the particular gene or gene group.

Using protein-coding sequence data compiled in Chapter 1 and the core network component genes identified in Chapter 2 it is feasible to test our hypothesis by comparing population genetic summary statistic values between the immune signaling network genes and all the genes.

## Tajima's D Summary Statistics

Tajima's D is a population genetic summary statistic used to test whether a particular gene, or a group of polymorphic sites, is under selection or demographic change (Tajima, 1989). In a neutral theory model, the expectations of the average pairwise difference, $\pi$, and of the number of polymorphic sites adjusted for the sample number, $\theta_w$ which is a method for estimating the population mutation rate (also known as Watterson's $\theta$ (Watterson, 1975)), are equal.

However, these expectations differ when the population is under selection and/or under demographic change. Tajima's D is the standardized value of the difference between observed $\pi$ and $\theta_w$ values. With an assumption of a well-mixed population of a constant size, when the two estimates differ there is evidence for selection. Positive Tajima's D signifies balancing selection, where the number of alleles for the gene of interest is low but the multiple alleles tend to be at relatively high frequencies. Negative Tajima's D signifies purifying selection, which means that a single allele of the gene is predominating in the population and the other alleles have low frequencies. A gene is under neutral selection when Tajima's D ~ 0 and this is when the DNA sequence contains random mutations which have no effect on the fitness and survival of an organism.

The neutral patterns of nucleotide variation expected at equilibrium can vary with change in population size. Rare frequency mutations are lost more readily following a reduction in population size (Nei et al., 1975) and positive Tajima's D values are expected (Tajima 1989). Following an increase in population size, there is a temporary

excess of new mutations segregating at rare frequencies, and negative D values are expected (Fay and Wu, 1999).

### Derived site frequency spectrum

The derived site frequency is defined for each nucleotide position with a Single Nucleotide Polymorphism (SNP) as the number of the occurrence of the derived allele among the individuals in the population of interest. In this study, 30 Arabidopsis accessions were used. For instance, if the derived allele occurs in 8 accessions at a SNP site, the derived site frequency is 8. The derived site frequency spectrum is a histogram of the derived site frequency in a group of SNPs, typically for a single gene.

An important difference between derived site frequency spectrum and Tajima's D is that while Tajima's D only uses information on diversity within the species in question, the derived site frequency spectrum uses information on divergence of the species as well. For example, when a single allele at a particular SNP site is predominating in a population, calculation of Tajima's D does not use information whether the predominating allele is the ancestral or derived allele while calculation of the derived site frequency spectrum does. Due to this difference, Tajima's D cannot distinguish a negative selection, in which the ancestral allele is selected, from a strong positive selection of a newly arisen allele while the derived site frequency spectrum can. In the derived site frequency spectrum, a negative selection enriches polymorphic sites with very low derived site frequencies, while a strong positive selection leads to enrichment of high derived site frequencies. Balancing selection leads to decrease of both low and high derived site frequencies. For these reasons, the derived site frequency spectrum was also

investigated in this study.

## Methods

### Calculation of Tajima's D Summary Statistic

Using COMPUTE tool from the LIBSEQUENCE library (Thornton, 2003) Tajima's D was calculated for each multiple aligned FASTA file. Using a shell script the value of Tajima's D for all protein-coding sequences were parsed from the output and reported into a table.

### Determining Significance of Tajima's D Results

Since a higher density of the Tajima's D value distribution around 0 was observed in the immune signaling network genes compared to all genes, the significance of this enrichment was tested using Fisher's exact test. The boundary values of (-0.4,0.6) were chosen post-hoc.

### Identification of *A. thaliana* Genes That Have Close Homologs

The Col-0 TAIR8 peptide sequence (Lamesch et al., 2010) library was searched with each of the *A. thaliana* genes as a query using BLASTP (Altschul et al., 1990). Each gene that had at least one homolog with E-value < 1e-10 was identified as a gene with a close homolog.

