Discovery of biological paths activating main regulators

Application on *Salmonella* infection and drug-drug interactions

Daniel Poglayen

DOCTORAL THESIS UPF / 2015

THESIS DIRECTOR:

Dr. Baldomero Oliva Miguel

Structural Bioinformatics Lab (SBI)
Research Program on Biomedical Informatics (GRIB)
Department of Experimental and Health Sciences (CEXS)

upf. Universitat Pompeu Fabra Barcelona

Ai miei genitori.

Acknowledgements

It is very difficult for me in these chaotic days to remember all the people to whom I owe gratitude. I apologize if I forget to mention any of you in this text but I will be glad to express my thankfulness in person.

El primer nom que ve a la meva ment és el del meu supervisor. No perquè mentre escric aquestes paraules encara estigui corrent alguns scripts, sinó perquè, gràcies a tu, Baldo, SBI ha estat un lloc molt estimulant on aprendre i créixer. I would also like to thank my historical lab mates: Jaume, Oriol, Javi, Emre, Joan and Manu. From you I learned a lot about science, Spain, Catalonia, Game of Thrones and I owe you all my nerd knowledge. Bernat, Dani, David, plus all the visitors: Thomas, Narcis, Reetesh, Attila, Billur… It has been a pleasure to share part of this journey with all of you.

I am grateful to the European project SHIPREC, and to the Ministerio de Empleo y Seguridad Social for partially finance this work. My gratitude goes also to Manuel Irimia and his lab to welcome me into the family, especially to Jon, Javi, Chris, Demi, Laura, Vicky and Barbara.

I would like to thank also Judith Klein-Seetharaman for the Pittsburgh experience. He will probably never read this words but I really would like to thank Nick for the time I spent in Pitt, meeting you has been a real pleasure. Thanks also to Kelly for the confidence and kindness, thanks to Filippo and his stories and suggestions, you guys made my stay there unforgettable (well, also my yellow banana helped with this task).

Volviendo en Barna quiero agradecer a todas las personas con las que he jugado a beach volley, y no son pocas, ha sido un placer, también cuando hemos perdido o me he lesionado!!

Una menzione la merita il mio club degli italiani del PRBB: inizieró con Jascha, grazie per avermi aiutato tanto soprattutto a muovere i primo passi, per i video golosi e le chiacchierate quasi giornaliere! Mi sarebbe piaciuto

finire insieme quest'avventura ma so che non mi invidi! Isa, intanto grazie per avermi lasciato la tua stanza, e poi anche di tutte le conversazioni e i consigli, professionali e non. Ale, coinquilina, amica, collega e grande bioinformática, sei cresciuta tanto e ancor più crescerai, grazie per i vari scripts, i consigli, i pranzi, le cene, le correzioni...grazie davvero di tutto!! Lucia, il tuo supporto e la tua carica mi hanno aiutato tanto. Luca... Ovviamente anche per te la mia riconoscenza, sei un grande amico e una grande persona, certe cose non si dimenticano (così come la mangiata che ci dobbiamo andare a fare tutti insieme con Carolina e Arianna).

Entre el PRBB y su sucursal, el Marítim, he hecho amistades verdaderas: Fran y Raúl, compañeros de acero (bueno con algunas lesiones xo igualmente aguantándolo todo) y Jordi, recién llegado en mi mundo pero el corazón más grande que haya visto nunca.

Mas personas del PRBB hicieron este viaje muy agradable: Max, Rocio, Elk, Viky, Miquel, gracias por cada momento.

Esta tesis no hubiera visto la luz sin que algunas personas especiales me hubiesen ayudado en varios momentos: Miri tu eres la primera! Has entendido mis momentos de dificultad mucho más que yo los tuyos, sin tantas vueltas me has ayudado y confortado. Siempre te acuerdas y te preocupas por mi, eres muy especial!

Amadis... Para ti no tengo muchas palabras, solamente gracias, gracias y gracias, de corazón!! Eres una persona extraordinaria, un grande profesional y un amigo de verdad! Sin tu ayuda no tengo claro cómo hubiera quedado esta tesis.

No puedo no nombrar en este apartado la que considero mi familia española: la yaya Kati y el yayo Kiko, Mari Carmen y el Rufi! Me habéis acogido y hecho sentir como en mi casa, sois realmente un ejemplo para mi!!

Voglio ringraziare Elena, perché la vita va avanti, le cose cambiano, abbiamo iniziato insieme questa avventura ormai sei anni fa ed oggi la finiamo ancora insieme!! Grazie per esserci sempre quando serve!! I miei amici di sempre, Beppe e Beppe, perché certe relazioni speciali sono uniche nella vita, purtroppo ormai ci vediamo poco ma ovunque ci troviamo sappiamo che possiamo contare sulla nostra amicizia!

Il mio pensiero va anche alle persone che oggi non possono essere qui, in un modo o in un altro è anche grazie loro che sono arrivato a questo punto.

Infine, ovviamente non in ordine d'importanza, desidero ringraziare la mia famiglia, specialmente i miei genitori, perché nessuno è perfetto, ma siete il mio riferimento, a voi va la mia gratitudine per essere sempre presenti, per appoggiarmi, per non avermi fatto mai mancare nulla sotto nessun punto di vista, questa tesi e questo sforzo sono dedicati a voi!!!

Abstract

Cellular behaviour is regulated by a very precise system in which the complex interplay between various types of biomolecules plays a crucial role for the proper functioning of the system itself. Proteins interacting with DNA are in charged of its replication, packaging repair and recombination, among them, transcription factors regulate gene expression, but in turn they are part of a larger network of proteins interactions.

In this thesis I addressed these aspects, in relation to a certain phenotype, in first instance by identifying common regulators of genes with similar expression signatures derived from high-throughput experiments, and then by computationally modelling the signal transduction by means of a message-passing algorithm. This allowed the identification of main regulators in the specific systems describing the infection process of *Salmonella spp.* in two different hosts: *Arabidopsis thaliana* and *Homo sapiens*. The same approach showed encouraging results in the pharmaco-dynamic study of drug-drug interactions.

Resum

El comportament cel·lular està regulat per un complex conjunt de relacions entre diferents tipus de biomol·lècules que juguen un paper fonamental per al correcte funcionament del propi sistema cel·lular. Diferents proteïnes s'encarreguen de la replicació, l'empaquetament, la reparació i la recombinació de l'ADN i en regulen la seva expressió per mitjà de factors de transcripció. Aquests modulen la seva activitat mitjançant interaccions amb altres proteïnes tot formant part d'una xarxa d'interaccions molt més extensa.

En aquesta tesi, m'interesso per aquests aspectes en relació a certs fenotips. Primer, identificant reguladors comuns de gens amb patrons d'expressió similars, obtinguts d'experiments d'alt rendiment; després, fent servir algoritmes de transmissió del missatge per al modelat computacional de la transducció de la senyal. D'aquesta forma he identificat reguladors principals en el procés d'infecció de *Salmonella spp.* en ambdós *Arabidopsis thaliana* i *Homo sapiens*. Una aproximació idèntica aporta resultats esperançadors en l'estudi de la farmacodinàmica de les interaccions entre drogues.

Preface

Taking into account that I wanted to be a medical doctor… I never expected to become a computational biologist when I started my bachelor degree many years ago.

For combinations of life I decided to start statistics, with the idea that translating events into numbers would have been a relatively easy task whose implications would have been extremely interesting. After the first year I still had the same idea but started to sweat having to study and justify very long formulas for, most of the times, easy concepts. After a few years, at the end of that tough path, I had clear in mind two things: I wanted to go back to study biological subjects and wanted to get away, at all costs, from everything that contained formulas. I read of a master degree in Bioinformatics and had no idea what it was. They sold it to me very well: a biology-oriented learning of informatics… They didn't mention any formulas. After one year and a half trying to fill my biological gaps I arrived at the structural bioinformatics lab of Baldo Oliva. New city, first time away from home but, most of all, for the first time I had to deal with real data analysis. Suddenly statistical formulas came back into my life, together with informatics and biology. I did not remember that much about my recent past but this time I was not overwhelmed by what I had to do. That is the point in which, in my opinion, I started learning. Of course it has not been an easy route to get here, but once you start seeing outcomes, your results start travelling, being questioned, validated (or not), then it all makes sense. Meanwhile the learning process never ends, things may turn out not to be as expected but understanding why becomes a new challenge. This work never gets boring. Maybe sometimes frustrating but that's the time to let experts in the field enter into play to help. This thesis collects my first steps into computational biology, my successes and my frustrations. I enjoyed this interdisciplinary field and I hope this little grain of knowledge can be useful and interesting, if not in its entirety, at least part of it. On my side, taking into account that I wanted to be a medical doctor, I am glad to have the opportunity to disclose this thesis.

