Genomics Proteomics Bioinformatics 11 (2013) 345–353

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

ORIGINAL RESEARCH

Identification of Immunity-related Genes in *Arabidopsis* and Cassava Using Genomic Data

Luis Guillermo Leal [1,#], Álvaro Perez [2,#], Andrés Quintero [2], Ángela Bayona [2], Juan Felipe Ortiz [2], Anju Gangadharan [3], David Mackey [3], Camilo López [2], Liliana López-Kleine [1,*]

[1] *Department of Statistics, Universidad Nacional de Colombia, Bogotá 111321, Colombia*
[2] *Department of Biology, Universidad Nacional de Colombia, Bogotá 111321, Colombia*
[3] *Department of Molecular Genetics, The Ohio State University, Columbus, OH 43210, USA*

Received 12 June 2013; revised 19 September 2013; accepted 22 September 2013
Available online 6 December 2013

KEYWORDS

Arabidopsis;
Cassava;
Functional gene prediction;
Genomic data;
Kernel canonical correlation analysis;
Plant immunity

Abstract Recent advances in genomic and post-genomic technologies have provided the opportunity to generate a previously unimaginable amount of information. However, biological knowledge is still needed to improve the understanding of complex mechanisms such as plant immune responses. Better knowledge of this process could improve crop production and management. Here, we used holistic analysis to combine our own microarray and RNA-seq data with public genomic data from *Arabidopsis* and cassava in order to acquire biological knowledge about the relationships between proteins encoded by immunity-related genes (IRGs) and other genes. This approach was based on a kernel method adapted for the construction of gene networks. The obtained results allowed us to propose a list of new IRGs. A putative function in the immunity pathway was predicted for the new IRGs. The analysis of networks revealed that our predicted IRGs are either well documented or recognized in previous co-expression studies. In addition to robust relationships between IRGs, there is evidence suggesting that other cellular processes may be also strongly related to immunity.

Introduction

Recent advances in genomic and post-genomic technologies have provided the opportunity to generate vast datasets. However, the data stored in genomic databases does not itself provide an understanding of biological processes and has not always been generated under biological conditions of interest. Nevertheless, available data could be combined with own data generated in-house for the biological condition of interest to improve results and generate more confident biological conclusions. The new challenge is to develop mathematical methods

* Corresponding author.
E-mail: llopezk@unal.edu.co (López-Kleine L).
# Equal contribution.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

ELSEVIER Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.gpb.2013.09.010

to assess biological problems or phenomena from a holistic or system-level perspective and to use own and other information available. Approaches to extract knowledge from genomic databases and combine these data with new experimental data should allow the integration, interpretation and analysis of genomic and post-genomic data and should represent the acquired biological knowledge in the form of gene or protein networks showing functional/co-expression relationships or other structured representation. This representation should reflect relationships at the individual and categorical levels, which would assemble genes/proteins of known, unknown and hypothetical functions.

Several different approaches have been developed in recent years to assess relationships between functionally known and unknown genes/proteins through biological networks and predict new functions of genes/proteins, especially in humans [1]. These methods are often supervised and allow the integration of multiple genomic data sources in different ways [2,3], thus generating reliable and robust results, often including the prediction of new protein functions [4]. Due to the specific and complicated characteristics of genomic data, proper analysis and generation of useful inference represent real mathematical and statistical challenges.

Predictions of function are better conducted using methods that allow the integration of prior knowledge (supervised methods), the identification of non-linear relationships and the fusion of heterogeneous genomic and post-genomic data. Kernel methods [5] have these characteristics and among them, kernel canonical correlation analysis (KCCA) can be useful in relating proteins of known function with those of unknown function to predict participation in processes of interest. Earlier studies have reported the use of KCCA methods to predict the functions of unknown proteins [4,6,7]. KCCA offers a rigorous mathematical but also intuitive framework to represent biological data through kernel functions [4,5]. KCCA provides a methodology for supervised network inference and does not require exhaustive data assumptions [8]. It is therefore in contrast to alternative strategies such as Naïve Bayes (NB) models [9], which require regularization methods and have challenges of computational efficiency in the presence of many data sets [10].

Losses caused by plant pathogens represent one of the most important limitations in crop production, which can compromise the food supply [11]. Plant immunity depends on the recognition of conserved microbe-associated molecular patterns (MAMPs) or strain-specific effectors by pattern recognition receptors (PPRs) or resistance (R) proteins, leading to MAMP-triggered immunity (MTI) and effector-triggered immunity (ETI), respectively [11,12]. Upon recognition, plants activate a complex network of responses that includes signal transduction pathways, novel protein interactions and coordinated changes in gene expression [13]. Detailed information concerning specific and punctual interactions between effector and resistance proteins has been accumulated in the recent years; in some cases, a global picture for some of these interactions has been established [9,14]. Immunity networks have been described for model plants such as *Arabidopsis* and rice primarily using yeast-two hybrid experiments [15,16].

In this study, we employed a kernel-based approach to reconstruct functional relationships between genes based on genomic and post-genomic data from various sources (primarily extracted from databases but also produced by laboratory experiments) for a group of well-characterized immunity-related genes (IRGs). We employed this approach to analyze *Arabidopsis* and cassava (*Manihotesculenta*), a staple crop with little genomic information available, following challenge with bacterial pathogens. This approach allowed us to identify a group of new IRGs in both species. Many of the identified genes were of unknown function. Based on our further detailed analyses and literature knowledge, we established a list of top gene candidates potentially related to immune responses. These results indicate that publically-available data can be combined with in-house generated data using novel data-mining methods to potentially answer challenging biological questions.

