R Script Approach to Infer *Toxoplasma* Infection Mechanisms From Microarrays and Domain-Domain Protein Interactions

Bioinformatics and Biology Insights
Volume 11: 1–10

DOI: 10.1177/1177932217747256

Ailan F Arenas[1], Gladys E Salcedo[2] and Jorge E Gomez-Marin[1]

[1]Grupo de Estudio en Parasitología Molecular (GEPAMOL), Universidad del Quindío, Armenia, Colombia. [2]Grupo de Investigación y Asesoría en Estadística, Universidad del Quindío, Armenia, Colombia.

ABSTRACT: Pathogen-host protein-protein interaction systems examine the interactions between the protein repertoires of 2 distinct organisms. Some of these pathogen proteins interact with the host protein system and may manipulate it for their own advantages. In this work, we designed an R script by concatenating 2 functions called rowDM and rowCVmed to infer pathogen-host interaction using previously reported microarray data, including host gene enrichment analysis and the crossing of interspecific domain-domain interactions. We applied this script to the *Toxoplasma*-host system to describe pathogen survival mechanisms from human, mouse, and *Toxoplasma* Gene Expression Omnibus series. Our outcomes exhibited similar results with previously reported microarray analyses, but we found other important proteins that could contribute to toxoplasma pathogenesis. We observed that *Toxoplasma* ROP38 is the most differentially expressed protein among toxoplasma strains. Enrichment analysis and KEGG mapping indicated that the human retinal genes most affected by *Toxoplasma* infections are those related to antiapoptotic mechanisms. We suggest that proteins PIK3R1, PRKCA, PRKCG, PRKCB, HRAS, and c-JUN could be the possible substrates for differentially expressed *Toxoplasma* kinase ROP38. Likewise, we propose that *Toxoplasma* causes overexpression of apoptotic suppression human genes.

KEYWORDS: R, host-pathogen interaction, *Toxoplasma*, domain-domain protein interaction, microarrays

RECEIVED: August 1, 2017. **ACCEPTED:** November 18, 2017.

TYPE: Methodology

FUNDING: The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by COLCIENCIAS through grant 111356933664 and Universidad del Quindío.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Ailan F Arenas, Grupo de Estudio en Parasitología Molecular (GEPAMOL), Universidad del Quindío, Carrera 15 Calle 12N, Armenia, 630001 Quindío, Colombia. Email: aylanfarid@yahoo.com

Introduction

Because infectious diseases are a major health problem worldwide, understanding the molecular mechanisms among pathogen-host interactions (PHIs) is key in addressing this concern. The design of an appropriate strategy to combat a specific pathogen depends on our understanding of the specific PHI. Most studies have focused on identifying protein-protein interactions (PPIs) within a single organism (intraspecies PPI prediction). It is difficult to infer new PPIs between 2 different species (interspecies PPI prediction) because the development of an interaction database depends on experimentally verified PHI data that are costly in time, equipment, and budget to produce.[1] Therefore, the design of computational strategies is worthwhile to elucidate infection mechanisms when experimental PHI data are scarce. The interactions between pathogen proteins and their hosts allow the pathogens to manipulate host cellular mechanisms for their own advantage, such as escaping from host immune responses. For instance, the *Toxoplasma gondii* pathogen can manipulate and control a variety of host processes due to secreted factors that interact with the host cell proteins.[2–5] For example, rhoptry proteins are vital for the *Toxoplasma* infection process and its survival. Most of the virulence of *T gondii* strains relies on their polymorphic rhoptry kinases, secreted protein effectors that target host transcription factors and other proteins with antimicrobial functions. The ROP18 and ROP5 cooperatively interact with the murine IFN-γ–induced immunity-related GTPases (IRGs).[6–8] ROP16, another polymorphic kinase, is correlated with virulence because it is involved in constitutive activation of the host STAT transcription factors.[3,4,9] Likewise, expression-level differences in ROP38, another secreted rhoptry kinase, mediate differences in gene activation along the MAP kinase pathway in the host cell.[10,11] After the genome sequence of *T gondii* became available, gene expression profiles at different developmental stages were investigated by microarray expression analyses.[12–15] This technology allows examining genome-wide expression changes in tissues under different conditions. This information was useful in identifying differential expression patterns in human and mouse cell cultures relative to infection by different *Toxoplasma* strains, revealing that either polymorphic or overexpression effector proteins from rhoptry or granule dense organelles are the main elements responsible for modulating host gene expression.[2,10,11] Although significant progress has been made regarding *Toxoplasma* infection mechanisms through microarray analysis, additional research is necessary to learn and decipher the interspecific PPI between toxoplasma and its hosts. Expression profiling indicates whether a particular gene is expressed in a particular condition (by measuring messenger RNA levels), but to determine whether it is involved in a particular cell process, the protein (the product of the gene) must also be examined. Protein

domains are the basic building blocks that determine the structure and function of proteins, and interactions between domains mediate (PPI). Domain-domain interaction (DDI)-based approaches are often used to predict both intraspecies and interspecies PPI. Several different databases store lists regarding experimentally confirmed and predicted DDIs, such as iPfam[16] and DOMINE[17]; this information would be useful if integrated with expression data to infer pathogen-host PPIs. Therefore, we have developed an R script that integrates differential gene expression calculations, enrichment analysis, and the crossing of interspecific DDI to predict interactions between host and pathogen proteins. We applied this script to examine the *Toxoplasma-host* system and identify human proteins that can potentially interact with a specific protein domain of *T gondii*. We focused on this parasite because it is a ubiquitous obligate intracellular protozoan that can invade and replicate in almost all cells of a broad range of warm-blooded animals and is estimated to infect approximately one-third of the world population.[18,19]

Materials and Methods

Data sets

All microarray data used in this work for *Toxoplasma* and host were downloaded from the NCBI (National Center for Biotechnology Information) Gene Expression Omnibus (GEO) database (www.ncbi.nlm.nih.gov/gds, *Toxoplasma* series GSE44189, GSE16115, GSE24905, GSE20145, GSE22315; human series GSE44191, GSE32104, GSE25468, GSE81016; mouse series GSE55298 and GSE27972). The series are interpreted as matrices in which columns are conditions and rows are gene expression values for each condition (all the series are already normalized). All GEO series are included as txt. file in Additional file 1.