Table of Contents

Acknowledgements ..... **i**

Abstract ..... **v**

Resum ..... **vi**

Preface ..... **vii**

Table of Contents ..... **ix**

List of Figures ..... **xiii**

List of Tables ..... **xix**

Prologue ..... **xxv**

1 Introduction **1**

1.1 Flow of genetic information **3**

1.1.1 Central dogma: from DNA to proteins 3

1.1.2 Transcriptional regulation of gene expression 5

1.1.3 Transcription factors 13

1.2 Microarrays analysis **21**

1.2.1 CATMA 21

1.2.2 Multiplex BeadArray Assays 22

1.2.3 Microarray data processing 22

1.3 Network biology **28**

1.3.1 General principles of network characterization 29

1.3.2 PPI networks 33

1.3.3 Gene regulatory networks 43

1.4 Networks and diseases **47**

1.4.1 Disease-gene prioritization and “guilt-by-association” principle 49

1.4.2 Host-pathogen PPIs 54
1.4.3 The drug-target space 56
1.5 Thesis motivation 60
2 Objectives 61
3 Salmonella infection in arabidopsis 65
3.1 Unravelling signaling pathways involved in *Arabidopsis* response to *Salmonella* infection using gene-expression and predicted cross-species protein-protein interactions. 68
3.1.1 Abstract 69
3.1.2 Author summary 70
3.1.3 Introduction 71
3.1.4 Materials and methods 74
3.1.5 Results 81
3.1.6 Conclusions and discussion 107
3.1.7 Bibliography 110
3.1.8 Supplementary Information 111
4 Salmonella infection in human 143
4.1 Unravelling signalling pathways involved in human response to salmonella infection leads to gene-specific drug targeting 146
4.1.1 Abstract 147
4.1.2 Introduction 148
4.1.3 Materials and methods 150
4.1.4 Results 154
4.1.5 Conclusions and discussions 162
4.1.6 Bibliography 163
5 Pharmaco-dynamic drug-drug interactions 165

5.1 On the use of protein interaction networks and message passing algorithms to study potential mechanisms of drug-drug interactions 168

5.1.1 Abstract 169

5.1.2 Introduction 170

5.1.3 Materials and methods 172

5.1.4 Results 175

5.1.5 Conclusions and discussions 179

5.1.6 Bibliography 181

5.1.7 Supplementary Information 182

6 Discussion 191

6.1.1 Future Perspectives 200

7 Conclusions 205

8 Appendix 209

8.1 Empirical assessment of causal network inference through a community-based effort 213

8.1.1 Abstract 214

8.1.2 Introduction 215

8.1.3 Results 220

8.1.4 Discussion 232

8.1.5 Acknowledgements 238

8.1.6 Author Contributions 239

8.1.7 Online Methods 240

8.1.8 Bibliography 253

8.1.9 Figures 256

8.2 Crowdsourced assessment of genetic contribution to predicting anti-TNF treatment response in rheumatoid arthritis 265

8.2.1 Abstract 269
8.2.2 Introduction 270
8.2.3 Results 271
8.2.4 Discussion 276
8.2.5 Methods 277
8.2.6 Author contribution 284
8.2.7 Bibliography 284
8.2.8 Figures 287

9 Bibliography 289

List of Figures

- **Figure 1-1** DNA and RNA structure. DNA and RNA have a very similar structure. The backbone consists in sugar molecules linked by phosphodiesteric bonds. In the case of DNA the sugar is deoxyribose while RNA has ribose. The double helical structure of DNA is kept together by hydrogen bonds between the pairing of four bases linked to the sugar molecules. The four bases are adenine (A), which can only bind to thymine (T) with a double hydrogen bond, and cytosine (C), which can only bind to guanine (G) with a triple hydrogen bond. In RNA thymine is substituted by uracil (U). 4
- **Figure 1-2** A two channels microarray experiment. The first step in a microarray experiment consists in the purification and isolation of the mRNA from the samples object of the study. Then reverse transcriptase is used to retrieve complementary DNA (cDNA). The cDNA is then coupled with fluorescent dyes to distinguish between the origin samples. Hybridisation with probes on a chip will cause the activation of the fluorescence that is detected by a laser and transformed into an image. This image is then processed and, to each probe, a value of intensity ratio is assigned. This, after normalization, allows the analysis and comparison between the samples. 10
- **Figure 1-3** Transcriptional Regulation. To initiate transcription, several factors are necessary. Activators bind to their target at the upstream activation sequence (UAS), recruiting the mediator complex. The TATA-binding protein (TBP) subunit of TFIID bind to the promoter TATA-box an recruits TFIIA and TFIIB. The mediator complex, together with the Polymerase II, TFIIF, TFIIE and TFIIH, at this point, are ready to start the transcription. Upon initiation RNA POLII is released from the complex and the transcription process takes place. Figure from

https://mutagenetix.utsouthwestern.edu/phenotypic/pfile.cfm/744/tran. 16

- **Figure 1-4** *In silico* representation of TFBS. A) A position frequency matrix (PFM). At each position reports the number of times each nucleotide has been found. B) The PFM is transformed into a logo. This helps for the identification of motifs at a glance. 19
- **Figure 1-5** Graph representation. On the right side an adjacency matrix and, on the left, the corresponding graphical representation. Interactions between node pairs are indicated with a 1 in the adjacency matrix, self loops, like for node 4, with ones in the diagonal, no interactions with zeros. Figure from: http://faculty.ycp.edu/~dbabcock/PastCourses/cs360/lectures/lecture15.html 30
- **Figure 1-6** Random and scale-free networks. On the left a representation of a random network. On the right a scale-free one. In both cases the number of nodes and edges is 32. Hub nodes, in the scale free net are highlighted in grey and they allow to have, on average, shortest paths lengths. Nodes degree distributions next to each network: for the random network it is clearly a Poisson while for the scale-free one is a power law. Probability distributions are from [136] while networks are taken from : https://en.wikipedia.org/wiki/Social_network. 31
- **Figure 1-7** Disease-genes attributes in PPI networks. A) Scale-free PPI network is characterized by the presence of peripheral nodes, hubs (nodes with a high degree), hubs-bottleneck (nodes with a high degree and high betweeness centrality), non-hubs-bottlenecks (node with low degree but high betweenes centrality). B) A protein mutation may result in two different effects in the network: node removal, typical of truncating mutation, or edge rewiring, aka edgetic perturbations, typical of in-frame mutations. C) Nodes with common biological processes identify functional modules in a network. In the same way disease modules can be identified by a

group of nodes involved in the same disease. The resulting groups may have some overlapping. D) The association between a drug and a potential adverse effect can be studied by examining the shortest path from the drug-target to the node associated with the adverse effect. Figure taken from [251]. 48

- **Figure 3-1**: Graphical representation of the set of hypothesis. Salmonella proteins can be known effectors or other proteins from the Salmonella proteome (in red). Some of them can act as one of the Arabidopsis transcription factors (TF). Arabidopsis proteins (in blue) can also transfer a signal if they are located in the plasma membrane. 81
- **Figure 3-2**: Clusters of Arabidopsis' genes producing similar time-expression profiles after Salmonella infection WT (A) and prgH-mutant (B). Significant clusters are shown in coloured background. The number on the top left corner of each coloured cell indicates the cluster/profile ID and in the bottom left corner is shown the number of genes contained in the cluster. 84
- **Figure 3-3**: Gene regulatory network of transcription factors predicted as MRs. Nodes in the graph represent either gene clusters of Arabidopsis (profiles) obtained with STEM in response to Salmonella WT infection (A) and Salmonella prgH- infection (B) or single TF predicted as MRs. Edges between profiles indicate that the MR of a cluster was found within another cluster (the code of the MR is shown in the edge). An arrow indicates the direction of the control: the profile containing the MR towards the profile it regulates. Similar clusters of both infections are shown in blue (specific clusters for WT infection in orange and for prgH- in green) and CCMRs are highlighted in red. 86
- **Figure 4-1**: Box plots pre (left) and post (right) invariant normalization of Salmonella infected human microarray data..... 151
- Figure 4-2: Toy example of MRs specific drugs ranking. A) In the cmap ranking of the genes according to their drug specific

differential expression profiles we search for our predicted MRs. B) We extract the minimum ranking value for each drug, in other words we search which drugs are reported to affect more the expression of each MR. C) We rank drugs in descending order according to the minimum value found earlier..... 154

- **Figure 4-3**: Predicted MRs and clusters regulated. In the figure we highlighted, by separating them, the common MRs between clusters 39 and 10..... 157
- **Figure 5-1**: Topology effects on the Δ_{AB} score. The figure shows the different topological combinations that our Δ_{AB} score reflects. It can assume negative values when the shortest path (SP) connecting the emitter of one drug to the receiver is shorter then the other. Thus the score assigned to the receiver will depend only on the message sent by the closer emitter. Δ_{AB} can assume a value of 0 when i) the emitter of one drug is not connected with the receiver or ii) the SP connecting the emitters from the two drugs and the receiver has exactly the same length. Finally we can observe positives scores when, using as seeds the combination of the two drugs targets, a node that was previously getting a lower score, now, being connected with both emitters, exhibits a higher score. The SP from this node to the receiver must of equal length than the previously found SPs..... 180
- **Figure 8-1**: Challenge schematic. (a) This analysis was performed in two phases. In the Competitive phase, an open competition was performed to formally evaluate and identify the best models in the world to address this research question. 73 teams representing 242 registered participants joined the challenge. Organizers evaluated model performance for test set predictions submitted by 17 teams. The 8 best performing teams were invited to join the collaborative phase. In this phase, a collectively designed experimental design was developed, in which each team independently performed analyses and challenge organizers performed a combined analysis.