Results

Exploratory analysis of categorical data

A total of 22 datasets were collected for *Arabidopsis* and cassava (see Materials and methods section for more details). Number of genes and the number of columns for each dataset are listed in Tables S1 and S2. To obtain a preliminary architecture of the data, we conducted classical descriptive multivariate analyses using multiple correspondence analysis (MCA), clustering and principal component analysis (PCA) [17] as a first step to evaluate the data structure, reveal unknown relationships and reveal clusters of genes potentially involved in immune responses. Our results showed that no groups of IRGs were clearly detected, indicating that functional relationships cannot be extracted using linear descriptive methods. Nevertheless, we were able to summarize the information of microarray data with fewer variables using an exploratory descriptive analysis. We found that most of the information contained in the microarrays is correlated and can be represented with two new variables (principal components). Accordingly, only a small portion of genes have different expression behaviors across experiments, which could be new IRGs. Furthermore, we found that RNA-seq data contains information that complements the microarray data. These results are useful and indicate that expression data contains valuable information to differentiate IRGs from non-IRGs if a more appropriate method is implemented.

All in all, the exploratory analyses showed that IRGs cannot be grouped together using only linear methods and methods such as KCCA (introduced in following section) are desired. For details on the procedure and the results of exploratory analyses, see the Supplementary File 1.

Relationship between genes/proteins obtained using KCCA

Since linear relationships between gene expression variables did not show any structure or pattern that allowed the grouping of IRGs based on either categorical or continuous data, we used non-linear kernel methods to integrate both types of data for extraction of relationships between genes. We used the supervised KCCA method [6] to predict functional relationships between genes. To do this, two reference datasets were used in the KCCA, including the real reference dataset and a random reference dataset of IRGs constructed by randomly placing a similar number of IRGs from the real reference in five categories to emulate five types of IRGs.

Table 1 Threshold and percentage of correct predictions using KCCA

Reference dataset	*Arabidopsis*		Cassava	
	Correct predictions (%)	Threshold	Correct predictions (%)	Threshold
Real	74	28.2	72	55.6
Random	61	53.0	57	77.5

The KCCA allowed us to project the genes in a new space and to assess distances between IRGs and other genes. We predicted "partners" or new IRGs per each known IRG. New IRGs were identified when they were projected closer to known IRGs with a chosen distance threshold (**Table 1**). This procedure provided a network and therefore a list of partners per IRG. Networks were drawn using Cytoscape 3.0 [18] for *Arabidopsis* and cassava (**Figure 1**).

Some types of interactions in the networks were identified (**Figure 2**). These include direct interaction between a known IRG with another known IRG or newly-predicted IRG (Figure 2A). Indirect interaction between two known IRGs was also observed via bridging effect of a newly-predicted IRG (Figure 2B). In addition, ternary direct interaction among known IRGs was also noticed (Figure 2C). The statistics of these types of interactions are summarized in **Table 2**. KCCA relationships and their interpretation showed differences between species. There are 19 partners on average for each IRG in *Arabidopsis* and 30 in cassava. However, in *Arabidopsis*, IRGs were mainly connected to other IRGs, where as such pattern was not observed in cassava. Moreover, the global clustering coefficient for *Arabidopsis* was much higher than that for cassava, showing high connectivity among IRGs. The results obtained using random reference datasets showed different patterns, compared to the real reference datasets of IRGs. In both species, the average number of partners per IRG regarding the total interaction was much higher. In addition, 20,305 non-IRGs in *Arabidopsis* were predicted to be partners of IRGs, which include almost all genes in this species (approximately 27,000 genes in total). In cassava, a less dense network was generated, possibly due to the low number of microarray experiments that only cover a small spectrum of conditions. These results indicate that our method showed higher selectivity with real datasets than with random datasets, thus the identified interactions would be unlikely random but instead specific.

The average degree of the nodes and the global clustering coefficients are plotted in **Figure 3**. High level of connectivity is detected when an IRG is excluded from the network and both the average node degree and clustering coefficient decrease (large downward peaks for the same IRG). According to our analysis, some IRGs and their predicted neighbors are highly connected with each other. IRGs that have a low number of partners (small downward peak in Figure 3, left panels for average node degrees) and belong to many triplets (large downward peaks in Figure 3, right panels for clustering coefficients) are thought to have highly specific interactions to form small clusters. In contrast, IRGs having many partners (large peaks in Figure 3, left panels) could be identified as hubs in the network. Median (middle line) in boxplots in Figure 3 shows that *Arabidopsis* IRGs (top panels) have fewer but more highly interconnected partners than cassava IRGs (bottom panels).

Common features in the predictions for both species

Fisher's exact test [19] was applied and it allowed us to identify 89 and 87 GO terms that were overrepresented in the networks from *Arabidopsis* and cassava, respectively. Among them, 61 terms were overrepresented in both networks, suggesting that the genes identified in both plants are functionally similar. The most overrepresented terms in both networks included various types of kinase activity, stress responses, immune responses and processes related to cell death. However, these