To predict DDI, we downloaded the collection of known and predicted DDIs from the database DOMINE v2.0 release 2010.[17] We converted the lists "INTERACTION" into comma-separated value files. This list is included in Additional file 1.

Because each *Toxoplasma* strain exhibits unique characteristics and gene expression signatures in the host cell, an appropriate way to exploit this information would be to identify the genes that are more variable for each strain compared with another across microarrays. Conversely, in microarray data, it is common to observe asymmetric gene expression distributions with extreme values. Generally, microarray expression data exhibit similar means (and medians) but heterogeneous dispersion. This fact suggests that dispersion measurements are appropriate to describe the gene expression profiles. For a set of n observations X, where $X=\{x_1,x_2,\ldots,x_n\}$, the standard deviation (s), the mean deviation (D_m), and the coefficient of variation $(CV=s/\bar{x})$ are 3 useful dispersion measurements when the chosen central tendency measurement is the mean $\bar{x}$. However, because the median is robust in asymmetric distributions with extreme values, 2 more appropriate dispersion measurements are the *Meda* and the coefficient of median variance. For the set of n observations X, their corresponding order statistics are given by $\tilde{X}=\{x_{(1)},x_{(2)},\ldots,x_{(n)}\}$, where $x_{(1)}=Minimum\{x_1,x_2,\ldots,x_n\}$ and $x_{(n)}=Maximum\{x_1,x_2,\ldots,x_n\}$, respectively. The middle value of $\tilde{X}$ is the median of X, denoted by Q_2 is that value which separates the upper 50% of values from the lower 50%. Considering now the set of deviations $D_Q=\{|x_{(1)}-Q_2|,|x_{(2)}-Q_2|,\ldots,|x_{(n)}-Q_2|\}$, the *Meda* is given by the median of D_Q, ie, remember $Meda=Median\{D_Q\}$ that the mean deviation of X is given by $D_m=\sum_{i=1}^{n}|x_i-\bar{x}|/n$. However, analogous to the coefficient of variation, the coefficient of median variation can be defined given by the quotient $CV_{med}=Meda/Q_2$ as a dispersion index based on the median.

The algorithm

We implemented 2 new functions in R named: "rowDM," which estimates the mean deviance for each row in a matrix and "rowCVmed," which estimates the median variation coefficient for each row in a matrix. The output of both functions is a numerical vector ordered from the highest variability to 0; then, a percentile threshold is defined for each vector to choose a set of genes with extreme variability, ie, a desired amount of differentially expressed genes, such that its variability is greater than the percentile chosen.

The procedure used to identify PHI based on microarray and DDI data is explained in the following steps:

Step 1. Functions rowDM and rowCVmed are applied on the GEO data set matrix and the set of genes with extreme variability are select according to appropriate percentile thresholds to obtain submatrices ranking from the most variable rows to the fewer ones (ID probes) in both the pathogen and host microarray experiments.

Step 2. The pathogen (toxoplasma) ID probes obtained from step 1 are mapped to the database "ToxoDB release 26" (www.toxodb.org).

Step 3. The host ID probe sets obtained from both functions (in Step 1) are mapped to gene symbols and Pfam entries using the hgu133plus2.db and mouse4302.db packages.[20,21] We also included a collection of human illumina IDs to map into gene symbol. This list is included in Additional file 1.

Step 4. Subsequently, functional enrichment analysis (FEA) was performed with the different gene sets obtained from Step 3, using the FGNet enrichment package.[22]

Step 5. We included the sqldf package in our R script, as well as the DDI list "INTERACTIONS" to map all domains obtained from the previous host enrichment gene sets (in Step 4), which could interact with the representative domains in *Toxoplasma* proteins. "INTERACTIONS" list

is included in Additional file 1.

Step 6. The result displays a vector list with gene symbols that can interact with a specific domain (for our case, PF00069 kinase domain from toxoplasma).

Step 7. Finally, we remapped the gene symbols for the (KEGG) signaling pathway (www.genome.jp/kegg) to identify all genes that could interact with PF00069 into a specific cell-signaling pathway.

The R script is provided in the Supplementary Information S11.

Results and Discussion

Application of rowDM and rowCVmed functions

Toxoplasma GEO series analysis. On implementing the program, we sought to evaluate its performance in a descriptive manner, comparing it against the GEO2R tool (www.ncbi.nlm.nih.gov/geo/geo2r), which was used by other authors. The GSE44191 and GSE16115 series include RNA expression values for 3 type 1 toxoplasma strains (RH-ERP, RH-JSR, and GT1) from human foreskin fibroblasts (HFFs) infected for 24 hours with each of these strains. To evaluate the performance of the R program, we applied this code script to identify the *Toxoplasma* genes most differentially expressed in both microarrays. Once we applied the rowDM and rowCVmed functions to the GSE44191 series with critical values of 1.0 and 0.142, respectively, we obtained 2 sets of 29 and 30 genes for each function with the most variable *Toxoplasma* genes, according to their RNA expression levels and observed that ROP8 and ROP38 were differentially expressed between the toxoplasma strains RH-ERP versus GT1. Similar results were observed by the original authors of this microarray data under GEO2R[23] in which ROP8 and ROP38 were found to be differentially expressed between RH-ERP and GT1 strains; likewise, 30% of the differentially expressed genes identified from our R script overlapped with the original results.[23] Furthermore, no remarkable differences were observed between the output data from rowDM and rowCVmed functions (Table 1 and Supplementary Information S1A and S1B). For the GSE16115 series, we applied the critical values 0.83 and 0.65 for both functions to obtain 104 and 113 *Toxoplasma* genes with the highest variability in RNA levels, respectively. We found similar results as the original authors, who proposed those hypothetical proteins and 3 members of ABC transporters as the most variable genes in their RNA expression among *Toxoplasma* type 1 strains by GEO2R.[24] By applying our functions, we also found the ROP38 with high variability in its expression for this microarray analysis (Table 1 and Supplementary Information S2A and S2B). Now, we analyzed the GSE24905 series that contain the RNA expression values from 49 recombinant progenies and their parental *Toxoplasma* type I (GT1) and type II (ME49) parasites. Using our R script, we created a submatrix with the parental type I and type II RNA expression data only. By applying critical values of 2.0 and 0.56 for rowDM and rowCVmed, 2 groups of 42 and 49 genes, respectively, were obtained. Interestingly, the ROP5 *Toxoplasma* protein was observed in both groups (Table 1 and Supplementary Information S3A and S3B). Although ROP5 has no kinase activity, it is known as one of the most important virulence factors in *Toxoplasma*, but it is less expressed in type I strains due to lack of some copies of this gene in type I strains. The authors of this series also highlight ROP5 as one of the most important genes differentially expressed between these 2 strains by means of GEO2R.[25] Likewise, we observed more ROP kinases, such as ROP8, ROP1, ROP29, ROP39, ROP21, and ROP16, as important representative groups with the most variable gene expression and possibly related to virulence strain dependence (Table 1 and Supplementary Information S3A and S3B). The ROP16 kinase has been associated with virulence in type I and III strains due to its ability to phosphorylate STAT3/6 host transcription factors.[3,4,9]