(b) Heritability estimates within the Primary Cohort. (c) Two datasets were used in the analysis: The discovery cohort and the CORRONA CERTAIN study.Participants were provided with 2.5 SNP genotypes + 5 covariates from two cohorts and with the response trait for 2031 individuals in the Discovery cohort ('Training Set'). At the completion of the 16 week training period, participants were required to submit a final submission containing predictions of response traits in a completely independent dataset, the CORRONA CERTAIN study ('Validation Test Set'). 287

List of Tables

- **Table 1-1** Experimental methods commonly used to gather information related with protein protein interactions. 36
- **Table 1-2**: Computational methods commonly used to determine information related with protein-protein interactions. Updated from [189]. 40
- **Table 1-3** Available data repositories for genetic variants and disease-gene associations (reproduced from [Capriotti et al., 2012]). 49
- **Table 1-4** Available disease-gene prioritization tools (adapted and updated from [254]). 53
- **Table 3-1** Summary of correlations (C) between profiles/clusters of Arabidopsis genes after response to infection with wild-type (WT) and prgH- forms of Salmonella. The number of common genes (#shared) is calculated with the p-value of significance based on a hypergeometric test (p-value). In addition, out of the common genes found, the common TFs are identified (Common TFs) and clusters with C>0.99 and sufficient common genes are renamed with the merged-code (Merge). 85
- **Table 3-2**: Common MRs of the merged clusters. The significance of common MRs is calculated with an hypergeometric test (p-value). 88
- **Table 3-3**: SDREM results. In A are the results for Salmonella WT infection and in B for prgH- form. The column "target" indicates if the node of the PIN is a TF of Arabidposis (Y) or not (N). We have included the degree of the node in the PIN (TAP network) and the score of SDREM based on the ratio of the number of oriented paths with highest confidence that go through the node (see methods). .. 90
- **Table 3-4**: Salmonella proteins that could act as Arabidopisis TF (AT4G38680) according to the criteria of: i) sequence similarity; ii) common Pfam domains involved in DNA binding; and iii) common interactions. Percentage of identical residues aligned (sequence

identity) and coverage of the aligned region with respect to the TF (Coverage of TF) and the Salmonella protein (Coverage of target) are shown for the criteria of sequence similarity. The name of the PFAM domains in common and the percentage of common interactions over the total of interactions of AT4G38680 are shown in the last two columns, respectively. 94

- **Table 3-5**: GUILD scores and ranking of Salmonella proteins. Scores of GUILD are calculated with the NetCombo approach (Netcombo scores) using the predicted CMRs as seeds. The second column indicates if the *Salmonella* proteins are known effectors (Y) or not (N). We show the results for *Salmonella* protein effectors with a positive score and the best 10 scores of Salmonella proteins. The third column shows the ranking among the total number of nodes in the PIN and the fourth column the ranking over the total number of Salmonella proteins in the PIN. 97
- **Table 3-6**: Predicted MRs among Arabidopsis TFs with best and positive GUILD scores (top 20% of TFs). We calculated the NetCombo GUILD scores using Salmonella effectors as seeds and the TAP network as the underlying PIN. In the second column (MR) we indicate if the TF was predicted for WT or prgH- infection, as CMR or CCMR. The third column shows the score and the next two columns show the ranking with respect to the total number of nodes in the network and the relative ranking with respect to the 20% of the total number of TFs in the PIN. 100
- **Table 3-7**: Integrated results for the prediction of CMRs of Arabidopsis. A Tick indicates that the TF was predicted as MR, a cross indicates that either the TF was not predicted as MR or it did not fulfil the requirement of the column. Main columns are then split in results for Salmonella WT and prgH- mutant form infections. The first column (TF) shows the name and TAIR code of Arabidopsis TFs. The second column (clustered) shows if the TF belongs to some of the clusters obtained with STEM. The third column (SP Plasma

membrane) indicates if a plasma membrane is found within a shortest path smaller than 4 steps to the TF. The fourth and fifth columns indicate if a Salmonella protein is found at a shortest path smaller than 4 steps when the Salmonella protein is a known effector (SP effector) or not (SP non-effector). The sixth column indicates if there is one or more similar Salmonella proteins under the less restrictive criteria (see Table S8 in Supplementary Information). The next two columns show if the TF was predicted as MR by SDREM or belonged to the top 20% best scored TFs with GUILD when using Salmonella effectors as seeds (GUILD). The last column (30') shows if there is a differential expression of the TF after 30' of infection by Salmonella WT..... 102

- **Table 4-1**: Genes clustered according to the similarity of their expression profiles..... 155
- **Table 4-2**: Putative MRs retrieved for the gene expression clusters retrieved with STEM. 156
- **Table 4-3**: Salmonella effectors that could act as Human TFs according to the criteria of sequence similarity. Percentage of identical residues aligned (sequence identity) and coverage of the aligned region with respect to the TF (Coverage of TF) and the Salmonella protein (Coverage of target) are shown for the criteria of sequence similarity..... 158
- **Table 4-4**: Salmonella proteins that could act as Human MRs of the STEM clusters according to the criteria of i) sequence similarity; ii) common Pfam domains. Percentage of identical residues aligned (sequence identity), coverage of the aligned region with respect to the TF (Coverage of TF) and the Salmonella protein (Coverage of target) are shown for the criteria of sequence similarity. In the third column we mention, eventually, the names of the Pfam domains in common. 160
- **Table 5-1**: GUILD scores of putative MRs in the case of additive DDI. In the table are shown the scores obtained by the predicted putative

MRs, derived from the cmap database (in the first column), using the targets of each drug individually as seeds (second and third columns) and for the drug combination (fourth column). Scores are calculated using the NetScore algorithm in GUILD. Section A) refers to the combination dorzolamide-timolol and section B) to hydrochlorothiazide-metoprolol. 176

- **Table 5-2**: GUILD scores of putative MRs in the case of antagonistic DDI. In the table are shown the scores obtained by the predicted putative MRs, derived from the cmap database (in the first column), using the targets of each drug individually as seeds (second and third columns) and for the drug combination (fourth column). Scores are calculated using the NetScore algorithm in GUILD. Section A) refers to the combination diphenhydramine-theophylline and section B) to aminophylline-theophylline. 177
- **Table 5-3**: GUILD scores of putative MRs in the case of synergistic DDI. In the table are shown the scores obtained by the predicted putative MRs, derived from the cmap database (in the first column), using the targets of each drug individually as seeds (second and third columns) and for the drug combination (fourth column). Scores are calculated using the NetScore algorithm in GUILD. Section A) refers to the combination enalapril-hydrochlorothiazide, section B) to imatinib-vorinostat and section C) to glipizide- metformin.. 179

Prologue

Chapter 1: Introduction. This chapter takes the reader from the Discovery of the DNA up to the specific context in which this thesis fits. It starts with a brief explanation of the flow of genetic information, from the "central dogma" of biology to the mechanisms of transcriptional regulation of gene expression and, in particular, the one carried out by transcription factors. A separate section is dedicated to microarrays because of their central role in the experimental part of all this work, especially for time series data. Some introductory principles of graph theory are the preamble of two sections in which the reader acquires a broader view concerning protein-protein and protein-gene interactions. Once this is achieved the focus is aimed at even more specific types of networks, as is the case of disease-related, host-pathogen and drug-drug interactions. In this introductory part I tried to maintain always a double vision from the experimental and computational point of view.

Chapter 2: Objectives**.** The objectives of this thesis are listed in this chapter with the indication on where they are addressed in this manuscript.

Chapter 3: Salmonella infection in arabidopsis. In this chapter is presented a system wide approach for the study of *Salmonella spp.* mechanisms of infection in *Arabidopsis thaliana*. From the clustering of time series microarray data a set of transcription factors regulating the expression of the genes in the same group (MRs) are computationally derived. A cross-species protein interaction network is inferred and used for the analysis of the shortest path between the MRs and i) plasma membrane proteins, ii) known *Salmonella* effectors. The predicted regulators are employed, then, to identify which pathways trigger the plant response under the bacterial infection. For this they are used i) as seeds of a message-passing algorithm through the host-pathogen interaction network, ii) as potential targets of the signalling pathway originated by *Salmonella* effectors. Finally the prediction on the key role played by a small set of host's proteins during the bacterial infection is experimentally validated.

Chapter 4: Salmonella infection in human**.** In this section is presented an analysis of salmonella-infected human data. With the same approach adopted in the previous chapter: from the clustering of time series microarray data a set of MRs is computationally derived. A cross-species protein interaction network is inferred and used for the analysis of those shortest path, between known salmonella effectors and the predicted MRs, that contain any plasma membrane protein. The hypothesis of a therapy targeting the predicted regulators based on a drug-specific genetic signature is then investigated and its results are used corroborate the MRs predictions.

Chapter 5: Drug- drug interactions. In this chapter is presented a new approach for the study of pharmaco-dynamic drug-drug interactions. Form public databases direct and indirect drug-targets are derived. In this context indirect are considered the most affected genes by the consumption of the chemical compound. This information is used then to predict which TFs are affected by the drug (MRs). The combination of experimentally validated TF-gene and protein-protein interactions into a single human "signalling network" allows the description of the mechanisms of signal transduction leading from direct to indirect drug targets. Based on the hypothesis that interacting drugs should act on the same paths we modelled computationally the signal transduction by means of a message-passing algorithm. The targets of both drugs are used as signal emitters and their gene profiles (through the proposed MRs) as receivers. Then we compared and analysed the scores retrieved by the common transcription factors and genes differentially expressed by both drugs for a few selected examples of interacting drugs.