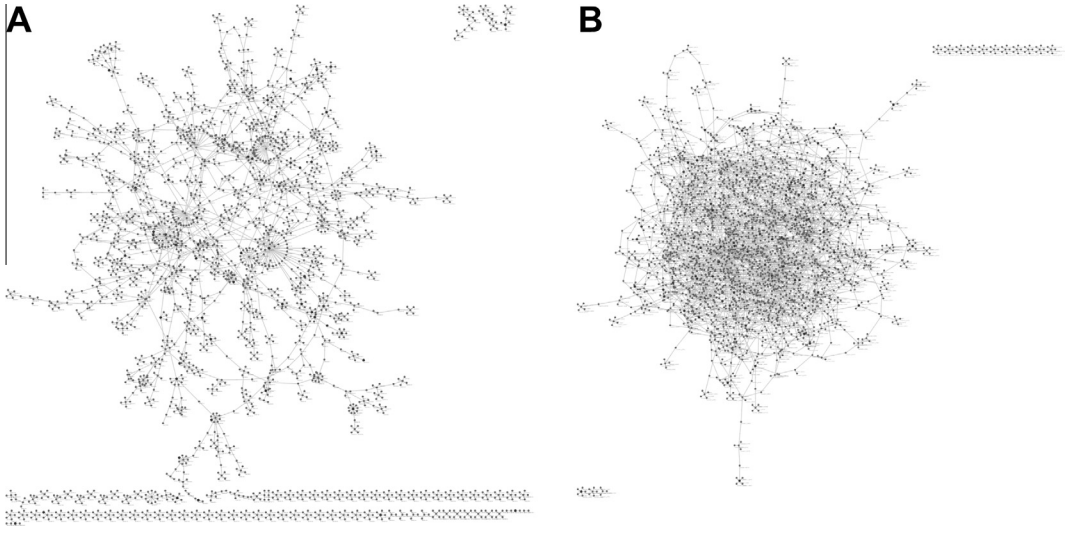

Figure 1 IRG networks for *Arabidopsis* and cassava

Network representation of functional relationships obtained for *Arabidopsis* (**A**) and cassava (**B**). Representations were plotted using Cytoscape 3.0. Genes coding for LRR or Pkinase-domain-containing proteins were excluded from representation. Only the top five closest partners of each gene are shown.

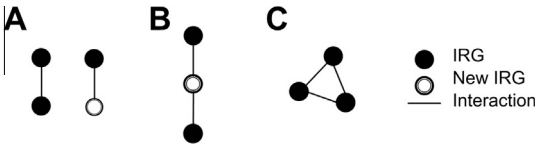

Figure 2 Types of interactions in IRG networks
There are mainly three types of interactions between genes, including simple interactions (**A**), indirect interaction (**B**) and triplets (**C**).

data should be taken with caution, since some genes sharing the same GO category could be orthologous genes.

Analysis of predicted relationships in Arabidopsis

We then focused on some of the most important IRGs and performed a detailed analysis of the predicted partners of 12 well-known IRGs described in the literature and gene network databases (**Table 3**). Based on the data, we were able to show that BRI1, for example, is one of the partners of BAK1. BAK1 is a regulator of the tradeoff between immunity and responses to hormones [20] and a co-receptor of FLS2 that triggers plant immunity after the recognition of flagellin [21]. Interestingly, previous studies indicated that BAK1 interacts with BRI1, a receptor for the growth hormone brassinosteroid [22]. Therefore, our prediction is strengthened with the biological data reported in the literature. Similarly, we found that CLV2 was among the top 10 partners for CLV1. CLV1 is a receptor kinase expressed in the center of the shoot apical meristem and interacts with CLV2 and other proteins to control meristem development [23]. Furthermore, a link between meristem development and plant immunity in the shoot was recently established [24]. A third example is extracted from the partners of CERK1. CERK1 is achitin receptor that triggers a response to fungi [25]. *Arabidopsis* plants expressing a mutant *CERK1* also exhibited compromised resistance to bacteria [26]. Among the top 10 partners for CERK1 are the genes *NDF4* and *GSTL2*, which code for an electron carrier and a glutathione S-transferase, respectively. The expression of these two genes is co-regulated, as reported in the CoEXpression network [27]. GSTL2 is a protein involved in the redox balance and the metabolism of reactive oxygen species (ROS), which are central to plant immunity. Besides, ROS production is one of the primary responses mediated by CERK1 [25].

In addition to analyzing the data presented in the literature, we also searched for the functions of the top 10 partners through BLASTX to complement the information about relationships of predicted IRG partners. Based on BLASTX results, a putative function in the immunity pathway was

Table 2 Statistics of the networks reconstructed from KCCA

Dataset	Species	No. of IRGs in the network	No. of new IRGs (predictions)	Ratio	Average No. of interactions per IRG		Total No. of interactions per IRG	GCC ($\times 10^{-4}$)
					With other IRGs	With new IRGs		
Real	*Arabidopsis*	1606	6085	1:4	5	14	19	7
Real	Cassava	2272	3340	1:2	2	28	30	1
Random	*Arabidopsis*	1606	20,305	1:13	11	137	148	39
Random	Cassava	2272	1464	2:1	100	65	165	224

Note: GCC stands for global clustering coefficient.

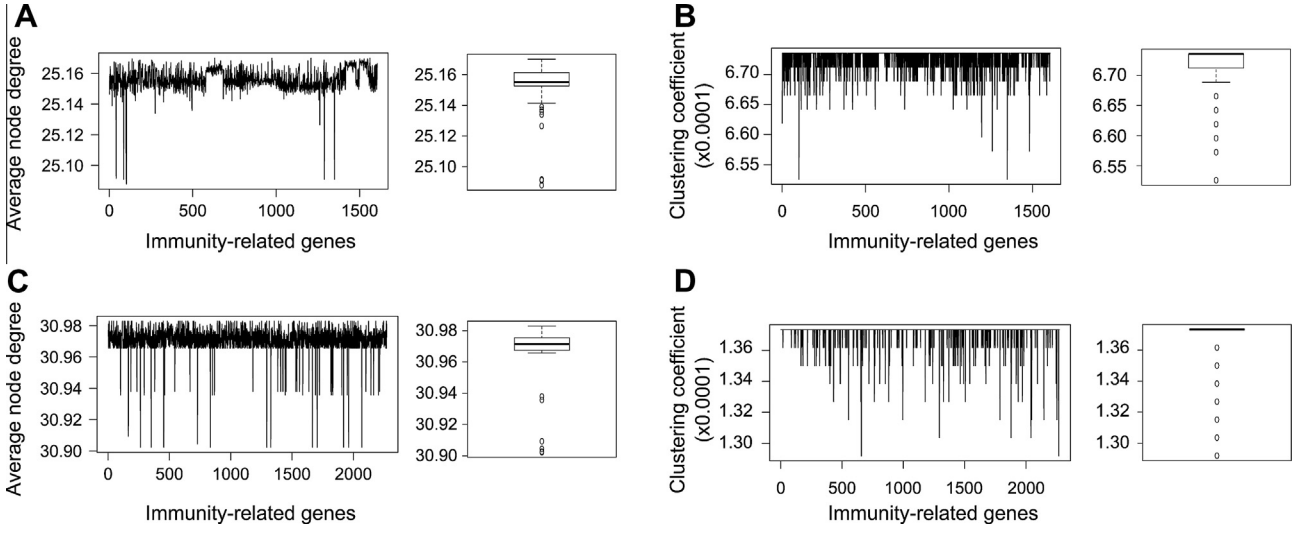

Figure 3 Average degree and clustering coefficient for *Arabidopsis* and cassava networks
Average node degree (left panels) and clustering coefficient (right panels) are shown in plot and boxplot. The top panels indicate the data for *Arabidopsis* (**A, B**) while the network analysis for cassava is shown in the bottom panels (**C, D**). Both variables were calculated by removing one IRG at each step. The negative peak appears when a highly connected or clustered IRG is removed from the network.