## Calculation of Summary Statistics for Synonymous and Non-Synonymous SNP Sites Separately

POLYDNDS tool in LIBSEQUENCE library (Thornton, 2003) was used to segregate the synonymous and non-synonymous sites from multiple aligned FASTA for all the protein-coding genes to calculate Tajima's D.

## A Significance Test for the Bimodal Curve

Random sampling was used to test whether the trough between two modes that was observed with the non-synonymous Tajima's D value distribution was significant by comparing the second peak value and the trough value. The test took samples of 55 genes randomly thousand times from the genomic non-synonymous data of all the genes to see how probable it is to get a bimodal distribution and, if so, how deep the trough would be compared to the second peak.

## Identification of Orthologous Genes between *A. thaliana* and *A. lyrata*

The reciprocal best BLAST hits criterion (Moreno-Hagelsieb and Latimer, 2007) was used to identify orthologous genes between *A. thaliana* and *A. lyrata*. NCBI BLASTN (Altschul et al., 1990) was run against a library of the coding sequences of the *A. thaliana* Col-0 sequence with each of the *A. lyrata* coding sequences (ver 7, Hu et al., 2011) as a query and *vice versa*. The BLASTN results were parsed using a custom Perl script to identify the reciprocal best BLAST hits.

In BLASTN results for some queries, the alignments were divided into multiple high-scoring segment pairs (HSPs) due to dispersed sequence regions of low homologies.

For the purpose of using the *A. lyrata* sequences to infer the *A. thaliana* ancestral sequences, it is not useful if the alignment between an orthologous gene pair is not very long compared to the *A. thaliana* sequence. Thus, the overall alignment length compared to the *A. thaliana* sequence needs to be determined to identify usable reciprocal best BLAST hits pairs. I observed that reducing gap opening costs to 2 from the default value of 5 using the BLASTN online tool (blast.ncbi.nlm.nih.gov) gave a large uninterrupted HSP in many of these cases. However, using the same parameter values in the standalone BLAST did not result in uninterrupted HSPs. Using a perl script I parsed the alignment results and joined the HSPs for queries where alignments resulted in multiple HSPs. For genes that were identified as the best reciprocal hits, alignment length, their percentage identity and percentage length based on the length of *A. thaliana* sequences was calculated.

## Detecting Derived Sites

Each *A. thaliana* multiple aligned FASTA of protein-coding gene with the identified *A. lyrata* ortholog, was aligned to the *A. lyrata* sequence using the multiple sequence alignment tool, ClustalW (Larkin et al., 2007). These sequences were aligned, by increasing default gap open values to 25 and gap distances to 10. The alignment for each gene was trimmed so that no extra nucleotide positions were included compared to the *A. thaliana* Col-0 reference coding sequence. The nucleotide information at the *A. thaliana* Col-0 positions for synonymous and non-synonymous SNPs were collected using a Perl script (Table 3.1). Such Col-0 positions had been determined using POLYDNDS as described above. For each *A. thaliana* SNP, the corresponding *A. lyrata*

nucleotide was used as the ancestral nucleotide and if any of the accessions had a different nucleotide at that position it was counted as a derived site (Table 3.2).

Table 3.1: Nucleotide calls at synonymous and non-synonymous sites for all accessions and the ancestral sequence

| **Multiple aligned FASTA file** | **Synonmous/ Non-synonymous variant** | **Variant Position** | **Nucleotide in *A. lyrata*** | **Nucleotide in *A. thaliana*** | **Count** | **Nucleotide in *A. thaliana*** | **Count** |
|---|---|---|---|---|---|---|---|
| **AT3G15520.1.msa** | syn | 741 | G | A | 2 | G | 28 |
| **AT3G15520.1.msa** | syn | 831 | A | A | 4 | G | 26 |
| **AT3G15520.1.msa** | syn | 891 | A | A | 27 | T | 3 |
| **AT3G15520.1.msa** | syn | 1260 | G | A | 1 | G | 29 |
| **AT1G73200.1.msa** | syn | 111 | G | T | 1 | G | 29 |