Thereafter, we examine the GSE22315 series, which contains RNA values for 6 more representative toxoplasma strains (type I: GT1 and RH, type II: ME49 and Prugniaud, and type III: CTG and VEG), taken 12 hours after infection in HFF cells. For this series, we used the critical values in both functions to obtain a list with the 100 most variable genes in their RNA values (Table 1). Among them, we found that the ROP18 protein considered along with ROP5 as the main virulence factors in the toxoplasma type I genetic background. The ROP18 gene has low expression in the type III strains, considered less virulent, at least in the mice hosts.[30] Similarly, as in the other toxoplasma series, ROP38 along with ROP8, ROP46, ROP20, and ROP19A were also found with high variability in their expression, which was also observed with GEO2R (Table 1 and Supplementary Information S4A and S4B). Finally, we analyzed the GSE20145 series, which compares the RNA values for the 3 canonical *Toxoplasma* strains (type I [RH], type II [Prugniaud], and type III [VEG]) after infecting HFF cells. As with the series examined above, ROP18 and ROP38 appear as the most differentially expressed genes among the 3 *Toxoplasma* strains during the infection process in HFF cells. We also observed other differentially expressed ROPs such as ROP14, ROP15, ROP1, ROP7, ROP40, ROP31, ROP20, ROP6, ROP11, and ROP29 confirmed with GEO2R (Table 1 and Supplementary Information S5A and S5B). In summary, agreeing with other authors, the most representative groups of proteins with the highest variability of RNA expression among canonical *Toxoplasma* strains are the ABC transporters, hypothetical proteins, and rhoptry kinase protein (ROP) family, especially ROP38, which was observed as a differentially expressed gene in most of the *Toxoplasma* strains compared (Table 1 and Supplementary Information).

Table 1. The most relevant differentially expressed gene candidates obtained with both functions from the *Toxoplasma* GEO set series.

GEO SET SERIES	ORGANISM	ROWDM CRITICAL VALUE	ROWCVMED CRITICAL VALUE	NO. OF PROTEIN CANDIDATES FOR BOTH FUNCTIONS, RESPECTIVELY	ENRICHMENT PATHWAY (FEA)	MORE RELEVANT CANDIDATES IN EXPRESSION VARIABILITY OBTAINED WITH BOTH FUNCTIONS	JOURNAL SUPPORT (IN REFERENCES)	SUPPLEMENTARY INFORMATION
GSE44191	*Toxoplasma*	1	0.142	29 and 28	NA	ROP8, ROP38, hypothetical proteins, and ABC transporters	Yang et al[23]	S1A and S1B
GSE16115	*Toxoplasma*	0.83	0.65	104 and 113	NA	ROP38, hypothetical proteins, and ABC transporters	Khan et al[24]	S2A and S2B
GSE24905	*Toxoplasma*	2	0.56	42 and 49	NA	ROP5, ROP8, ROP1, ROP29, ROP39, ROP21, and ROP16	Behnke et al[25]	S3A and S3B
GSE22315	*Toxoplasma*	0.92	0.24	97 and 95	NA	ROP8, ROP18, ROP46, ROP38, ROP20, ROP19A		S4A and S4B
GSE20145	*Toxoplasma*	1.6	0.25	97 and 105	NA	ROP18, ROP14, ROP15, ROP38, ROP1, ROP7, ROP40, ROP31, ROP20, ROP6, ROP11, and ROP29		S5A and S5B
GSE44189	Human	0.51	0.093	80 and 81	Type I interferon	IRF7, ISG15, ISG20, MX1, MX2, OAS1, OASL, and RSAD2	Yang et al[23]	S6A and S6B
GSE25468	Human	1.877	1.115	120 and 120	Immunity related (NF-κB)	ILB1, IRF1, and NFKB1	Rosowski et al[26]	S7A and S7B
GSE32104	Human	2.455	1.125	5 and 5	NA	MEOX1, MMP10, SERPINB3, SERPINB4, IL1RN	Behnke et al[27]	
GSE55298	Mouse	1	0.25	50 and 50	Immunity related	MARCKS, HBEGF, SLC7A2, SOCS2, EGR3, c-MYC, SOCS2, SERPINB9, ITGAX, CISH, C3	Franco et al[28]	S8A and S8B
GSE27972	Mouse	1.52	0.308	72 and 77	JAK-STAT	CISH, SOCS1, SOCS2, and SOCS3	Blader and Saeij[29]	S9A, S9B, and S9C
GSE81016	Human	2.718 (2 h)	1.221 (2 h)	250 and 200	Regulation of MAPK cascade	CDK5RAP3, CSK, FOXM1, NDRG2, PRKCA, SORL1, SPRY2		S10B
GSE81016	Human	2.178 (6 h)	1.219 (6 h)	220 and 210	Regulation of macroautophagy	CTTN, MAP1LC3B, RAB33B, ULK1, ZDHHC8		S10B
GSE81016	Human	2.718 (24 h)	1.222 (24 h)	150 and 170	Regulation of apoptotic process	ARNT2, BIRC5, DFFA, F2R, FNIP1, JUN, MTDH, NME2, RNF34, SLC39A10, SOCS2, SQSTM1, THOC6		S10B

Abbreviations: FEA, functional enrichment analysis; GEO, Gene Expression Omnibus; NA, not applicable; NF-κB, nuclear factor κB.