Table 3 A selection of well-known *Arabidopsis* IRGs described in the literature and gene network databases

Name of IRG	ID of IRG	Summary	The 10 closest partners
RPS2	*AT4G26090*	Confers resistance to *Pseudomonas syringae* strains that carry the avirulence gene *avrRpt2*	*AT3G28620, TRFL1, ATCG00570, ATATG18G, DREB1A, PDF1A, AT3G46070, ACOS5, AT1G61190, AT4G08850*
RPM1	*AT3G07040*	Confers resistance to *Pseudomonas syringae* strains that carry the avirulence genes *avrB* and *avrRpm1*	*AT1G72580, AT5G11700, AT3G06035, ATCNGC7, AT4G00940, AT3G56130, VPS28–2, HAP13, CRA1, scpl29*
WRKY31	*AT4G22070*	Member of WRKY transcription factor family; group II-b	*AT3G27090, anac052, AT3G51050, NBP35, ATEXO70E1, AT1G14150, AT2G03500, AT3G14800, AT4G18250, AT5G39020*
MPK9	*AT3G18040*	Expressed preferentially in guard cells and appears to be involved in reactive oxygen species-mediated ABA signaling	*AT3G06610, AT3G13980, IQD12, AT3G05280, HCF153, AT3G46610, ERF1–2, AT2G41820, AT5G09890, AT4G31110*
BAK1	*AT4G33430*	Leucine-rich receptor serine/threonine protein kinase; component of BR signaling that interacts with BRI1 *in vitro* and *in vivo* to form a heterodimer	*AT5G01350, AT4G03820, AT4G33780, AT1G23280, SPK1, SULTR3;2, AT4G37090, AT5G56790, AT3G22800, BRI1*
WRKY33	*AT2G38470*	Regulates the antagonistic relationship between defense pathways mediating responses to *P. syringae* and necrotrophic fungal pathogens	*AT1G52100, AT1G19370, AT1G77090, AT1G76070, DegP13, ATRER1A, AT1G57570, AT1G36980, AT5G58790, AT5G47435*
FLS2	*AT5G46330*	LRR receptor-like serine/threonine-protein kinase, recognizes peptide from flagellin	*AT5G61520, AT1G47710, AT5G65960*
CERK1	*AT3G21630*	LysM receptor-like kinase; essential in the sensing and transduction of the chitin oligosaccharide elicitor	*AT2G18720, AOX1A, AT3G07020, AT4G03153, AT1G72430, NDF4, ATPANK1, GSTL2, AT2G30940, AtRLP24*
CLV1	*AT1G75820*	Putative receptor kinase with an extracellular leucine-rich domain. Controls shoot and floral meristem size	*AT1G72070, AT5G17760, UBC9, AT5G26360, CLV2, AT4G09150, AT1G63280, AT2G33600, AT3G43890, AT1G24650*
RPS4	*AT5G45250*	TIR-NBS-LRR class disease resistance protein	*APX3, BAM1, AT1G20530, ATKDSA2*
ER	*AT2G26330*	Homologous to receptor protein kinases; contains a cytoplasmic protein kinase catalytic domain, a transmembrane region and an extracellular LRR	*AT2G20110, TUBG1, TSD2, AT4G04170, AT2G38000, AT1G17210, AT3G06540, SWI2, AT1G63110, AT2G21160*
RPP13	*AT3G46530*	Confers resistance to the biotrophicoomycete, *Peronosporaparasitica*; encodes an NBS-LRR type R protein with a putative amino-terminal leucine zipper	*CRF1, AT2G25610, AT1G18700, AT2G47970, AT2G34300, NUB, AT5G37570, AT5G43680, BZIP34, AT5G47250*

Note: The summary of IRGs was obtained from NCBI (http://www.ncbi.nlm.nih.gov/guide/genes-expression/).

assigned to 51 of 107 predicted partners in *Arabidopsis* (see Supplementary File 2). This information is also useful to filter the partners and to select the best ones for future experiments that employ the corresponding *Arabidopsis* mutants. The BLASTX results allowed us to reinforce the results of our predictions. An example is given for RPS2. RPS2 is a very well-known NBS-LRR resistance protein that mediates resistance to strains of *Pseudomonas syringae* expressing the avirulence protein avrRpt2 [28]. We identified important genes that are related to plant immunity among the top 10 partners for this particular R gene. These include ACOS5, transcription factor DREB1A and *AT1G61190*, an NBS-LRR-coding gene. ACOS5 is a protein carrier involved in the reinforcement of the cell wall and in vesicular trafficking [29]. Vesicular trafficking is proposed to transport specific enzymes involved in the production of compounds such as 1,3 b-glucans to reinforce the cell wall and prevent colonization by the pathogen [29,30]. In addition, DREB1A is involved in the response to dehydration and in the response of *Arabidopsis* to *Hyaloperonospora arabidopsidis* [31], indicating that this protein plays a role in responses to both biotic and abiotic stresses. In addition, *AT1G61190*, an NBS-LRR-coding gene was also among the top 10 partners of RPS2, suggesting a network connection between proteins of this large class of resistance proteins.