Table 3.2: Derived site nucleotide at a site is the nucleotide different from the ancestor

| **Multiple aligned FASTA file** | **Synonmous/ Non-synonymous variant** | **Variant Position** | **Nucleotide in *A. lyrata*** | **Nucleotide at derived site** | **Derived site frequency** |
|---|---|---|---|---|---|
| **AT3G15520.1.msa** | syn | 741 | G | A | 2 |
| **AT3G15520.1.msa** | syn | 831 | A | G | 26 |
| **AT3G15520.1.msa** | syn | 891 | A | T | 3 |
| **AT3G15520.1.msa** | syn | 1260 | G | A | 1 |
| **AT1G73200.1.msa** | syn | 111 | G | T | 1 |

## Results and Discussion

### The Tajima's D Distribution Suggests Enrichment of the Genes Relieved from Strong Selection in the Immune Signaling Network

I compared Tajima's D distribution for all genes and the core immune signaling network genes to investigate whether evidence can be detected that suggests enrichment of genes under reduced selection among the immune signaling network genes. The distribution of Tajima's D values for all *A. thaliana* genes is unimodal with a mode around -1.5 (Figure 3.1, red curve). Note that all Tajima's D values in this chapter were calculated for each gene in the group of interest. This suggests that on average the *A. thaliana* genes are under purifying selection if the effective population size is assumed to be constant.

This observation corroborates a general belief that most mutations have deleterious effects and are removed from the population rapidly (Crow, 1997). When the Tajima's D distribution for the immune signaling network genes are overlaid (Figure 3.1, gray rectangles), it appears that genes with Tajima's D ~ 0 are enriched among the immune signaling network genes.

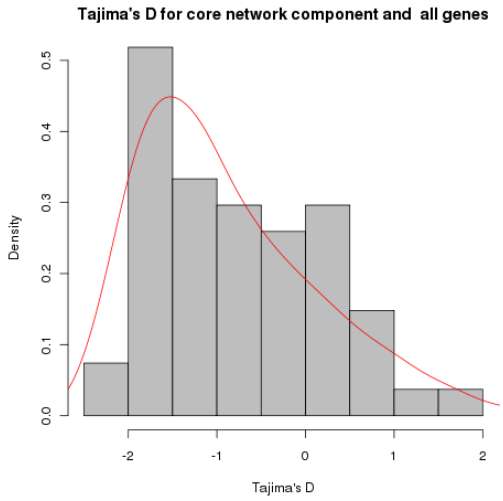


Figure 3.1: Tajima's D distributions for the core component genes of the immune signaling network (gray rectangles) and for all genes (red curve).

In contrast no tendency of enrichment was observed with the Tajima's D distribution for the bottom 10% genes (Figure 3.2) from the list of components, ranked by the betweenness centrality. It appears that the enrichment tendency is associated with highly ranked immune signaling network gene candidates, which are considered to be enriched with the network core component genes.

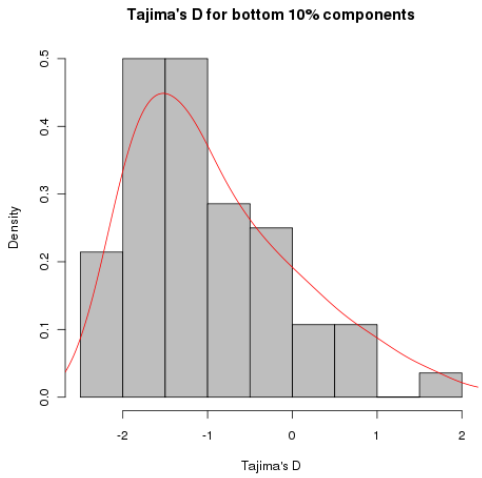


Figure 3.2: Tajima's D distributions for bottom 10% component immune network genes of the immune signaling network (gray rectangles) and for all genes (red curve).