Chapter 6: Discussion**.** This section includes a final summary of the work presented in this thesis packed with some observations about it and about possible future directions that can continue the path initiated in this book.

Chapter 7: Conclusions

Finally, in the Appendix are the resulting manuscripts from the participation, during the thesis period, to the HPN-DREAM breast cancer network inference and to the HPN-DREAM rheumatoid arthritis responder challenges that do not directly belong to its content.

1 INTRODUCTION

1.1 Flow of genetic information

1.1.1 Central dogma: from DNA to proteins

A milestone in modern genetics is the work published by Watson and Crick in 1953 [1]. For the first time they described the structure of the Dexyrobose Nucleic Acid (DNA). They describe it as "two helical chains each coiled around the same axis". Such double helical structure is kept together by the pairing of four nitrogen-containing bases, namely Adenine (A), Thymine (T), Cytosine (C) and Guanine (G). A will only pair with T in a double hydrogen bond, while C will bind exclusively with G with a triple hydrogen bond, making this last bind stronger then the previous one. The backbone of DNA consists in molecules of a monosaccharide sugar (deoxyribose), to which the bases are attached, one for each sugar molecule, and phosphates joined by phosphodiesteric bonds. The description of the structure of DNA briefly lead to the "central dogma of molecular biology" that consists in a first explanation of how, in a biological system, the genetic information contained in the double stranded DNA (dsDNA) is translated into proteins passing through the phase of transcription into single stranded ribonucleic acid (ssRNA[1]). ssRNA is a nucleic acid with a structure very similar to the one of DNA but while DNA contains deoxyribose, RNA contains ribose [2] and the complementary to adenine is not thymine but uracil (U), an unmethylated for of thymine [3]. A cartoon is represented in Figure 1-1.

[1] For the sake of this introduction we have to mention that it exists also a double stranded RNA (dsRNA), which is composed by two complementary strands, just like DNA. dsRNA can be found inside some viruses and it has been demonstrated that, like siRNA, it can trigger RNA interference in eukaryotes, as well as interferon response in vertebrates[412]–[415].

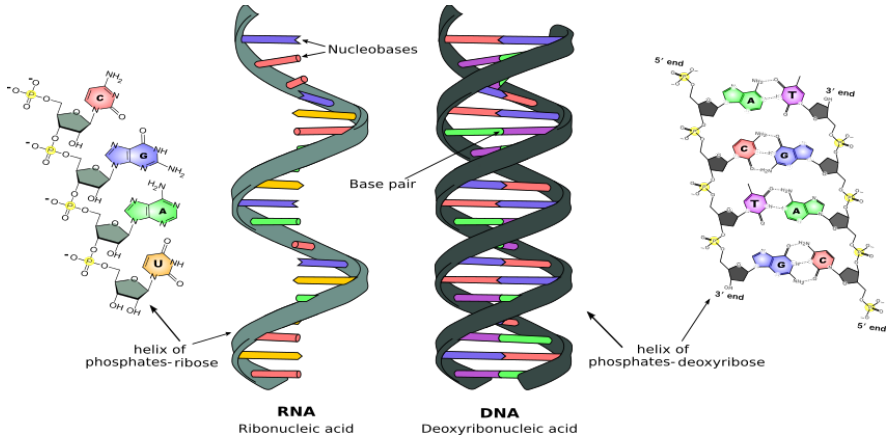


Figure 1-1 DNA and RNA structure. DNA and RNA have a very similar structure. The backbone consists in sugar molecules linked by phosphodiesteric bonds. In the case of DNA the sugar is deoxyribose while RNA has ribose. The double helical structure of DNA is kept together by hydrogen bonds between the pairing of four bases linked to the sugar molecules. The four bases are adenine (A), which can only bind to thymine (T) with a double hydrogen bond, and cytosine (C), which can only bind to guanine (G) with a triple hydrogen bond. In RNA thymine is substituted by uracil (U).

The first hypothesis on the translation of genetic information into protein was born originally in 1941 by Beadle and Tatum [4] and was then stated by Crick in 1958 [5] and revised with a publication in Nature in 1970 [6]; it linked the genotype, the genetic material that each individual inherits, with the phenotype, the specific characteristics of each individual. In this context the locus, or region, of the DNA that encodes a functional RNA or a protein product, is called gene. After many decades of investigation nowadays we known that, according to the controversial data of the ENCODE project, in human, the 93% of the DNA is transcribed into primary RNA products [7], [8]. Transcription is regulated by a category of proteins, called transcription factors (TFs), that are in charge of promote (in the case of activators) or inhibit (in the case of repressors) the recruitment of the RNA polymerase II (Pol II) complex. RNAs can be classified in two categories:

- **coding RNA** In this category we find the messenger RNA (mRNA) which is used for the translation into proteins.
- **non-coding RNA** (ncRNA) Normally this type of RNAs are involved in translation, post-transcriptional regulation and chromatin remodelling. The category includes small interfering RNA (siRNA) and micro-RNAs (miRNA).

The relevance of ncRNAs in controlling transcription and translation and their influence on gene expression have lead to the addition of an extra layer of regulation to the central dogma of molecular biology. In this thesis we will not cover the gene expression regulation performed by ncRNAs, we will thus focus on the action of TFs.

1.1.2 Transcriptional regulation of gene expression

Transcriptional regulation of genes is not an easy task, several components play crucial roles in the characterization of its setting and many regulatory signals define the final transcription production. It is now widely known that each DNA sequence defines a unique landscape for molecules to bind and, in turn, all the molecules that interact and bind to DNA show a unique distribution of binding configurations for each sequence. [9] The first evidence of proteins controlling the gene expression was the lac operon discovery in bacteria by Jacob and Monod [10]. Since then, this mechanism has been widely studied and we now know that basal TFs, recruited by transcription adapter proteins, form the RNA pre-initiation complex (PIC). The PIC binds to DNA regions, called gene promoter regions, located 5' upstream of the transcription start site (TSS) and positions the RNA polymerase II complex. The region in which this last complex is recruited is called core promoter. The remaining part of the promoter region is the

transcription factor binding site (TFBS) and there resides the specificity of the transcription process. From the work of Kawaji et al., [11] we learned that, for the majority of core promoters, there is not a single TSS, thus two positional distributions of TSSs can be found in core promoters: the ones with a single dominant peak (this derives from a single TSS or a group of TSSs located in less then 10bp) and the ones with a more extensive distribution derived form a group of initiation sites closely located. As an example of this last condition, in mammals, some regions of DNA, with length comprised between 200bp and 1Kb, have a frequency of Cs close to a Gs higher then 50%; those regions are called CpG islands (CGIs). Transcription from those regions initiates from multiple weak start sites in regions that are about 100 bp wide [12], [13]. In addition there can be genes with alternative promoters. In this case core promoters are far away in the genome and the use of one promoter region or the other depends on the different cell conditions. Alternative promoters differentially regulated are a common feature in protein-coding genes [14].

1.1.2.1 Experimental methods for testing gene expression

Several experimental methods have been developed in the past decades to quantitatively measure gene expression. They can be divided in two main categories: low and high-throughput methods. In this introduction I will summarize, in chronological order, those techniques but the core of this thesis is based on just one of them: microarrays. Among the low-mid-throughput techniques:

- **Reporter gene.** This technique consists in using a so called "reporter gene" to test whether another gene of interest is expressed or not in a certain condition. This implies to create a construct, called gene fusion, to introduce the reporter gene and the gene of interest in the organism or in the cell culture. The two genes should

have the same promoter elements so that they can be transcribed into a single mRNA molecule. This is then translated into proteins. The two resulting proteins should be able to properly fold into their active conformations and interact with their substrates despite being fused. To obtain this, a DNA segment coding for a flexible polypeptide linker region is included in the artificial DNA construct. In this way the interference of the reporter gene with the one of interest is minimalized. The same technique is also used, in a more complex way and on a larger scale, to test protein-protein interaction in two-hybrid screenings [15].

- **Northern blot.** This technique was developed in 1977 at Stanford University [16]. It is based on the electrophoretic separation of RNA (or isolate mRNA) from different samples. Then a detection step, involving the use of an hybridization probe complementary to all or to a part of the target sequence, is applied. Thanks to this very specific technique even small changes in gene expression can be detected and the false positive results are minimized [17], [18].
- **Western blot.** This technique was developed in 1979 at the Friedrich Miescher Institute [19]. It is used to detect specific proteins in a sample. In all its variants the basics steps are the same. After an electrophoretic separation on gel of the proteins, depending on their structure or length (if denatured), a membrane stained with specific antibodies is in charged of capturing the target proteins.
- **Fluorescent *in situ* hybridization (FISH).** This technique was developed in 1982 [20]. It is used to detect specific DNA sequences on chromosomes using specifically designed fluorescent probes that bind only those parts of the chromosome with a high degree of sequence complementarity. It can be used also to detect and localize specific RNA targets in different types of cells (including tumour

cells) and tissue samples, helping then in the characterization of spatial-temporal patterns of gene expression within cells and tissues.

- **Reverse transcription PCR (RT-PCR).** This technique is used to clone expressed genes by reverse transcribing the RNA of interest into its DNA complement through the use of reverse transcriptase. The newly synthesized cDNA is then amplified using traditional polymerase chain reaction (PCR). Quantitative PCR, also called qPCR, can be added to the RT-PCR for RNA quantification using fluorescent probes. This technique, also called quantitative RT-PCR (qRT-PCR). It is considered to be the most powerful, sensitive, and quantitative assay for the detection of RNA levels. It is frequently used in the expression analysis of single or multiple genes, and expression patterns for identifying infections and diseases [21].