Analysis of predicted relationships in cassava

We analyzed the predicted IRGs in cassava to identify their roles in defense. In particular, we analyzed the gene *RXam2* (cassava4.1_031234 m), which is an NBS-LRR gene that colocalizes with a quantitative trade loci (QTL) associated with resistance to *Xanthomonas axonopodis* pv *manihotis* [32,33]. RXam2 was predicted to be associated with a serine-threonine protein kinase (cassava4.1_027765), an NAD-dependent epimerase (cassava4.1_023284) and a transcription factor (cassava4.1_014914). Associations with this type of gene were commonly found for known immunity genes in *Arabidopsis*. Six of the *Arabidopsis* partners analyzed in detail were associated with transcription factors and serine threonine kinases, and among them, CLAVATA was associated with an NAD-dependent epimerase. These interactions suggest that RXam2

could be involved in pathways in cassava that are similar to those in *Arabidopsis* (Supplementary File 3).

Discussion

We were able to gather a large amount of genomic and post-genomic information from public and private databases on both *Arabidopsis* and cassava, with each dataset represented as a kernel matrix to apply KCCA. In the newly-projected space, we calculated relationships between IRGs and other genes to predict a potential IRG network. We showed, after a functional analysis, that these relationships are useful as a starting point to predict potential IRGs or genes coding for proteins strongly related to immune processes in both plant species.

The quality of genomic information depends primarily on genome annotation and the connection of this information to other databases, including literature data. The data on the model plant *Arabidopsis* are much more reliable than the data on cassava due to more genomic data available and better genome annotation. The quality of predictions and the ability to interpret them biologically also depends on the quality of the data. The level of confidence in the predictions, independent of the estimated statistical error, is higher for *Arabidopsis*. Predictions for *Arabidopsis* can be verified and explored to generate biological hypotheses for validation, although predictions for cassava should be taken more cautiously. Nevertheless, the predictions and findings for cassava are very valuable because they constitute one of the first predictions for this plant species.

The quantity of data is often directly proportional to the amount of redundancy found in the databases [17], which can be observed in the preliminary cluster analysis and PCA. In both cases, the original variables were reduced more efficiently for *Arabidopsis* than for cassava, suggesting that variables of *Arabidopsis* are more correlated. In the PCA we found that microarray and RNA-seq data behave orthogonally, meaning that, for the same genes, the gene expression levels measured using these two techniques are different. Although very few RNA-seq experiments were used, this result could indicate that these gene expression measurements are indeed complementary and do not contain the same type of information. RNA-seq is a novel strategy to obtain information about gene expression, and although not much information has been generated by this strategy, some data suggest that there is not a direct relationship between the data from microarrays and the data from RNA-seq. The percentage of information obtained by the linear PCA was more or less maintained in the KCCA, where RNA-seq was calibrated to a weight of 0.01, in contrast to the microarray data, which had a weight of 0.6. The applied procedure suggests that the reconstruction of biological networks is successful when data from different sources are used but it is important to weigh the importance of each variable.

We investigated the importance of the reference category by conducting predictions based on a random reference category. We obtained ~61% of correct predictions (Table 1) when the random reference was used in *Arabidopsis*. Accordingly, the predictions are almost random and less accurate when a random category is used.

Other studies [15,16,27] have shown that the accuracy of functional predictions is relatively high. For example a rice network proposed by Lee and colleagues [16] allowed the prediction of 14 genes involved in XA21-mediated immunity, and 3 of which were in fact validated biologically to be important for plant defense against *Xanthomonasoryzae* pv. oryzae. Here we obtained a prediction precision of five new genes among the top 50 candidates for each biological process. For *Arabidopsis*, a pathogen stress network modeled by Atias et al. [27] was theoretically validated through GO enrichment and a cluster was revealed, in which 8 of 45 genes were associated with the "response to biotic stimulus" and "defense response" GO terms. Additionally, Mukhtar et al. [15] experimentally validated 9 of 18 proteins predicted to be targets of effectors from two pathogens for *Arabidopsis*. Although our strategy was different, we expect that a relatively high percentage of the predicted genes in this study are most likely important in plant immunity.

The overall shape and topological features of the obtained IRG networks were different between *Arabidopsis* and cassava. The average number of interactions between IRGs was higher in *Arabidopsis* than in cassava (Table 2). Furthermore, in *Arabidopsis*, a higher global clustering coefficient of IRGs was observed (Figure 3C and D). These data could have a biological meaning suggesting that genes involved in immune processes are a much more defined group in *Arabidopsis* and other cellular functions (*i.e.*, metabolic functions) are less involved in immune processes in *Arabidopsis* than in cassava. Thus, in cassava, the IRGs seem to be connected with more non-IRGs. Nevertheless, this conclusion needs to be used with caution, since it could be due to the lack of information on non-IRGs in cassava.

Some summary statistics were calculated when each IRG was removed from the network (Figure 3). The relatively fewer negative peaks for degree of nodes are an indicative that *Arabidopsis* network seems to have fewer hubs than cassava network, while IRGs from *Arabidopsis* appear to be better clustered than in cassava as shown by the higher clustering coefficient. Again, biological explanations of these topological differences should be taken cautiously. Effector proteins are directed to hubs of plant immunity networks [15]. Consequently, an interpretation is that *Arabidopsis* network has fewer hubs but removing of them as is done using the tolerance algorithm does not obviously affect the overall connectivity. Thus, the immunity network in *Arabidopsis* might be considered as a robust or tolerant network against attacks, where other IRGs can be imputed the same functional relationships of those IRGs suppressed. On the other hand, the clustering coefficient of the cassava network is reduced when IRGs are removed. Therefore, immunity processes in cassava could be more vulnerable to be fragmented when hubs are preferentially attacked.

In-depth analysis of for 12 well-known IRGs and their predicted interaction partners in *Arabidopsis* yielded interesting findings. Our results indicate that strong relationships between IRGs exist and that other cellular processes are also strongly related to immunity. Partners either well documented or proposed in previous co-expression studies were verified by our predictions. The putative functions of some partners were recognized inside the immune pathway based on BLASTX searches and biological annotations.