Host GEO series (human and mouse) and FEA. Continuing with the descriptive performance analysis of our script, we examined the host gene response during *Toxoplasma* infection. We chose the GSE44189 series that includes RNA expression values for HFF infected with 3 type 1 toxoplasma strains. The GSE44189 series was also structured as a matrix in which columns were treatments (*Toxoplasma* infections with 3 strains) and rows were RNA expression values for each human gene for each treatment. After applying both functions with critical values rowDM > 0.51 and rowCVmed > 0.093, we obtained 2 HFF gene sets (ID probes) with the greatest variability in the microarray experiment for each function with 80 and 81 genes, respectively. We observed differentially expressed human genes because of *Toxoplasma* type I infections; these are IRF7, ISG15, ISG20, MX1, MX2, OAS1, and OASL. After applying an FEA for the 2 gene sets, we confirmed that these genes are altogether activated by interferon type I, as proposed by Yang et al[23] (Table 1 and Supplementary Information S6A). Differential expression for those genes was also observed by GEO2R (Supplementary Information S6B). We also analyzed the GSE25468 series which comes from HFF cells infected with the type II (Prugniaud) and type III (CEP) canonical *Toxoplasma* strains. We applied both functions with critical values to obtain 2 subsets with 120 genes each. In this array, we found variably expressed human genes such as ILB1, IRF1, and NFKB1 which are important molecules in the host inflammatory response against pathogens. It seems to be modulated by differential expression genes among *Toxoplasma* strains (Table 1 and Supplementary Information S7A). The differential expression for these 3 genes was also confirmed by GEO2R (Supplementary Information S7B). The original authors for the GSE25468 series reported that toxoplasma type II strains interfere in the nuclear factor κB (NF-κB) pathway.[26] Likewise, it was shown that the activity of this transcription factor is modulated during *Toxoplasma* infection.[29] Now, we examine the GSE32104 series that contains the HFF RNA level in 2 infections: one of them is a wild-type RH type I strain and the other is the same strain but knockout for ROP5 gene. We did not find interesting human variably expressed genes with our script for this series. The author for the GSE32104 series only reported the SERPINB3 as the most differentially expressed gene related to the knockout ROP5 condition.[27] This gene was also observed in our set among the first 5 genes with the highest variability in their RNA expression (MEOX1, MMP10, SERPINB3, SERPINB4, and IL1RN) (Table 1). Although ROP5 alleles are significantly related to infection in the host specifically because of interaction with IRGs,[6–8] the ROP5 gene does not seem to modulate host gene expression.

After exploring *Toxoplasma* infection in mouse macrophages, the data are contained in the GSE55298 series show RNA values from RAW264.7 cells infected with *Toxoplasma* RH strain versus uninfected cells. We looked for the first 5 most variable genes in this array and found the MARCKS and HBEGF genes also proposed by the original authors of this microarray as the most differentially expressed genes because of toxoplasma RH infections in mouse cells.[28] By applying our functions, we expanded the search for the first 50 genes with the greatest variability in RNA expression and found the c-Myc transcription factor in this gene set, which was reported as a gene regulated by *Toxoplasma* RH infection, producing differential expression for the following genes: MARCKS, HBEGF, SLC7A2, SOCS2, EGR3, and others (Table 1 and Supplementary Information S8A). The differential expression for these 5 genes was also corroborated via GEO2R (Supplementary Information S8B).[28]

After that we examine the GSE27972 series that compares the RNA levels from mouse bone marrow–derived macrophages (BMdMs) infected with *T gondii* type I RH strain for 6 hours versus the BMdM uninfected. By taking the first 70 most variably expressed genes in this series, we found the cytokine signaling suppressor groups, such as CISH, SOCS1, SOCS2, and SOCS3, which are involved in inhibiting the JAK-STAT signaling pathway (Table 1, Supplementary Material S9A and S9B); the highest differential expression for those 4 genes were also observed via GEO2R (Supplementary Material S9C). Evidence exists that *Toxoplasma* mediated the induction of the suppressor cytokine signaling protein 1 (SOCS1), which contributes to the inhibition of IFN-β immune response, proven to be critical to control parasite replication in the host.[29] Finally, we executed our script in the GSE81016 series that contains RNA values for WERI-Rb-1 human retinal cells infected with toxoplasma for 2, 6, and 24 hours compared with uninfected control. Thus far, no information has been reported about this series. We found variability in the RNA levels after 2 hours of *Toxoplasma* infection for genes related to both regulations of MAPK cascade and kinase activity. In addition, we observed genes involved in macroautophagy regulation after 6 hours of infection; interestingly, autophagy has been demonstrated to be an antitoxoplasmacidal cell process.[31] We also observed that after 24 hours of infection, apoptotic and cell death processes were also altered. Apoptosis has also been shown as a cell immune mechanism to control *Toxoplasma* growth in the host cell[31] (Table 1 and Supplementary Information S10B).

Mapping toxoplasma Gene Ontology terms to Pfam entries

It was observed that ROP kinases in the *Toxoplasma* genome were the most differentially expressed genes among the *Toxoplasma* strains when they infect and grow inside the host cell (Table 1). This means that *Toxoplasma* strains have different molecular mechanisms to survive, which is correlated with the infectiveness of the strain. The outstanding protein domain for *Toxoplasma* was the protein kinase domain "Pkinase" (Pfam entry: PF00069), present in active ROP kinases. The ROP38

seems to be the most interesting gene found with high expression variability among strains in 4 of 5 toxoplasma arrays analyzed (Table 1). It has been demonstrated that ROP38 can activate host genes associated with the MAPK signaling pathway and NF-κB,[10,23] indicating that ROP38 may interact with some host proteins.