Among the high-throughput techniques we find:

- **Serial analysis of gene expression (SAGE).** We can date back this technique to 1995 [22]. Ten years later the most recent version: SuperSAGE [23]. The final output is a list of short sequence tags and the number of times this is observed. The aim of this technique is to obtain a picture of the mRNA present in a sample and the small tags correspond to fragments of transcripts. The improvements in the technique allowed a better identification of the source gene by obtaining longer tags.
- **Microarrays.** This technique was described for the first time in 1991 [24] and designed in 1995 [25]. It physically consists in a slide (chip) with a collection of DNA spots that are used to measure gene expression levels on a large scale. In a standard microarray, the probes are synthesized and then attached to a chemical matrix on a solid surface by a covalent bond. The solid surface can be glass or a silicon chip, in which case they are

colloquially known as an *Affy chip* when an Affymetrix chip is used. Other microarray platforms, such as Illumina, use microscopic beads, instead of the large solid support. Alternatively, microarrays can be constructed with a direct synthesis of oligonucleotide probes on the solid surface. DNA arrays are different from other types of microarray only in that they either measure DNA or use DNA as part of their detection system. The very first arrays used a two-channel technology for comparing two different conditions, disease *vs.* control, (also known as comparative hybridization). Each sample is distinguished by a fluorochrome of different colour: green (Cy3) and red (Cy5). If genes in the sample labelled with red (or green) are over-expressed, the spot on the microarray will appear red (or green). The spot will appear yellow or orange if the ratio of gene expression between both samples is the same. Single channel microarrays, thus using only one fluorochrome and one chip per sample, were also introduced. Gene expression difference between the samples is computed using the expression ratio of the corresponding hybridized chips. The most used single channel arrays are the Affymetrix "Gene Chip", the Illumina "Bead Chip" and the Agilent single-channel arrays. Many types of arrays, each aimed at a specific biological aspect, have been created. Among others array-comparative genomic hybridization (aCGH) [26] to study copy-number variations (CNVs), single nucleotide polymorphisms (SNP) arrays [27] to detect polymorphisms in a sample population and ChIP-on-chip (ChIP: Chromatin-Immunoprecipitation) to find interactions between proteins and DNA [28]. Microarrays have been successfully used to identify gene signatures, to detect important biomarkers and aid in cataloguing the diverse molecular patterns underlying biological and physiological processes [29]–[31]. Typically, after the "wet" part of a microarray experiment, there is a

pre-processing phase and then a subsequent analysis that may include clustering, differential gene expression analysis and overrepresentation analysis (e.g. gene ontology (GO) enrichment or gene set enrichment (GSE) analysis) and classification. In Figure 1-2 is represented a typical two channel microarray experiment.

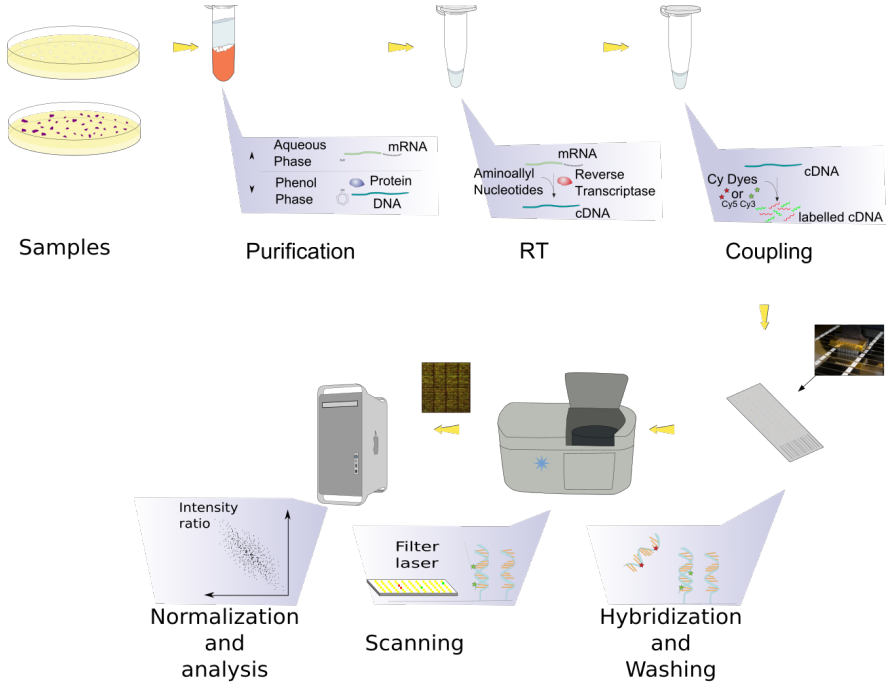


Figure 1-2 A two channels microarray experiment. The first step in a microarray experiment consists in the purification and isolation of the mRNA from the samples object of the study. Then reverse transcriptase is used to retrieve complementary DNA (cDNA). The cDNA is then coupled with fluorescent dyes to distinguish between the origin samples. Hybridisation with probes on a chip will cause the activation of the fluorescence that is detected by a laser and transformed into an image. This image is then processed and, to each probe, a value of intensity ratio is assigned. This, after normalization, allows the analysis and comparison between the samples.

- **Tiling arrays (ChIP-chip and ChIP-seq).** They are very similar to microarrays; they also work by hybridization DNA or RNA to probes fixed to a solid surface. They differ from the previous

technique because instead of using probes of known (or predicted) genes that can be dispersed in the genome, they are meant to work for sequences in contiguous regions, allowing the characterization of regions whose sequence is known but not the specific local functions. The most used are ChIP-chip in which chromatin immunoprecipitation allows the identification of binding sites of proteins. Experiments can be done genome-wide. Formaldehyde is used to cross link proteins and DNA in direct contact and then cell lysate is added and DNA fragmented via sonication. All the resulting fragments are immunoprecipitated and specific antibodies are used to capture only the protein of interest together with the crosslinked DNA fragments. These are then reversed so that the bound DNA can be amplified and characterized using microarrays (ChIP-chip) [32] or high-throughput sequencing (ChIP-seq) [33].

- **RNA sequencing (RNA-seq).** Also known as whole transcriptome shotgun sequencing, is a next generation sequencing technique that allows the study of whole transcriptomes, at a given moment in time, at an incredible depth. In addition to measure the presence and amount of RNA, this technique is currently used also for studies on alternative splicing [34], host-pathogen interactions [35], and fusion genes [36]. Several technologies are currently available from different manufacturers and, thanks to the results derived by the extensive usage made of this method by the ENCODE (encyclopedia of the regulatory elements) [37] and TCGA (the cancer genome atlas) projects, a lot of emphasis is currently put in this next generation sequencing technique.

RNA-seq vs microarrays

The main difference between the two techniques is that, in microarrays, gene expression levels are measured using fluorescence after hybridization, in RNA-seq is the number of fragments that can be mapped to a gene, an exon

or a transcript. Since RNA-seq does not depend on genome annotation, it can be used for the detection of novel expressed regions, alternative isoforms, allele-specific expression or, as mentioned, fusion genes. In addition species lacking a reference genome can be sequenced [38]. Specific arrays have been introduced for splicing but the sequencing technique seems to better detect exon/exon junction [39]. Microarrays have biased signals due to cross-hybridization and limited dynamic range due to the saturation of the fluorescence signals [40]. Nowadays microarrays are still a good choice because of the well-established protocols and their relatively low cost compared to RNA-seq. In addition the analysis of the resulting RNA-seq data needs more infrastructural and computational resources [41] than the ones for microarray results. The tendency for RNA-seq protocols is to be every day more standardized and, in a near future, costs for sequencing, storage and computation will be significantly lowered and probably RNA-seq will supplant microarrays but then the great strides of technology will, probably, make us live new revolutions in the field.

1.1.2.2 Epigenetic regulation of gene expression

As mentioned, the transcription complex is formed by specific TFs, general TFs, co-factors, and RNA polymerase II [42] but the process of transcription is not only dependent on those proteins and their interaction but also on the chromatin structure. In eukaryotes, in order to fit almost 2 meters of genetic material (in the case of human) into a single cell [43], DNA is tightly packed into chromatin. This is a complex of macromolecules composed by proteins, called histones, wrapped by DNA. The single macromolecule is called nucleosome. Arrays of nucleosomes, formed by histones wrapped into a 30 nm fibre, represent the most condensed form of chromatin: heterochromatin. This form of chromatin is too compact and transcription will not take place. However some molecular changes can revert the situation. These are known as epigenetic factors.

Regulatory proteins of the chromatin remodelling complex are attracted or repelled by specific patterns of histone tail modifications [44] . These covalent modifications are caused by chromatin-remodelling factors and enzymes *ad hoc* recruited by gene specific TFs and lead to the binding of other regulatory factors. These, together with the chromatin, create a favourable or non-favourable environment for gene expression [45]. During RNA synthesis, the mentioned chromatin-modifying factors are situated ahead of RNA pol II, this generates a permissive context for transcription.