Significance of this enrichment tendency among the immune signaling network genes was tested by dividing the genes into two categories according to Tajima's D values: those within (-0.6, 0.4) and the others. These boundary values were chosen *post-hoc* based on visual inspection of Figure 3.1. Fisher's exact test was applied to find the probability that the immune signaling network gene Tajima's D distribution is a random sample of the all gene distribution to be 0.025. This concludes that the enrichment of genes with Tajima's D ~ 0 among the immune signaling network genes is significant although there could be a possible overestimation of the significance due to the *post-hoc* nature of this test.

Existence of a gene functionally redundant with a particular gene would likely relieve the gene from strong purifying selection, as the deleterious effect of a mutation in the gene could be compensated by the other gene.

It was important to investigate whether or not the observed enrichment of genes with Tajima's D ~ 0 among the immune signaling network genes is due to enrichment of genes with functionally redundant genes in this group. Functionally redundant genes encode proteins that generally share a high level of sequence similarity. I identified genes that share close similarity between their encoded protein sequences, using BLASTP (Altschul et al., 1990) with a threshold of alignment E- value < 1e-10. From 27,025 genes, 6735 genes had close homologs (25%). Among all the genes, 4587 genes had Tajima's D < 0.6 and > -0.4 and 1132 genes from this subgroup were designated to have close homologs (25%). Of 55 core immune signaling network genes, 11 genes had close homologs (20%). Of 16 core network component genes with -0.4 < Tajima's D < 0.6, 3 genes had close homologs (18%). Therefore, neither the immune signaling network genes nor the network genes with Tajima's D ~ 0 have genes with close homologs enriched compared with all genes. Thus, I exclude the possibility that the reason these groups of genes have genes with Tajima's D ~ 0 is due to enrichment of genes with close homologs.

The Tajima's D Distribution Only Using the Non-Synonymous SNP Sites Substantiates the Enrichment Trend.

When the nucleotide change observed at a SNP site does not change the encoded amino acid residue, the site is called a synonymous SNP site, while when it does, it is called a non-synonymous SNP site. Although it is possible a synonymous change could affect the expression level of the protein (Hunt et al., 2009), generally a synonymous change is considered not to have a strong association with phenotypic change (Kimura, 1977). On the other hand, a non-synonymous change has a much higher chance to affect phenotype (Stenson et al., 2003). Therefore, I separated synonymous and non-synonymous SNP sites to calculate Tajima's D for each gene and examined the effect of this separation.

Figure 3.3 shows that Tajima's D has a clear bimodal distribution with the immune signaling network genes when only the non-synonymous sites are used for analysis (orange rectangles), while the corresponding distribution with all genes (Figure 3.3 red curve) does not change much from the unimodal distribution observed when non-synonymous and synonymous sites were not separated. One mode in the distribution with the immune signaling network genes is closely associated with the mode for the all-gene distribution, and the other mode corresponds to Tajima's D ~ 0.

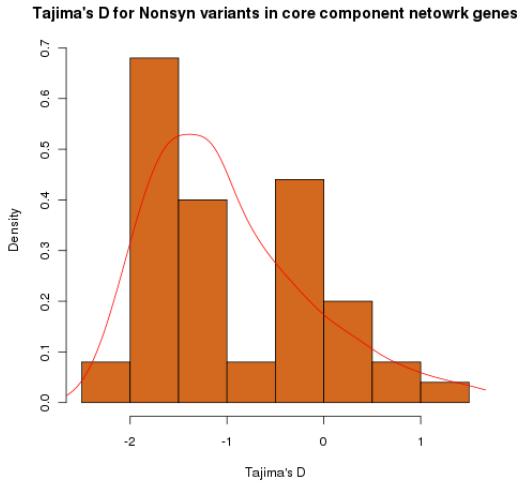


Figure 3.3: Tajima's D distributions for Non-synonymous variants in the core component genes of the immune signaling network (orange rectangles) and all genes (red curve).

Although the Tajima's D distribution with the synonymous sites in the immune signaling network genes may still have the bimodal tendency (Figure. 3.4), the tendency is not as clear as that with the non-synonymous sites. The distribution with the synonymous sites of all genes was unimodal and similar to that for the non-synonymous sites in all genes.