DDIs: "Pkinase PF00069 versus gene set domains" and mapping to the KEGG signaling pathways 04630 (JAK-STAT), 04064 (NF-κB), and 04010 (MAPK)

The previous human and mouse gene sets obtained with our functions from the GSE44189, GSE25468, GSE27972, and GSE81016 series were also mapped to Pfam domain entries to identify functional domains that could interact with the *Toxoplasma* Pkinase domain PF00069 found in *Toxoplasma* ROP38.

In addition, because the differential expression of ROP38 influences HFF gene expression associated with the MAPK, JAK-STAT, and NF-κB signaling pathways,[10,23] we remapped the gene sets obtained with both functions from the GSE44189, GSE25468, GSE27972, and GSE81016 series to the KEGG pathway IDs 04630, 04064, and 04010 to identify possible targets involved in some of these signaling pathways.

We observed interesting transcription factors such as the NFKB1 p105 subunit in the GSE25468 series and the NFKB inhibitor zeta (NFKBIZ) in the retinal human GSE81016 series. These 2 proteins contain inhibitory ankyrin repeat domains, which have been shown to interact with kinase activity proteins[32] (Table 2). Likewise, we also found suppression of SOCS2 and SOCS5 cytokine signaling in the GSE81016 series; these proteins are regulators of the JAK/STAT signaling pathway. Both SOCS2 and SOCS5 contain SH2 domains that can be phosphorylated by JAK proteins[33,34] (Table 2). We saw important kinase proteins, such as PIK3R1, PRKCA, PRKCG, PRKCB, and the GTPase HRAS, that have been reported as key molecules in the MAPK signaling, which activate antiapoptotic genes.[35] In our list, we evidenced the presence of inflammasome activator MAP2K3 and the proapoptotic proteins NFKBIZ, MAP3K5, as well the c-JUN transcription factor (Table 2).

After all of these results, our hypothesis is that the lower expression of ROP38 in type I *Toxoplasma* is correlated with the high expression of these survival genes, which could suppress the activation of the c-JUN transcription factor and avoid cell death via the activation of apoptosis suppressors.[36] Likewise, proapoptotic genes such as NFKBIZ, MAP3K5, and c-JUN were downregulated, compared with the uninfected controls in the human retinal cells (Figure 1). Unlike type I, overexpression of ROP38 in *Toxoplasma* type III upregulates proapoptotic factors leading *Toxoplasma* clearance in the host.[10] ROP38 may interact transiently with some members of this group of MAPK kinases or mimic these proteins and, consequently, the alteration of the MAPK signal pathway in retinal human cells. In summary, according to these results, we propose that *T gondii* can control immunity mechanisms, such as the apoptosis keeping overexpression of apoptotic repressors through the secretion of ROP kinases such as ROP38 (Figure 1).

Because there is no gold standard consensus to identify differentially expressed genes in array experiments,[37] most of the methodologies to identify differentially expressed genes in array experiments are based on a variety of statistical tests, such as analysis of variance[38–40] or false discovery rate.[41–43] We tried to evaluate the performance of our R script differently; first, we proposed critical values to obtain a free chosen output of differentially expressed genes; second, we compare the outputs obtained through our functions against those reported by other authors. It was shown that with our functions we could reproduce similar outcomes to those previously reported, even with a smaller gene set than that proposed by the original authors (Table 1). Similarly, all the differentially expressed gene sets obtained with our functions were also evidenced by the GEO2R tool. The aforementioned demonstrated that our R script can be used by other researchers, even those with no knowledge of R programming, and can be applied to other pathogen-host coexpression experiments to predict more PHI for specific domains. Our R script does not seek to compete with other approaches, such as GSEA,[44] or other R packages, such as limma or GSEABase[45,46]; rather, we wrote an R script code to identify plausible pathogen genes involved in pathogenesis, particularly those with high expression variability. Furthermore, using their functional domains, we aim to identify the host domains that are differentially expressed and potential interactors, as well as map the different cellular pathways.

Conclusions

Predicting interactions between host and pathogen proteins is a never-ending problem with important implications for public health. We have presented an R script that integrates differential gene expression calculations, enrichment analyses, and the crossing of interspecific DDI to predict interactions between pathogen and host proteins (PHI). When applied to the *Toxoplasma*-host interaction system, the R script outcome exhibits similar results with previously reported microarray analyses, thus validating our approach. The R script for human retinal cells suggested that apoptosis inhibitors, such as PIK3R1, PRKCA, PRKCG, PRKCB, and HRAS, or the c-JUN transcription factor directly could be the possible substrate for the differentially expressed ROP38 *Toxoplasma* kinase. This approach will help researchers determine the number of interspecific candidates for each interaction and reduce the number of experiments required for confirmation; in addition, mapping with the different pathways and cell processes could help to infer the parasite's survival mechanisms in the host cell.

Table 2. Domain-domain interaction.