Methylation (the addition of a methyl group) and acetylation (the addition of an acetyl group) are types of histone modification that are known to control gene expression. Acetylation opens the chromatin and eases the access for TFs to the DNA. Methylation, essentially, leads to a repression of transcription by interfering with the binding sequence of TFs and through the binding of methyl-CpG binding proteins (MBD) [46]. In vertebrates we known that CGIs often contain unmethylated CpG dinucleotides and this transcriptionally active genomic regions contain multiple TSSs [12].

Once the genomic region is uncoiled, epigenetic factors can bind to histones and stretch DNA at the TFBS.

1.1.3 Transcription factors

As already pointed out TFs are DNA binding proteins that bind to short DNA sequences (5-20 bp). They regulate the recruitment of RNA pol II to the promoter region of a gene acting alone or as part of a protein complex. It has been proved that, in a promoter region, we can find from 10 to 50 binding sites corresponding to 5 to 15 different TFs [47]. TFs are so crucial that their mutations is directly associated to many diseases among which we find cancer [48] and one third of the human developmental disorders [49].

Some studies related the number of TFs among different species and their proportion with respect to the number of genes and it has been found that their number grows proportionally to the genome size [50]. For eukaryotes this proportion is around the 5-10% of the total number of genes [51]–[55]. In human the absolute number of TFs should be around 2,000 [56].

1.1.3.1 Regulatory motifs

The DNA region responsible for the regulation of gene expression is called cis-regulatory element. Cis-regulatory elements, in turn, are organized in cis regulatory modules, as in the case of TFBSs. On the other side we have TFs that are trans-regulatory elements that, interacting with cis-regulatory modules (CRMs), carry out their regulatory function. The functionality of the cis-regulatory elements depends on their accessibility and on the relative amount of active TFs. We already described the mechanisms by which DNA regions become accessible or not for transcription, we will now focus our attention on the mechanisms by which, in the cell, the concentration of TFs is controlled. The two main basic mechanisms by which this happens are synthesis and degradation. Alternative splicing [57] and translational regulation [58] give rise to TF isoforms that may also have different regulatory functions. In this context is important to highlight that, from a metabolic point of view, protein synthesis is very expensive and probably it does not have a response quick enough for the regulation of inducible gene response. The fastest mechanisms to regulate the function of TFs are represented by protein phosphorylation, protein-ligand binding and protein-protein interactions (PPIs).

Apart from gene repertory variance, phenotypic differences between organisms may come from differences in the regulation of gene expression [59]. For example the same family of TFs can have different functions in eukaryotes while others are specific to particular lineages. For the majority

of TFs it has been demonstrated that the DNA-binding domain is highly conserved among eukaryotes, while the remaining protein sequence, the one that give rise to PPI and activation domains for example, is often very divergent [60]. Slipped-strand mispairing can quickly change the length of, for example, amino acid (AA) tandem repeats, that are included among these very divergent domains [61]. All this evidences lead the community to strongly accept the idea that changes in regulatory networks will more likely affect the cis part (TFBS) then the trans part (TFs), because of the weaker effects in the first case [62].

1.1.3.2 Recognition of DNA binding sites

A typical example to illustrate how TFs work in the formation of the PIC is the TATA-box promoter. When the sequence TATAAA is present in the promoter region of a gene it is called TATA-box. TATA-box binding proteins (TBP) recognize this motif, bind to the promoter and modify the structure of the DNA in order to recruit more TBP-associated factors (TAFs). On one hand activators increase the binding of TBPs to the TATA-box during transcriptional activation [63] . On the other hand, negative factors are in charged of suppressing the binding activity of the TBP (for example Mot1 or the Taf1 N-terminal domain) [64]. TFs are among the TBPs. The interaction between the TBP and the DNA is initially stabilized by TFIIA that has to compete with NC2, Mot1 and Taf1 for binding to the TBP. In yeast, in addition, the interaction of TFIIA with TAF40, allows the addition of TFIID to the complex [65]. A stable PIC is formed when TFIIB binds to the flanking regions of the TATA-box [66]. A loop from the N-terminal region of TFIIB, named "B-finger", has been found to interact with the DNA but also with the nascent RNA in the catalytic centre of the polymerase [67]. More recently it has been discovered that is the C-terminal of TFIIB that, being located above the polymerase active centre cleft, that guides the "B-finger" towards the catalytic centre. The DNA active centre is

opened thanks to the TFIIB helix/strand interaction with the polymerase rudder. Once this happens a helical region of the TFIIB, the "B-reader" helps in finding the DNA start site while this slides into the cleft, in the catalytic centre of the polymerase. Once the transcript reaches the length of 5 nucleotides in forms a stable complex with the "B-finger" and, once it reaches the length of 7 nucleotides, it collides with the "B-finger" causing the displacement of TFIIB from the promoter [68], [69]. The binding, at this point, of the RNA pol II and TFIIF stabilizes the PIC and allows the recruitment of two general TFs (TFIIH and TFIIE) that, together with the mediator complex, initiate the transcription process. The role of TFIIH has been recently linked with the control of an ATP-dependent transition from the closed to open PIC, a fundamental step for a successful transcription initiation [70]. A cartoon representing the transcription initiation is reported in Figure 1-3.

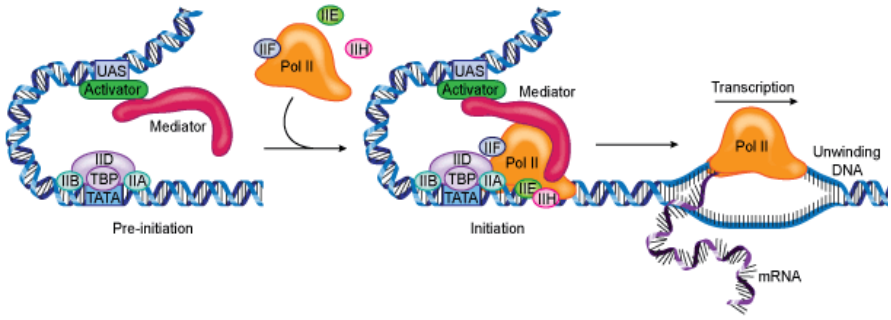


Figure 1-3 Transcriptional Regulation. To initiate transcription, several factors are necessary. Activators bind to their target at the upstream activation sequence (UAS), recruiting the mediator complex. The TATA-binding protein (TBP) subunit of TFIID bind to the promoter TATA-box an recruits TFIIA and TFIIB. The mediator complex, together with the Polymerase II, TFIIF, TFIIE and TFIIH, at this point, are ready to start the transcription. Upon initiation RNA POLII is released from the complex and the transcription process takes place. Figure from https://mutagenetix.utsouthwestern.edu/phenotypic/pfile.cfm/744/tran.

TATA-box promoter is just an example, we known that in yeast only 20% of genes present a TATA-box in their promoter regions and most of them are associated with stress response. Most of eukaryotic promoters

don't have a typical TATA-box promoter. TATA-less genes, represented by most of the housekeeping genes [71], show other elements useful for the recognition of the promoter by the transcription complex, for example downstream promoters and Initiators [72].

1.1.3.3 *In silico* prediction of regulatory motifs

While protein coding sequences have an easy connection with their phenotype, represented by a determined sequence of amino acids, regulatory sequences have a context dependant relationship with their phenotype: a particular profile of transcription [62]. Methods for the *in silico* prediction of regulatory motifs, the TFBS, can be divided in two main categories: pattern matching and pattern discovery algorithms; all of them assume that TFs tend to bind to similar DNA sites, in other words a set of transcriptionally co-regulated genes under specific conditions is likely to be regulated by a set of common TFs.

Pattern matching.

From a set of TFBS a regular motif is derived. The first step is a multiple sequence alignment of the TFBS, each column of the alignment is then represented by a letter of the International Union of Pure and Applied Chemistry (IUPAC) notation in such a way to have an idea of the relative importance of each nucleotide. With such a representation, the information on the relative frequencies of nucleotides at each position is lost. Position matrices have been introduced in order to have the number (with position frequency matrices PFMs) and probability (position weight matrices PWMs) of the normalized frequencies of the four possible nucleotides at each position. Thanks to this approach, given a DNA sequence, it is possible to calculate a quantitative score based on the observed nucleotides at each position. In 1998 Stormo et al. demonstrated that for large and

representative collections of binding sites, the scores are proportional to the binding energies [73], [74] but it does not take into account methylation and acetylation events (for example there are TFs that influence acetylation by recruiting acetylases) [75]. Suffix trees have also been used to predict TFBS but, although criticized [76], PWMs are still the most popular method for this kind of predictions.

Pattern discovery

This strategy consists in finding a common motif in a group of sequences but without aligning them. Two type of detection are possible:

- *De novo* methods. Those strategies compare sequences putatively bound by the same TF with each other (these can be derived for example by a group of genes co-expressed in a microarray experiment or from orthologous promoter sequences) [77]. Hidden Markov Models (HMMs), expectation maximization (EM) and neural networks are extensively used to refine the initial matrix or binding sequence. From the first algorithm of this kind, in 1985, by Galas et al. [78], other discovery methods in this category include the Gibbs-sampling [79], MEME (which uses multiple EM for motif elicitation) [80], AnnSpec [81] and Dispom [82].
- Scanning tools. These methods use collections of PWMs, such as TRANSFAC [83] or JASPAR [84], which will be better described in the next paragraph. They move the PWM along the DNA sequence and, for each position calculate a score. This score is then compared to the one calculated for background sequences, like intergenic regions or CGIs, for determining its significance. CGIs can also be used as a second reference due to the histone acetylation of those regions. Among these tools we mention, in

chronological order, STORM [85], TOUCAN [86], MotifViz [87] and PEAKS [88].