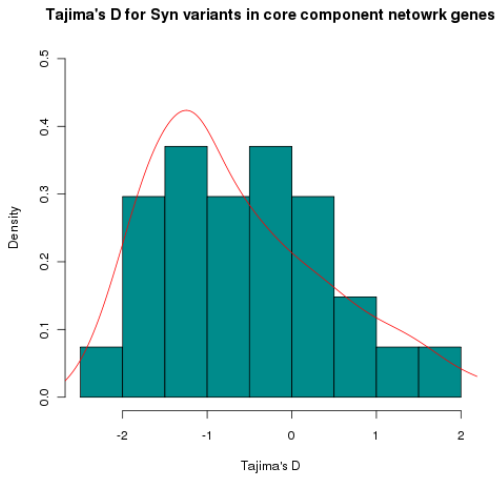


Figure 3.4: Tajima's D distributions for Synonymous variants in the core component genes of the immune signaling network (blue rectangles) and all genes (red curve).

From these results it can be interpreted that there are at least two groups selection-wise among the immune signaling network genes, which correspond to two modes of the Tajima's D distribution with the non-synonymous sites: one group is under purifying selection similar to the average of all genes, and the other group is under neutral selection. Since the linkage disequilibrium (LD) decay in the *A. thaliana* population is slow (Cao et al., 2011), the selective characteristics of the non-synonymous sites are probably represented among the synonymous sites at a reduced level: hence, the Tajima's D distribution with the synonymous site appears to have a weak bimodal trend

The significance of the bimodal distribution for the non-synonymous sites in the immune signaling network genes was tested. In this case, we chose to test for the bimodality instead of using the boundary values chosen *post-hoc*. The bimodality was

defined as the distribution that satisfy two conditions: there are at least two modes detected by a sliding window with a width of 0.6 and a step of 0.1; the difference between the second mode and the trough being the same as that in the distribution at the non-synonymous sites in the immune signaling network genes.

By random sampling with the sample size being the same as the number of the immune signaling network genes from the distribution with non-synonymous sites in all genes, the probability to obtain a sample to have as a clear bimodal pattern as the immune signaling network genes was 0.0043. Therefore, the bimodality of the Tajima's D distribution with non-synonymous sites in the immune signaling network genes is significant. We conclude that enrichment of genes with Tajima's D ~ 0 among the immune signaling network genes when analyzed with non-synonymous sites only is significant.

No Clear Difference in the Derived Site Frequency Spectrum Was Observed between the Immune Signaling Network Genes and All Genes (Figure 3.5 and 3.6). This may be because all the SNPs from the genes in each group were combined without adjusting the gene length, because the SNP sites at which the ancestral allele cannot be determined were removed from the analysis, and/or because the SNP sites at which information for some accessions are missing were removed from the analysis. Since I made a table in which information at all the SNP sites are compiled (Table 3.2), these potential issues can be investigated in the future.