The methodology applied in this study allowed constructing networks which would be useful for functional prediction in

Arabidopsis and cassava. Although the data quality was very different from the beginning, predictions in both species would facilitate generating new biological hypotheses for further investigation.

Materials and methods

Construction of genomic datasets

A reference dataset of IRGs was constructed with the genes coding for canonical immune protein domains (WRKY, TIR, NBS, kinase and LysM). These domains were downloaded from Pfam [34] and searched in the proteomes of *Arabidopsis* and cassava. HMM search [35] was used to examine the occurrence of these domains using an e-value of 1E-10 and the default parameters [36]. The reference dataset for *Arabidopsis* was complemented with a graph of 119 IRGs extracted from BAR (http://bar.utoronto.ca/welcome.htm). To test the reliability of the reference dataset, we constructed a random reference dataset by randomly assigning genes to five non-sense categories (emulating five types of IRGs).

Other genomic datasets were obtained as follows. Categories obtained from KEGG (http://www.genome.jp/kegg/) were used to construct a dataset indicating the participation of a gene product in all metabolic categories from this database. The cellular localization of the proteins was assessed by searching the proteomes of *Arabidopsis* and cassava for signal peptides using the program TargetP, including ChloroP and SignalP [37]. The assigned GO ID (gene ontology), KOG ID (eukaryotic orthologous groups) and PfamID (protein families) for proteins of both species were queried using a BioMart tool accessible from the Phytozome project version 7.0 (http://www.phytozome.net). GO annotations for cassava are not currently available. Data for experimentally validated miRNA target genes from *Arabidopsis* were obtained from the MPSS *Arabidopsis* PARE Database [38]. Target genes for the identified miRNAs in cassava were obtained from predictions made by Pérez-Quintero and colleagues [39]. Sequences up to 1000 nucleotides upstream of each identified cassava and *Arabidopsis* gene was extracted from the Phytozome database v7.0 (www.phytozome.net) for identification of transcription factor binding sites (TFBS) as described by Megraw and Hatzigeorgiou [40] using Java scripts.

Construction of post-genomic datasets

The *Arabidopsis* microarray dataset related to pathogen resistance was obtained by downloading GEO datasets, which are publicly available at the National Center for Biotechnology Information (NCBI) website (http://www.ncbi.nlm.nih.gov/). In addition, genes present in these datasets were assessed with the TAIR gene annotation (also publicly available at http://www.arabidopsis.org). A total of 51 datasets related to pathogen resistance were collected for analysis (see Supplementary File 3). The cassava microarray dataset was obtained from a previous study [41].

The RNA-seq data for *Arabidopsis* were generated at the Ohio State University using libraries obtained from *Arabidopsis* plants (wild-type Col-0 (C) and the R-gene mutant (rps4–1)). Plants were either hand-inoculated with *Pseudomonas syringae* pv *phaseolicola* strain NPS3121, which expresses the AvrRps4 resistance protein recognized by RPS4, at 1×108 CFU/mL [42] or mock inoculated and plant stem tissues were collected at three different time points post-inoculation (6, 12 and 24 h). RNA-seq data for cassava were generated by the Manihot Biotech research group at the Universidad Nacional de Colombia (Muñoz et al., unpublished) using libraries obtained from cassava leaf tissues inoculated with *Xanthomonas axonopodis* pv *manihotis* (*Xam*) or a *Xam* strain lacking the TAL effector, $TALE_{Xam1}$.

The quality of RNA-seq libraries was evaluated using FastQC and in-house Perl scripts (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads were unpaired and 50 bp in length. Adapters and reads containing "N"s were removed. Reads with more than 50% of their bases having Phred scores lower than 30 were also removed.

Sequencing reads were mapped onto coding sequences from the respective plant genome obtained from Phytozome v7.0 (www.phytozome.net) using seq-map [43] with no mismatches, and the final expression values (RPKM) for each gene were obtained using R-seq with default parameters [44]. RPKM values were used for the network.

Data preprocessing

Microarray and RNA-seq RPKM data were normalized using the R [45] vsn library through the glog transformation proposed by Huber [46].

Kernel canonical correlation analysis

Kernel canonical correlation analysis (KCCA) uses kernels (similarity matrices between objects, here genes) for each type of data to conduct a regularized canonical correlation analysis [5]. We used polynomial kernels for categorical data, Gaussian radial basis function (RBF) kernels for continuous data and diffusion kernels for graphs as described previously [4,6]. The sigma parameters for the RBF kernel, regularization for the diffusion kernel and degree for the polynomial kernel were obtained by leave-one-out cross validation following Yamanishi et al. [6] (Tables S1–S3).

To combine genomic data, we fused kernels by weighted addition [6]:

$$K = w_1K_1 + w_2K_2 + ... + w_iK_i, \quad (1)$$

where K_i denotes the kernel and w_i denotes the weight of each kernel. The weights must add up to 1 to preserve the value on the diagonal of the final kernel.

KCCA was conducted and their parameters were obtained by cross validation (Tables S3, S4 and S5). We calculated distances between genes in the new space obtained by KCCA. Gene "predictions" were retained when they were under a chosen distance threshold, which was calculated as the 25% of the maximum distance between IRGs (Table 1).

A theoretical percentage of correct predictions was calculated with the KCCA results (Table 1). The percentage reflects how closer the IRGs in the new space are. Therefore, a high percentage is an indicative that KCCA accurately

reconstructs the functional relationships between IRGs. The percentage was calculated as follows. (i) The distances between genes were arranged in increasing order and ranked. For example, if 100 genes (30 IRGs and 70 other genes) were projected in the new space, we calculated and ordered the 10,000 distances between them. (ii) A rank threshold for IRGs distances was assessed. For example, 900 distances were obtained if 30 IRGs were projected and thus, the rank threshold is 900 in this example. (iii) The distances between IRGs that were under the threshold rank were considered correct. For example, if only 90 distances between IRGs were found under the rank of 900, we concluded that 10% of the predictions were correct.