Most of the method we mentioned here, being them based directly on word counting (k-mers), pattern matching or pattern discovery algorithms, predict common motifs using phylogenetic footprinting or over-representation methods, this leads, most of the times, to functionally relevant predictions as the background is taken into account. However we should not forget that low affinity binding sites, non conserved functional binding sites [89] and alternative recognition motifs are very common in mammalian TFBS and, at the same time, very difficult to identify with currently available methods. Multiple transcripts, from the same gene, can be generated from different TSS and this, together with distal and proximal TFBS, depending on their distance from the TSS, increases the complexity of the fundamental task of TFBS identification [90]. The way in which TFBS are represented *in silico* is depicted in Figure 1-4.

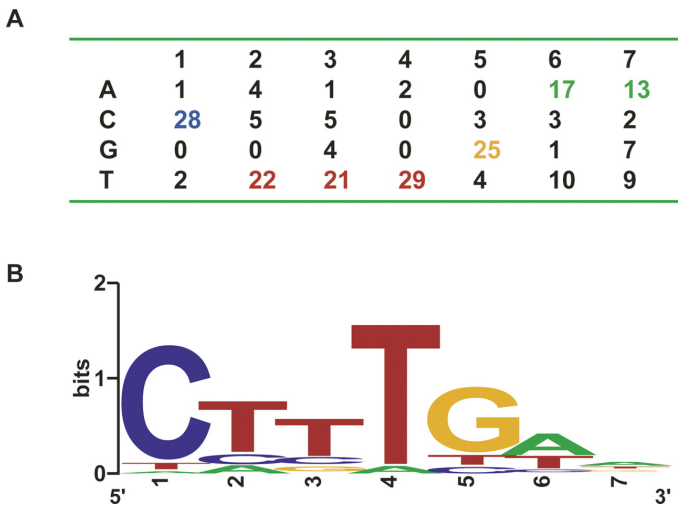


Figure 1-4 *In silico* representation of TFBS. A) A position frequency matrix (PFM). At each position reports the number of times each nucleotide has been found. **B)** The PFM is transformed into a logo. This helps for the identification of motifs at a glance.
Figure from: https://sites.google.com/site/iiserbioinformatics/tutorials.

1.1.3.4 Databases of regulatory elements

As previously mentioned, scanning tools for TFBS discovery use databases of annotated TFBSs, PWMs and sequence logos. These data repositories differ among them for their contents, validation procedures and commercial nature. Among the most famous ones, and widely used, we mention JASPAR [84], [91]. It is a curated database that contains literature-derived, non-redundant motifs in the form of PWMs and sequence logos. It is constantly extended and freely available (http://jaspar.genereg.net). A manually curated commercially available database, with a limited contents free version, is represented by TRANSFAC [83], from the Biobase company (http://www.biobase-international.com). CisRED [92] contains predictions of TFBS derived from neutral evolution, phylogenetic foot printing and homology based analyses (http://www.cisred.org). RegulonDB [93], specific for *Escherichia coli*, provides comprehensive data including regulatory network, binding sites and interactions between TFs (http://regulondb.ccg.unam.mx). Finally, CIS-BP (http://cisbp.ccbr.utoronto.ca) [94] is a library of TF-DNA binding motifs and specificities that also offers on-line tools for scanning DNA sequences and looking for putative TFBS, assigning a DNA sequence where a given TF could bind and, given a DNA motif, assign a TF that may recognize it.

1.2 Microarrays analysis

Many different types of microarrays are available depending on the scientific problem addressed. General principles of microarray experiments are contained in the previous section. A part from single channel and two channels arrays, there is variability in the chip material, cost, manufacturer and way of production. Oligonucleotides can be spotted onto glass using fine pins, using photolithography and pre-made masks or dynamic micro-mirrors, or using electrochemistry on microelectrode arrays. In this chapter the two types of microarrays used for the biological experiments contained in this thesis are explained, together with the theory behind the computational analysis of the resulting data.

1.2.1 CATMA

As mentioned in the prologue, in this thesis we analysed microarray time series data, specifically from the two channels complete Arabidopsis transcriptome microarray (CATMA). It was constructed in the frame of the CATMA European program framework and is based on Genes Specific Tags (GST), designed with "Specific Primers & Amplicons Design Software" (SPADS) [95], which are short and specific sequences for most Arabidopsis genes designed based on "Eugene" annotation software [96]. The AGRIKOLA (Arabidopsis Genomic RNAi Knock-out Line Analysis) European project, focusing on the large-scale systematic RNAi silencing of Arabidopsis genes (http://www.agrikola.org/), is fundamentally based on those GSTs. An exhaustive benchmark study established the CATMA array as a mature alternative to the Affymetrix and Agilent platforms [97].

1.2.2 Multiplex BeadArray Assays

Another type of two channels microarray technology that we used for analysing human time series data is the Multiplex BeadArray Assays developed by Illumina [98]. In this case the chip consists in a multicore optical imaging fiber or in a planar silica slide. This is engraved to obtain micron-sized wells on its surface that allow thousands of 3-micron silica beads, covered by different oligonucleotide capture sequences, to self-assemble in its interior. To distinguish the position of the beads with respect to the wells a decoding process is carried out. Complementary oligonucleotides present in the sample will bind to the beads and the bounding activates a fluorescent label.

1.2.3 Microarray data processing

1.2.3.1 Pre-processing / normalization

The first processing step of the raw data includes image quantification, quality control, background correction, normalization and summarization. This pre-processing is fundamental taken into account that technical noise may heavily influence the results. One of the most used tools in this context is the AffyPLM package [99], [100]. This quality control is performed on the probe level and chips that do not pass this step are filtered out from the analysis. AffyPLM fits models on probe set level to identify the chips of lower quality. Two measures are used in this phase of the analysis:

- Relative log expression (RLE). It is calculated by comparing probe expression on each array against the median expression across all arrays.

- Normalized unscaled standard error (NUSE). It is the standard error estimate for each gene, standardized across all arrays.

Robust multichip average (RMA) [99], [101] is a very useful tool for background correction, normalization, summarization of the raw intensities and scaling of the expression values to a proper scale. Background correction is fundamental to remove the effect of non-specific binding of the fluorophore on the array. Normalization is a crucial step in order to have the same distribution of values on each chip and make those values comparable. One of the most used and efficient is the quantile normalization method [99]. This involves sorting the values per sample in ascending order, substitute the value with the average of each gene across samples and then reorganize with the original order for each sample. Then, to each chip the same mean value is set. Summarization is the fundamental step to pass from probe's values to genes. Each probe only matches to a part of the sequence of a gene and, on the array, probes are grouped into sets thus it is fundamental to combine those sets into a single signal. This step can use one or multiple chips, in the first case median or mean background corrected and normalized probe intensities values are calculated. RMA, based on the idea that the same probe sets respond similarly over different chips, uses a multichip approach. Different probes on the same chip have a higher variability than the same probe on different chips [102]. The multi-chip model includes probe and chip response parameters, to account for the probe effect as well as the relationship of concentration and gene expression [100]. In this thesis we applied the invariant set normalization [103] that allows the arrays to have a similar overall brightness. The procedure is based on the selection of a non-differentially expressed set of genes, called invariant set. This is calculated iteratively until it reaches a stable number of points that are used to compute a piece-wise linear running median curve that will be used for normalization.

1.2.3.2 Batch Effects

Batch effect, generally speaking, is a systematic bias introduced in a biological experiment. Typically batch effects emerge from the laboratory where data are produced and processed: different people may have performed the same experiment, or different parts of it, of different experiments to be compared, the variability of the chips, the specific day, or days in which experiments are done. All these aspects increase the variability in the study thus lowering the confidence on results reflecting real biological signals and even the best-planned and organized study on earth will be affected by technical variation. As Baggerly *et al.* point out: "Batch effects are common in large-scale expression studies, but are not commonly addressed" [104]. Most of the methods for correcting batch effects are addressed to microarray experiments [105]. Batch effects in microarrays have already been reported with the emergence of the first microarray experiments [106].

Among the methods that can be used to address batch effects we find single value decomposition (SVD) [107] and distance weighted discrimination (DWD) [108]. A minimum of 25 samples within each batch is needed by both methods in order to identify which variables explain the batch effect variation. Specifically DWD, with the hyper plane it tries to find in order to separate two batches, only works by pairwise analysis; it then calculates the batch mean and subtracts it in order to obtain corrected values. To overcome the lower limit of 25 samples the approach by Johnson *et* al., (2007) [109], which applies empirical Bayes methods, called ComBat, can be used. It has been proved that this approach is among the ones that better address the problem [110]. If the technical variables influencing the expression are not known, surrogate variable analysis (SVA) [111] can be used to identify these hidden components [112] and then create a linear model that will be used during the analysis to adjust for batch effects.