GEO SET SERIES	TYPE CELL AND GENE SET	FUNCTION	PROTEINS THAT COULD INTERACT WITH PKINASE DOMAIN PF00069	JAK/STAT (KEGG 04630)	NF-κB (KEGG 04064)	MAPK (KEGG 04010)
GSE44189	HFF	rowCVmed	PABPN1, KRT19, MX1, CDH3, VCAM1, RGS4, GAP43, MX2, CENPE, EIF2S3, SMAD3, ISG15, OASL, SLC18A2, CACNB2 HIST1H3C, MCAM, HGF, MAP3K7, CIT, PCLO, CASP4, ITGBL1, PPFIA4, COL7A1, PSAT1, GDF15, NTRK2, SCLY		MAP3K7	CACNB2, MAP3K7, NTRK2
GSE25468	HFF	rowCVmed	HSPA1A, HSPA1B, HSPA1L, GSTP1, RAD23A, PSMC3, SFRP1, CSNK1E, CFB, TRIP10, UBE2S, FKBP2, TRMT1, RGS4 USP13, DCLK1, HSD11B1, ABLIM3, IL7R, SLC18A2, UBD, GABBR1, EPHA3, SELE, SERPINB7, NR0B1, RAB2A, SART3 NFKB1, SERPINB3, SERPINB4, SLC43A3, CASP1, MSH6, STIP1, SLC16A3, CSNK1D, PCLO, HIST2H2AA3, HIST2H2AA4 CDH6, FUS, ANKRD11, MRPS18A, HNRNPUL2	IL7R	NFKB1, CASP1	HSPA1A, HSPA1B, HSPA1L, NFKB1
GSE27972	BMdM	rowCVmed	DUSP6, SATB1, VDR, HBEGF, SOCS2, VCAN, AHR, PLEKHF1, SERPINA3G, PTPN2, EDIL3, GEM, TRIB3, EGR2, CDC42EP2 ABTB2, MCF2L, RASGRP1, DIXDC1, ITGB8, NR4A3, NR4A2, CDH1, CISH, F10, SOCS1, DUSP2, BATF3, SOCS3	SOCS2, CISH, SOCS1, SOCS3		DUSP6, RASGRP1, DUSP2
ERI-Rb-1 (2h) GSE81016		rowCVmed rowDM	ZNF337, ROBO1, NFKBIZ, MIXL1, DDR2, SPAG7, HIPK2, KLK1, CC2D1A, STK24, SLC22A7, DGKK, CNTNAP5, SLC25A24 KIF19, ZNF483, ULK1, WDSUB1, PRSS33, UTP3, SUPV3L1, IDH2, KLHDC7A, AHCYL1, CNTN4, PRKCA, ARL13B, ACADM IL1RL1, PAX6, NHLH1, HMHA1, SETD1B, MCM5, KLK9, STK17B, SEZ6L2, SDCBP, FERMT3, LTB, HNRNPA1L2, SORD PLEKHA1, SMARCD3, CASP2, ARL17B, ZDBF2, ZNF583 TIAM1, DNAJC10, WWP2, GABBR2, ZNF337, KHSRP, PRKAR1B, SPAG7, HIPK2, OGT, CC2D1A, PTGR1, H2AFV, KPNA3 CCNY, SF1, ZNF786, ERAL1, SHC1, FGFR3, TRIM36, FOSL2, USP7, SRC, ZNF483, ULK1, WDSUB1, ZCCHC7, UTP3, SGK3 SUPV3L1, IDH2, CSK, R3HDM1, SORL1, PIK3R1, ARL5A, PRKCA, PSPH, UBE2Q1, ACADM, ARHGEF9, HSPD1, SOCS5 GTF3C3, NHLH1, HMHA1, RAF1, SETD1B, MICALL1, PI4K2B, MCM5, IRS2, REL, RGL2, ZNF14, RBM7, DHX37, PTPN12 SLC25A42, LGALS8, ELK4, PTGS1, FOXM1, DDX59, TNFSF14, SDCBP, CASP6, UBE2V1, DYRK2, SYT7, HNRNPA1L2 PLEKHA1, MARK2, ARL17B, KAT2B, ZNF583, ZSWIM7	PIK3R1, SOCS5	NFKBIZ FGFR3	PRKCA, NFKBIZ FGFR3, PRKCA, RAF1, ELK4

(Continued)

Table 2. (Continued)

GEO SET SERIES	TYPE CELL AND GENE SET	FUNCTION	PROTEINS THAT COULD INTERACT WITH PKINASE DOMAIN PF00069	JAK/STAT (KEGG 04630)	NF-κB (KEGG 04064)	MAPK (KEGG 04010)
GSE81016	WERI-Rb-1 (6h)	rowCVmed rowDM	LRIG2, DMD, DEF8, NBEAL2, SLC25A43, DCC, TRIM21, TMPRSS3, SLC25A4, LRSAM1, NFATC3, BRCC3, AARSD1, PSKH1 MASP2, PDZD2, ZNF71, PRKCG, MAP2K3, GP9, TDRD3, RXRG, STXBP6, CABP2, CRB1, FBXW2, PICALM, EPHB2, ZDHHC8 STK39, MXI1, ABCA6, SNX7, RAB5A, ZFHX3, TRIM25, SHANK2, PCDHB10, GIPC2, ZNF14, SAR1B, FOXR1, USP21, CLK3 PUM1, SNRNP40, PPM1K LRIG2, OAT, DEF8, NBEAL2, HNRNPA1, FKBP14, SNX5, ME2, MPP6, PDIA6, ADCY6, CHD4, NFATC3, RALY, NCL, BRCC3 ANKRD28, TRAP1, AARSD1, ULK1, PCNA, EED, SASH1, ZNF420, RGS12, PLCB2, WDR53, PPM1A, RAB33B, PDZD2 TRIM41, PLCL2, SLTM, CCNH, ZNF85, BANF1, KLHL20, ZNF526, MAP4K5, DDX49, SUPT4H1, HRAS, FBXW2, PICALM ZDHHC8, STK39, ZNF787, CRK, ZNF721, ZFHX3, TRIM25, CORO2A, NOL3, BMPR2, TSSC1, DNAJB4, PPWD1, ARNT USP21, ZNF436, EIF4G1, SYT7, CLK3, LRFN1, KLHL28	MAP2K3	PRKCG TRIM25, TRAP1	PRKCG, MAP2K3 PPM1A, HRAS, CRK, MAP4K5
GSE81016	WERI-Rb-1 (24h)	rowCVmed rowDM	SHH, ZNF580, FMNL3, SQSTM1, BDH2, ZNF785, THOC6, AFAP1L1, PRPF4B, HOXC11, RASA2, PARD3, ARNT2, DUSP19 CD4, RORC, CLTCL1, UBE2Z, PTPRT, UBQLNL, TMTC4, ARNT, FOXP3, ZNF468, PRRX2, SPAST, SLC16A5, SFN, UBE2E3 PHLPP1, HPCA, CLK2, BMPR2, FOXJ3, EIF2AK1, USP48, CCNG1, CACNB2, CD79A, PRKCB, ZNF69, PACSIN2 GNB2, SLC2A1, SQSTM1, GAB2, ZNF785, THOC6, MPP6, BAZ1B, PPA2, ZNF2, DTX1, ATP1A3, RALY, WBSCR16, PTPN9 AARSD1, BRD8, HIATL1, PFKL, ERMAP, PICALM, ARNT2, AKAP10, MAP3K5, L3MBTL3, BRD1, ZNF680, OSR2, MKLN1 IRF2BP1, SYNE2, SP3, RAPGEFL1, LRFN2, TMTC4, ANKHD1, ZNF43, NUDT2, ZNF468, CDH23, PRSS8, ZNF91, BIRC5 SEMA4F, JUN, LHX2, RXRG, PHLPP1, HPCA, BMPR2, DHX29, LMNB1, SOCS2, SERPINF1, BRSK1, USP48, ZNF565, PTPRO CORO2A, PKIA, TRAF5, NUDT6, SNRNP70, SENP7, FAM13A, ZNF69, PACSIN2	SOCS2	SOCS2	RASA2, CACNB2, PRKCB MAP3K5, JUN