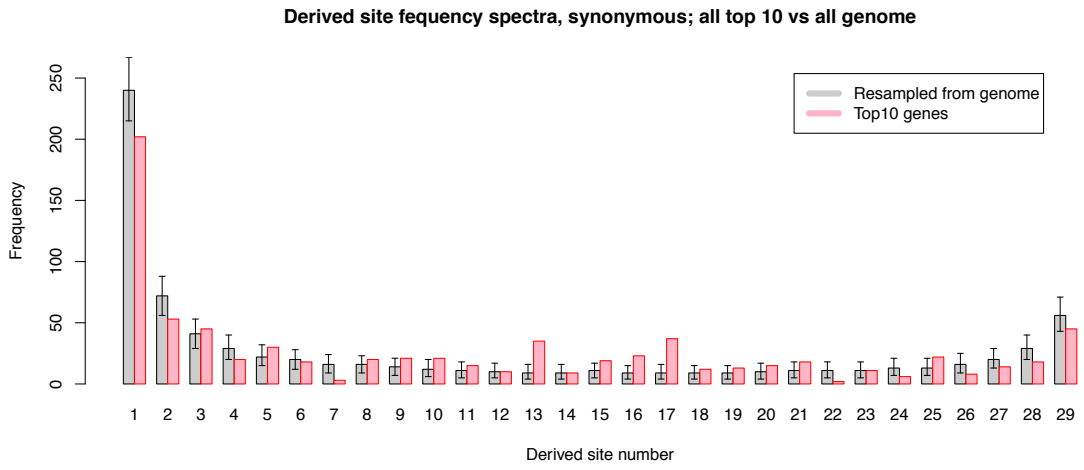


Figure 3.5: Derived site frequency spectrum for synonymous variants of all genes and core component immune network genes

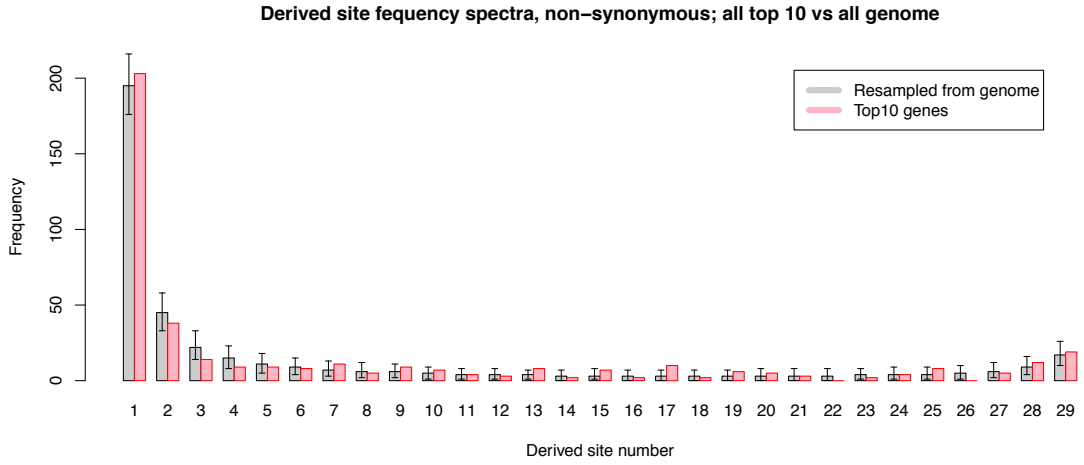


Figure 3.6: Derived site frequency spectrum for non-synonymous variants of all genes and core component immune network genes

## Conclusions

The Tajima's D distribution of non-synonymous sites suggests that the immune signaling network genes are enriched with genes that are selectively neutral, which is in line with the hypothesis of my thesis research.

The enrichment of neutrally selected genes in the network could mean that under a constant selection pressure, the immune signaling network can accumulate diversity, such as different network structures, within a single population without requiring isolation of subpopulations. In other words, the robustness of the network may buffer substantial changes of the network.

A crucial factor for a potentially pathogenic microbe to become a true pathogen of a particular host plant species is its ability to sufficiently impair plant's immunity. As microbes evolve much faster than plants in general, it is conceivable that plants suddenly have to face new pathogens that attack the plant immune system in different ways and that such sudden selection pressure changes may occur frequently in the evolutionary time scale. If the robustness of the immune signaling network, as discussed above, allows accumulation of a high level of network diversity in a single population, some subpopulations with certain network structures may be able to fight off such new pathogens well, which results in survival of the plant population. If this is the reason the plant immune signaling network maintains a high level of robustness against network perturbations, no other signaling networks in plants are expected to have a comparable

level of robustness.

My thesis work implicates that population genetic summary statistics for groups of genes composing different networks can be used to compare the robustness of the networks. My thesis work established a new, general approach that enables comparison of the robustness among different networks in this way. This was achieved by bringing population genomics data and functional gene network models together. Furthermore, particularly for *A. thaliana*, I compiled the genome diversity data set for 30 accessions, which will facilitate broad applications of this approach in the future.