All of the IRGs and their partners under the chosen distance threshold (Table 1) were assumed to be part of a network. Genes were represented by nodes, which were joined by edges. These edges were obtained by using the KCCA.

The network was described by an adjacency matrix A:

$$A = (a_{u,v}) \tag{2}$$

where u and v are two genes, and $a_{u,v} = 1$ if and only if u and v are joined by a prediction; in other cases, $a_{u,v} = 0$.

From the adjacency matrix, the average node degree was assessed [47]:

$$\overline{deg}(v) = \frac{1}{n_v}\sum_{A} a_{u,v} \tag{3}$$

where n_v is the number of genes in the network.

The global clustering coefficient (CC) was used to determine how many IRGs are clustered together based on IRG triplets:

$$\mathrm{CC} = \frac{3 \cdot n_{\Delta}}{n_{\mathrm{IRG}} \frac{(n_{\mathrm{IRG}}-1)}{2}} \tag{4}$$

where n_{IRG} is the number of IRGs and n_{Δ} is the number of triplets formed only by IRGs.

To determine whether the cassava network had features similar to that of the *Arabidopsis* network, a singular enrichment analysis (SEA, Fisher's exact test, $P < 0.005$) was performed to identify overrepresented GO terms in the sets of genes related to resistance in the genome using AgriGo [19].

Through a bibliographic search using PubMed, the functions of 12 well-known IRGs and their 10 closest partners were established. In addition, a search for co-expression networks in the public *Arabidopsis* databases ATTED-II [48], CoEXpression [27] and GeneCat [49] was carried out. A BLASTX search of the first 10 predicted partners was performed.

Authors' contributions

All authors were involved in gathering and preprocessing of data. AG and DM generated the RNA-seq data for *Arabidopsis*. LL implemented the KCCA method. AQ, ÁB and JFO participated in the analysis of predictions regarding the literature and other known networks. CL analyzed the predictions and proposed the putative function of genes in the immunity pathways. LGL and AP participated in data preparation, data exploratory analyses, the topological analyses of networks and the GO enrichment analyses analysis and contributed equally to all the mentioned steps. LGL, AP, CL and LL drafted and revised the manuscript. All authors have read and approved the final manuscript.

Competing interests

The authors have no competing interests to declare.

Acknowledgements

This work is financially supported by the Dirección de Investigación Sede Bogotá of the Universidad Nacional de Colombia (Grant No. 201010016738).

Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.gpb.2013.09.010.

References

[1] Wang X, Gulbahce N, Yu H. Network-based methods for human disease gene prediction. Brief Funct Genomics 2011;10:280–93.
[2] Qi Y, Klein-Seetharaman J, Bar-Joseph Z. Random forest similarity for protein–protein interaction prediction from multiple sources. Pac Symp Biocomput 2005:531–42.
[3] Werhli AV, Husmeier D. Reconstructing gene regulatory networks with Bayesian networks by combining expression data with multiple sources of prior knowledge. Stat Appl Genet Mol Biol 2007:6. Article 15.
[4] Kleine LL, Monnet V, Pechoux C, Trubuil A. Role of bacterial peptidase F inferred by statistical analysis and further experimental validation. HFSP J 2008;2:29–41.
[5] Vert JP, Tsuda K, Schölkopf B. A primer on Kernel methods. In: Schölkopf B, Tsuda K, Vert J-P, editors. Kernel Methods in Computational Biology. Cambridge: MIT press; 2004. p. 35–70.
[6] Yamanishi Y, Vert JP, Nakaya A, Kanehisa M. Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis. Bioinformatics 2003;19:i323–30.
[7] Bleakley K, Biau G, Vert JP. Supervised reconstruction of biological networks with local models. Bioinformatics 2007;23:i57–65.
[8] Yamanishi Y, Vert JP, Kanehisa M. Protein network inference from multiple genomic data: a supervised approach. Bioinformatics 2004;20:i363–70.
[9] Pop A, Huttenhower C, Iyer-Pascuzzi A, Benfey PN, Troyanskaya OG. Integrated functional networks of process, tissue, and developmental stage specific interactions in *Arabidopsis thaliana*. BMC Syst Biol 2010;4:180.
[10] Huttenhower C, Haley EM, Hibbs MA, Dumeaux V, Barrett DR, Coller HA, et al. Exploring the human genome with functional maps. Genome Res 2009;19:1093–106.
[11] Jones JD, Dangl JL. The plant immune system. Nature 2006;444:323–9.
[12] Chisholm ST, Coaker G, Day B, Staskawicz BJ. Host–microbe interactions: shaping the evolution of the plant immune response. Cell 2006;124:803–14.