Abbreviations: BMdM, bone marrow–derived macrophage; GEO, Gene Expression Omnibus; HFF, human foreskin fibroblast; NF-κB, nuclear factor κB; TF, transcription factor.
The most differentially expressed gene sets obtained from the functions rowDM and rowCVmed which have domains that can interact with the Pkinase domain PF00069. Those genes were also mapped for 3 signal pathways JAK/STAT, NF-κB, and MAPK. Red: survival factors HRAS, PIK3R1, PRKCA, PRKCG, and PRKCB that activate antiapoptotic genes. Green: MAP2K3 is related to inflammasome activation. Blue: MAP3K5, NFKBIZ, and JUN (TF) are apoptosis activators.

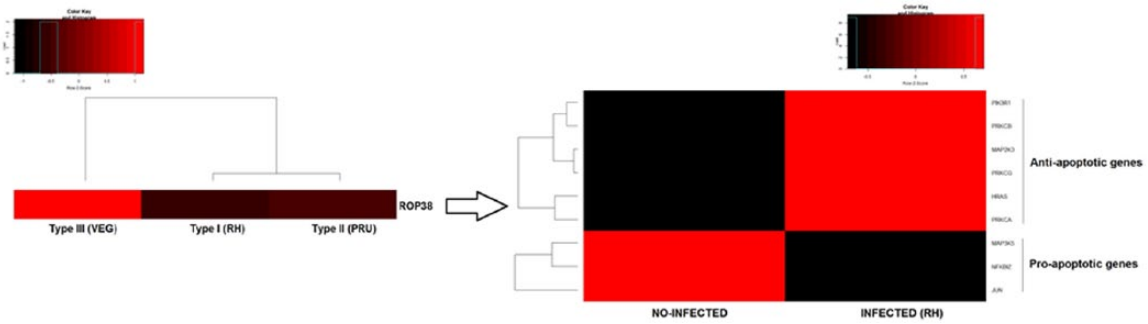

Figure 1. Suggested proteins belonging to the MAPK signaling pathway (KEGG 04010) that could potentially interact with the Pkinase domain (PF00069), as observed in ROP38. On the left is a heat map showing ROP38 toxoplasma; this was the most differentially expressed gene among type I (RH), type II (Prugniaud), and type III (VEG). On the right, up/downregulated genes in human WERI retinal cells (GSE81016 series); these proteins contain domains to interact with PF00069 and are mapped for the MAPK signaling pathway.

Author Contributions

AFA and GES designed the approach and implemented the algorithm. JEG-M contributed to the discussion and revision of the manuscript. All authors read and approved the final manuscript.

Availability of Data and Material

Additional file 1 contains all the additional information.