# Bibliography

A. Block, G. Li, Z. Q. Fu, J. R. Alfano (2008), Curr. Opin. Plant Biol. 11, 396

Altschul S.F., Gish W., Miller W., Myers E.W. and Lipman D.J. (1990) Basic local alignment search tool. J. Mol. Biol. 215: 403-410.

C. Zipfel (2009), Early molecular events in PAMP-triggered immunity. Curr. Opin. Plant Biol. 12, 414

Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C et al. (2011) Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nat Genet 43: 956–963

Clark RM, Schweikert G, Toomajian C, Ossowski S, Zeller G, Shinn P, Warthmann N, Hu TT, Fu G, Hinds DA, Chen H, Frazer KA, Huson DH, Schölkopf B, Nordborg M, Rätsch G, Ecker JR, Weigel D (2007) Common sequence polymorphisms shaping genetic diversity in *A. thaliana*. Science, 317:338-342

Crow JF (1997) The high spontaneous mutation rate: is it a health risk? Proc Natl Acad Sci U S A 94: 8380–8386

Dangl, J.L. & Jones, J.D (2001). Plant pathogens and integrated defence responses to infection. Nature 411, 826−833

F. Katagiri, K. Tsuda (2010) Understanding the plant immune system Mol. Plant Microbe Interact., 23

Hu TT, Pattyn P, Bakker EG, et al.(2010) The A. lyrata genome sequence and the basis of rapid genome size change. Nat. Genet. 43:476–481.

Hunt R, Sauna ZE, Ambudkar SV et al. (2009) Silent (synonymous) SNPs: should we care about them? Methods Mol Biol: 578: 23–39.

Fay, J. C. and Wu C-I (1999) A human population bottleneck is compatible with the discordance between patterns of mitochondrial vs. nuclear DNA variation. Mol. Biol. Evol. **16**:1003-1005

Jones, J.D. & Dangl, J.L. (2006) The plant immune system. Nature 444, 323–329
Lamesch et al, (2011) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Research doi: 10.1093/nar/gkr1090.

Kimura, M. (1977) Preponderance of synonymous changes as evidence for neutral theory of molecular evolution. Nature 267, 275–276

Larkin M.A., Blackshields G., Brown N.P., Chenna R., McGettigan P.A., McWilliam H., Valentin F., Wallace I.M., Wilm A., Lopez R., Thompson J.D., Gibson T.J. and Higgins D.G. (2007) ClustalW and ClustalX version 2. Bioinformatics 23(21): 2947-2948.

Moreno-Hagelsieb G, Latimer K (2008) Choosing BLAST options for better detection of orthologs as reciprocal best hits. Bioinformatics, 24:319-324.

Nei M, Maruyama T, and Chakraborty R (1975) The bottleneck effect and genetic variability in populations. Evolution **29**:1–10.

Nielsen R, Wakeley J. (2001) Distinguishing migration from isolation. A Markov chain Monte Carlo approach. Genetics 158:885-896.

Ossowski, S. Schneeberger K, Clark RM, Lanz C, Warthmann N, Weigel D (2008). Sequencing of natural strains of *A. thaliana* with short reads. Genome Res. 18, 2024–2033.

Quinlan AR and Hall IM, (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26, 6, pp. 841–842.

Rice,P. Longden,I. and Bleasby (2000), A EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics 16, (6) pp276—277

Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, et al. (2003) Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat 21: 577–581

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. Genetics 105: 437–460

Tajima, F. (1989). The effect of change in population size on DNA polymorphism. Genetics **123**:597–601.

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

Thornton, K. (2003). Libsequence: A C++ class library for evolutionary genetic analysis. Bioinformatics 19: 2325-2327.

Tsuda K, Katagiri F (2010) Comparing signaling mechanisms engaged in pattern-triggered and effector-triggered immunity. Curr. Opin. Plant Biol. 13: 459–465.

Watterson, G.A. (1975), "On the number of segregating sites in genetical models without recombination.", Theoretical Population Biology 7 (2): 256–276

Weigel D, Mott R. (2009) The 1001 genomes project for *A. thaliana*. Genome Biol 10:107