[13] Dodds PN, Rathjen JP. Plant immunity: towards an integrated view of plant–pathogen interactions. Nat Rev Genet 2010;11: 539–48.
[14] Pritchard L, Birch P. A systems biology perspective on plant–microbe interactions: biochemical and structural targets of pathogen effectors. Plant Sci 2011;180:584–603.
[15] Mukhtar MS, Carvunis AR, Dreze M, Epple P, Steinbrenner J, Moore J, et al. Independently evolved virulence effectors converge onto hubs in a plant immune system network. Science 2011;333:596–601.
[16] Lee I, Seo YS, Coltrane D, Hwang S, Oh T, Marcotte EM, et al. Genetic dissection of the biotic stress response using a genome-scale gene network for rice. Proc Natl Acad Sci U S A 2011;108: 18548–53.
[17] Lebart L, Piron M, Morineau A. Statistique exploratoire multimensionnnelle. 3rd ed. Paris: Dunod; 1995.
[18] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003;13: 2498–504.
[19] Du Z, Zhou X, Ling Y, Zhang Z, Su Z. AgriGO: a GO analysis toolkit for the agricultural community. Nucleic Acids Res 2010;38:W64–70.
[20] Chinchilla D, Shan L, He P, de Vries S, Kemmerling B. One for all: the receptor-associated kinase BAK1. Trends Plant Sci 2009;14:535–41.
[21] Chinchilla D, Zipfel C, Robatzek S, Kemmerling B, Nürnberger T, Jones JDG, et al. A flagellin-induced complex of the receptor FLS2 and BAK1 initiates plant defence. Nature 2007;448: 497–500.
[22] Li J, Wen J, Lease KA, Doke JT, Tax FE, Walker JC. BAK1, an *Arabidopsis* LRR receptor-like protein kinase, interacts with BRI1 and modulates brassinosteroid signaling. Cell 2002;110:213–22.
[23] Guo Y, Han L, Hymes M, Denver R, Clark SE. CLAVATA2 forms a distinct CLE-binding receptor complex regulating *Arabidopsis* stem cell specification. Plant J 2010;63:889–900.
[24] Lee H, Chah OK, Sheen J. Stem-cell-triggered immunity through CLV3p-FLS2 signalling. Nature 2011;473:376–9.
[25] Miya A, Albert P, Shinya T, Desaki Y, Ichimura K, Shirasu K, et al. CERK1, a LysM receptor kinase, is essential for chitin elicitor signaling in *Arabidopsis*. Proc Natl Acad Sci U S A 2007;104:19613–8.
[26] Willmann R, Lajunen HM, Erbs G, Newman M-A, Kolb D, Tsuda K, et al. *Arabidopsis* lysin-motif proteins LYM1 LYM3 CERK1 mediate bacterial peptidoglycan sensing and immunity to bacterial infection. Proc Natl Acad Sci U S A 2011;108:19824–9.
[27] Atias O, Chor B, Chamovitz DA. Large-scale analysis of *Arabidopsis* transcription reveals a basal co-regulation network. BMC Syst Biol 2009;3:86.
[28] Bent AF, Kunkel BN, Dahlbeck D, Brown KL, Schmidt R, Giraudat J, et al. RPS2 of *Arabidopsis thaliana*: a leucine-rich repeat class of plant disease resistance genes. Science 1994;265: 1856–60.
[29] Collins NC, Thordal-Christensen H, Lipka V, Bau S, Kombrink E, Qiu JL, et al. SNARE-protein-mediated disease resistance at the plant cell wall. Nature 2003;425:973–7.
[30] Nomura K, Debroy S, Lee YH, Pumplin N, Jones J, He SY. A bacterial virulence protein suppresses host innate immunity to cause plant disease. Science 2006;313:220–3.
[31] Huibers RP, de Jong M, Dekter RW, Van den Ackerveken G. Disease-specific expression of host genes during downy mildew infection of *Arabidopsis*. Mol Plant Microbe Interact 2009;22: 1104–15.
[32] López CE, Quesada-Ocampo LM, Bohórquez A, Duque C, Vargas J, Tohme J. Mapping EST-derived SSRs and ESTs involved in resistance to bacterial blight in *Manihot esculenta*. Genome 2007;1088:1078–88.
[33] Contreras E, López C. Identification of polymorphisms in RXam2 a cassava bacterial blight resistance gene candidate. Rev Colomb Biotecnol 2011;13:63–9.
[34] Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, et al. The Pfam protein families database. Nucleic Acids Res 2012;40:D290–301.
[35] Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 2011;39:W29–37.
[36] Eddy SR. Profile hidden Markov models. Bioinformatics 1998;14:755–63.
[37] Emanuelsson O, Brunak S, Von Heijne G, Nielsen H. Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc 2007;2:953–71.
[38] Nakano M, Nobuta K, Vemaraju K, Tej SS, Skogen JW, Meyers BC. Plant MPSS databases: signature-based transcriptional resources for analyses of mRNA and small RNA. Nucleic Acids Res 2006;34:D731–5.
[39] Pérez-Quintero ÁL, Quintero A, Urrego O, Vanegas P, López C. Bioinformatic identification of cassava miRNAs differentially expressed in response to infection by *Xanthomonas Axonopodis* pv. *manihotis*. BMC Plant Biol 2012;12:29.
[40] Megraw M, Hatzigeorgiou AG. MicroRNA promoter analysis. Methods Mol Biol 2010;592:149–61.
[41] López C, Soto M, Restrepo S, Piégu B, Cooke R, Delseny M, et al. Gene expression profile in response to *Xanthomonas Axonopodis* pv. *manihotis* infection in cassava using a cDNA microarray. Plant Mol Biol 2005;57:393–410.
[42] Kwon SI, Koczan JM, Gassmann W. Two *Arabidopsis* srfr (suppressor of rps4-RLD) mutants exhibit avrRps4-specific disease resistance independent of RPS4. Plant J 2004;40:366–75.
[43] Jiang H, Wong WH. SeqMap: mapping massive amount of oligonucleotides to the genome. Bioinformatics 2008;24:2395–6.
[44] Salzman J, Jiang H, Wong WH. Statistical modeling of RNA-Seq data. Stat Sci 2011;26:62–83.
[45] The R Development Core Team. R: a language and environment for statistical computing. Team RDC, editor. R Foundation for Statistical Computing. 2011. p. 409.
[46] Huber W, Von Heydebreck A, Sültmann H, Poustka A, Vingron M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics 2002;18:S96–S104.
[47] Chen P, Deane C, Reinert G. A statistical approach using network structure in the prediction of protein characteristics. Bioinformatics 2007;23:2314–21.
[48] Obayashi T, Kinoshita K, Nakai K, Shibaoka M, Hayashi S, Saeki M, et al. ATTED-II: a database of co-expressed genes and *cis* elements for identifying co-regulated gene groups in *Arabidopsis*. Nucleic Acids Res 2007;35:D863–9.
[49] Mutwil M, Øbro J, Willats WGT, Persson S. GeneCAT – novel webtools that combine BLAST and co-expression analyses. Nucleic Acids Res 2008;36:W320–6.