REFERENCES

1. Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*. 2001;409:533–538.
2. Melo MB, Jensen KD, Saeij JP. *Toxoplasma gondii* effectors are master regulators of the inflammatory response. *Trends Parasitol*. 2011;27:487–495.
3. Saeij JPJ, Boyle JP, Coller S, et al. Polymorphic secreted kinases are key virulence factors in toxoplasmosis. *Science*. 2006;314:1780–1783.
4. Saeij JPJ, Coller S, Boyle JP, Jerome ME, White MW, Boothroyd JC. Toxoplasma co-opts host gene expression by injection of a polymorphic kinase homologue. *Nature*. 2007;445:324–327.
5. Behnke MS, Khan A, Wootton JC, Dubey JP, Tang K, Sibley LD. Virulence differences in *Toxoplasma* mediated by amplification of a family of polymorphic pseudokinases. *Proc Natl Acad Sci U S A*. 2011;108:9631–9636.
6. Fentress SJ, Behnke MS, Dunay IR, et al. Phosphorylation of immunity-related GTPases by a *Toxoplasma gondii* secreted kinase promotes macrophage survival and virulence. *Cell Host Microbe*. 2010;8:484–495.
7. Steinfeldt T, Konen-Waisman S, Tong L, et al. Phosphorylation of mouse immunity-related GTPase (IRG) resistance proteins is an evasion strategy for virulent *Toxoplasma gondii*. *PLoS Biol*. 2010;8:e1000576.
8. Niedelman W, Gold DA, Rosowski EE, et al. The rhoptry proteins ROP18 and ROP5 mediate *Toxoplasma gondii* evasion of the murine, but not the human, interferon-gamma response. *PLoS Pathog*. 2012;8:e1002784.
9. Rosowski EE, Saeij JP. *Toxoplasma gondii* clonal strains all inhibit STAT1 transcriptional activity but polymorphic effectors differentially modulate IFNγ induced gene expression and STAT1 phosphorylation. *PLoS ONE*. 2012;7:e51448.
10. Peixoto L, Chen F, Harb OS, et al. Integrative genomic approaches highlight a family of parasite-specific kinases that regulate host responses. *Cell Host Microbe*. 2010;8:208–218.
11. Melo MB, Nguyen QP, Cordeiro C, et al. Transcriptional analysis of murine macrophages infected with different toxoplasma strains identifies novel regulation of host signaling pathways. *PLoS Pathog*. 2013;9:e1003779.
12. Blader IJ, Manger ID, Boothroyd JC. Microarray analysis reveals previously unknown changes in *Toxoplasma gondii*-infected human cells. *J Biol Chem*. 2001;29:24223–24231.
13. Radke JR, Behnke MS, Mackey AJ, Radke JB, Roos DS, White MW. The transcriptome of *Toxoplasma gondii*. *BMC Biol*. 2005;3:26.
14. Bahl A, Davis PH, Behnke M, et al. A novel multifunctional oligonucleotide microarray for *Toxoplasma gondii*. *BMC Genomics*. 2010;11:603.
15. Behnke MS, Radke JB, Smith AT, Sullivan WJ Jr, White MW. The transcription of bradyzoite genes in *Toxoplasma gondii* is controlled by autonomous promoter elements. *Mol Microbiol*. 2008;68:1502–1518.
16. Finn RD, Miller BL, Clements J, Bateman A. iPfam: a database of protein family and domain interactions found in the Protein Data Bank. *Nucleic Acids Res*. 2013;42:D364–D373.
17. Yellaboina S, Tasneem A, Zaykin DV, Raghavachari B, Jothi R. DOMINE a comprehensive collection of known and predicted domain-domain interactions. *Nucleic Acids Res*. 2011;39:D730–D735.
18. Flegr J, Prandota J, Sovičková M, Israili ZH. Toxoplasmosis—a global threat. Correlation of latent toxoplasmosis with specific disease burden in a set of 88 countries. *PLoS ONE*. 2014;9:e90203.
19. Gómez-Marin JE, de-la-Torre A, Angel-Muller E, et al. Colombian multicentric newborn screening for congenital toxoplasmosis. *PLoS Negl Trop Dis*. 2011;5:e1195.
20. Carlson M. hgu133plus2.db: Affymetrix human genome U133 Plus 2.0 array annotation data (chip hgu133plus2). *R package version 3.2.2*; 2016.
21. Carlson M. mouse4302.db: Affymetrix Mouse Genome 430 2.0 Array annotation data (chip mouse4302). *R package version 3.2.3*; 2016.
22. Aibar S, Fontanillo C, Droste C, De Las Rivas J. Functional gene networks: r/Bioc package to generate and analyze gene networks derived from functional enrichment and clustering. *Bioinformatics*. 2015;31:1686–1688.
23. Yang N, Farrell A, Niedelman W, et al. Genetic basis for phenotypic differences between different *Toxoplasma gondii* type I strains. *BMC Genomics*. 2013;14:467.
24. Khan A, Behnke MS, Dunay IR, White MW, Sibley LD. Phenotypic and gene expression changes among clonal type I strains of *Toxoplasma gondii*. *Eukaryotic Cell*. 2009;8:1828–1836.
25. Behnke MS, Khan A, Wootton JC, Dubey JP, Tang K, Sibley LD. Virulence differences in *Toxoplasma* mediated by amplification of a family of polymorphic pseudokinases. *Proc Natl Acad Sci U S A*. 2011;108:9631–9636.
26. Rosowski EE, Lu D, Julien L, et al. Strain-specific activation of the NFKB pathway by GRA15, a novel *Toxoplasma gondii* dense granule protein. *J Exper Med*. 2011;208:195–212.
27. Behnke MS, Fentress SJ, Mashayekhi M, Li LX, Taylor GA, Sibley LD. The polymorphic pseudokinase ROP5 controls virulence in *Toxoplasma gondii* by regulating the active kinase ROP18. *PLoS Pathog*. 2012;8:e1002992.
28. Franco M, Shastri AJ, Boothroyd JC. Infection by *Toxoplasma gondii* specifically induces host c-Myc and the genes this pivotal transcription factor regulates. *Eukaryotic Cell*. 2014;13:483–493.
29. Blader IJ, Saeij JP. Communication between *Toxoplasma gondii* and its host: impact on parasite growth, development, immune evasion, and virulence. *APMIS*. 2009;117:458–476.
30. Shwab EK, Jiang T, Pena HF, Gennari SM, Dubey JP, Su C. The ROP18 and ROP5 gene allele types are highly predictive of virulence in mice across globally distributed strains of *Toxoplasma gondii*. *Int J Parasitol*. 2016;46:141–146.
31. Krishnamurthy S, Konstantinou EK, Young LH, Gold DA, Saeij JPJ. The human immune response to *Toxoplasma*: autophagy versus cell death. *PLOS Pathog*. 2017;13:e1006176.
32. Kummer L, Parizek P, Rube P, et al. Structural and functional analysis of phosphorylation-specific binders of the kinase ERK from designed ankyrin repeat protein libraries. *Proc Natl Acad Sci U S A*. 2012;109:E2248–E2257.
33. Yasukawa H, Misawa H, Sakamoto H, et al. The JAK-binding protein JAB inhibits Janus tyrosine kinase activity through binding in the activation loop. *EMBO J*. 1999;18:1309–1320.
34. Croker BA, Kiu H, Nicholson SE. SOCS regulation of the JAK/STAT signalling pathway. *Semin Cell Dev Biol*. 2008;19:414–422.
35. Reyland ME. Protein kinase C isoforms: multi-functional regulators of cell life and death. *Front Biosci*. 2009;14:2386–2399.
36. Bossy-Wetzel E, Bakiri L, Yaniv M. Induction of apoptosis by the transcription factor c-Jun. *EMBO J*. 1997;16:1695–1709.

37. Zhang ZH, Jhaveri DJ, Marshall VM, et al. A comparative study of techniques for differential expression Analysis on RNA-Seq data. *PLoS ONE*. 2014;9:e103207.
38. Draghici S, Kulaeva O, Hoff B, Petrov A, Shams S, Tainsky MA. Noise sampling method: an ANOVA approach allowing robust selection of differentially regulated genes measured by DNA microarrays. *Bioinformatics*. 2003;19: 1348–1359.
39. Nadon R, Shoemaker J. Statistical issues with microarrays: processing and analysis. *Trends Genet*. 2002;18:265–271.
40. Tarca AL, Romero R, Draghici S. Analysis of microarray experiments of gene expression profiling. *Am J Obstet Gynecol*. 2006;195:373–388.
41. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Royal Stat Soc B*. 1995;57:289–300.
42. Pawitan Y, Michiels S, Koscielny S, Gusnanto A, Ploner A. False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics*. 2005;21:3017–3024.
43. Reiner A, Yekutieli D, Benjamini Y. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*. 2003;19:368–375.
44. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102:15545–15550.
45. Ritchie ME, Phipson B, Wu D, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nuc Acid Res*. 2015;43:e47.
46. Morgan M, Falcon S, Gentleman R. GSEABase: gene set enrichment data structures and methods. *R package version 1.32.0*.