1.2.3.3 Clustering and the special case of short time series data

Several clustering algorithms have been applied to gene expression data [113]. In general the grouping is preceded by the calculation of a pairwise coefficient, such as Pearson correlation or Euclidean distance, and subsequent application of a clustering method on that measure. In this paragraph we will focus on the clustering of time series expression data. Three of the most popular algorithms in the field are hierarchical clustering [114], k-means [115], and self-organizing maps (SOMs) [116]. All these methods are not specifically designed for time series data thus they ignore data sequentiality and treat the observations, at each time point, as if they were independent of each other. Nevertheless interesting biological results have been retrieved. Other well established algorithms for clustering time series data have been implemented. Schilep *et al.,* (2003) proposed a clustering method based on a mixture of HMM [117]. In an EM style algorithm genes are associated with the HMM most likely to have generated their time courses, then the parameters of the HMMs are estimated based on the genes associated with them. This algorithm requires the number of time points to be much larger than the number of states (or nodes in each Markov chain). Thus, while this algorithm works well for long time series datasets it is not appropriate for short ones. The method proposed by Bar-Joseph *et al.* (2003) [118], is based on a continuous representation of profiles. This algorithm requires the estimation of a few parameters related to the class and, for each gene other five parameters. All this will clearly overfit if the dataset contains only a small number of points ad fail in the resulting clustering. The method by Ramoni *et al.* (2002), is based on gene expression dynamics [119]. It relies on regression and aims to cluster genes whose dynamics can be expressed with almost the same auto-regressive equation. This approach fails in separating clusters of short time series. For example using a regression with the minimum number of parameters to distinguish between "up" and "down" trend, that is two, in a 5 points time series in can

use only the last three and this may lead to over fitting and a deficient cluster separation.

As mentioned, the common problem of all this algorithm, in case the data represent only a few time points, is the over fitting and the difficulty in discerning between real clusters related to a significant biological responses, and random patterns that may occur just by chance. These last are very probable in the case of short series data because of the basal noise and the small number of points studied. Zhao *et al.* (2001) [120] and Lu *et al.* (2004) [121] used a set of predefined profile shapes. This requires the a priori knowledge of the shape of the curve, thus of the gene behaviour in time, that, in most cases is not available. In the method by Möller-Levet et al. (2003) [122] a comprehensive set of profiles is calculated and then genes are clustered by assigning them to the matching profiles. The number of potential profiles grows exponentially with the number of time points making this algorithm efficient only in the case of very few time points. In addition, with such approach, is impossible to differentiate between patterns that arise just by chance and real biological responses. Inequality constraints for the selection of expression profiles have been proposed by Peddada *et al.*, (2003) [123]. Their statistical analysis, based on several repeats, assigns genes to the profile that they best matched. In this case the availability of replicates and the fact that the user has to specify the set of profiles of interest, are crucial aspect for the method to properly work. The approach of De Hoon et al. (2002) [124], fits linear splines with the aim to leverage the statistical power of the different repeats to better estimate gene profiles and their differential expression, when few time points and several repeats are available. In the light of these considerations, for our analysis, where only a few time points and replicas where available, we preferred to use an algorithm specifically designed for clustering short time-series data that outperforms the others. The short time-series expression miner (STEM) [125], can work even if no repeats are available by leveraging the statistical

power obtained from the large number of genes being profiled simultaneously. First a set of model profiles is selected, and then genes are assigned to the profiles that better represent them among the preselected profiles. This step of selection of model profiles independently from the data allows the algorithm to calculate the significance of the different clusters.

1.3 Network biology

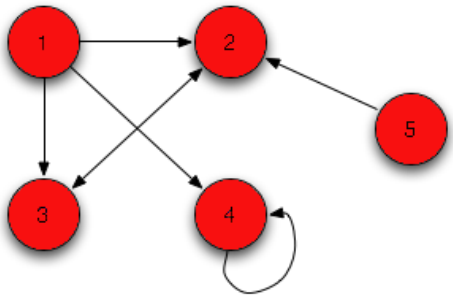
Biomolecules inside the cell (DNA, RNA, proteins and other small molecules) do not act alone. It is clear that, in order to direct cell's activities, it must be established a complex, large and well-.organized system of interactions. The advances in biotechnology and bioinformatics allowed the massive collection of an impressive amount of biological data. Thanks to this, a new perspective emerged in which the characterization of a phenotype is not anymore gene-centred; the new core is represented by the interactions between biomolecules. The dynamics of the biological system are studied analysing the interactions between the components of the system but contemplating it as a whole. Protein sequences, gene expression data, protein-protein interactions (PPI) are just a few examples about the sources of data that can be integrated in order to understand how the different functions of an organism are handled. Thanks to this we are now aware of groups of molecules working together: PPI networks, regulatory networks (represented by gene-protein interactions), genetic interaction networks (represented by gene-gene interactions). These networks are interconnected through common biomolecules. The study of these interconnections, object of a relatively new brunch of science called system biology, aims to identify, understand and model, in a quantitative way, the topological and dynamic properties of biological networks [126]. In his famous book titled "Foundations of System Biology", Hiroaki Kitano defined and made explicit the final aim of system biology: "[…] a new field in biology that aims at system-level understanding of biological systems" [127] but the real origin of system biology can be dated back to 1968 with the system theory of Ludwig von Bertalanffy [128]. To pursue the objective, many massively parallel experimental techniques have been developed and adopted, in different fields, the so called *–omics* (for example genomics and proteomics). The complete set of results of these techniques for each specific context, are

described by terms ending in *–ome* (genome, proteome…). It is clear that the key, for such a holistic understanding of biological systems, is the interplay between the experimental world and modelling methods. Many of the approaches involved have their origin in physics or other natural sciences, for example the formulation and analysis of non-linear models, the theory of complex networks, the characterization of stochastic phenomena, the idea of noise or the statistical methods to identify a model. Thanks to this, the system is converted in a dynamic interaction network and its properties arise directly from the network topology and its dynamic behaviour. Under this new light we can think about diseases as perturbations in the normal biological networks that characterize the cell processes; thus the role of system biology becomes evident to understand and fight against pathologies. Although with modern tools significant progresses have been made in very short time, this branch of science is still in its beginning. Before getting deeper into this, a quick *excursus* on networks is required.

1.3.1 General principles of network characterization

With the term network we refer to a set of elements with connection, or interactions, between them. The formal representation of a network is achieved through the mathematical concept of graph. A graph is an object, consisting in vertices and edges, representing elements and connections respectively. Thus a graph G =(V, E) consist of a set of vertices (aka nodes or points) V and a set of edges (aka arcs or links) E, where each edge is assigned to two (not necessarily disjoint) vertices. The traditional representation of a graph uses a point for each of the vertices and a line for each edge connecting interacting nodes. Two nodes, u and v are neighbours (or adjacent) if they are connected by an edge e, also represented with $\{u, v\}$. This concept of adjacency leads to the representation of a network as an

adjacency matrix. Each row (and each column) represents a node and, when *u* interacts with *v*, the corresponding cell at the intersection of row *u* and column *v*, is filled with a 1. A node that interact with itself, a self-loop, will contain 1 on the diagonal, otherwise, when no interactions or no self-loops are present, the cells are filled with zeros (see Figure 1-5).



	1	2	3	4	5
1	0	1	1	1	0
2	0	0	1	0	0
3	0	1	0	0	0
4	0	0	0	1	0
5	0	1	0	0	0

Figure 1-5 Graph representation. On the right side an adjacency matrix and, on the left, the corresponding graphical representation. Interactions between node pairs are indicated with a 1 in the adjacency matrix, self loops, like for node 4, with ones in the diagonal, no interactions with zeros. Figure from: http://faculty.ycp.edu/~dbabcock/PastCourses/cs360/lectures/lecture15.html.

The first foundations of graph theory can be dated back to 1735 with the famous Königsberg problem enunciated by Euler [129]. The river Pregel divides the two main areas of the city of Königsberg. Seven bridges connect the two islands and the problem consists in finding a walk that crosses every bridge once and only once and that ends exactly where it started. Euler proved that the problem was unsolvable and his explanation was based on the concept of node degree (or connectivity). The degree of a node is the number of edges having the node on one extreme. The conclusion of Euler's work is that "a graph has a path traversing each edge exactly once if exactly two vertices have an odd degree" [129]. If all edges of a walk are distinct then it is called a path. Shortest path between two nodes is the path of minimal length connecting the two vertices. The distance between the two nodes is the length of the shortest path between them.

1.3.1.1 Models and measures

Graphs can be directed or undirected and the direction represent the flow of information in the graph, when this is known. For this reason, in a directed graph, nodes have an in-degree and an out-degree. These describe, respectively, the number of edges pointing at the node and the number of edges that have the node as source. Calculating the degree of all nodes in a network allows the degree distribution calculation [130]. This measure represents the probability distribution of the degrees over the whole network and can be used distinguish between different types of networks. The first mathematical model of random networks is from 1960 [131] and it assumes that each node has the same probability to be connected to another, thus the connections in the network occur by chance. In this case most of the nodes have degrees very close to the average degree of the network, thus the node degree distribution P(k) will have a uniform Poisson distribution [132]. What has been discovered in late 1990's is that most of the real natural networks are far away from being normal. In such webs the majority of nodes seem to have only a few links while only a few of them, called hubs, have very high degrees. Usually, in biological networks, those hubs are essential, for example it has been proved that removing one of these nodes in utero leads to embryonic lethality [133], [134]. This type of networks is called scale-free. In this case it's not possible to use a single node to characterize the network [135]. Scale-free and random networks are compared in **Figure 1-6**.

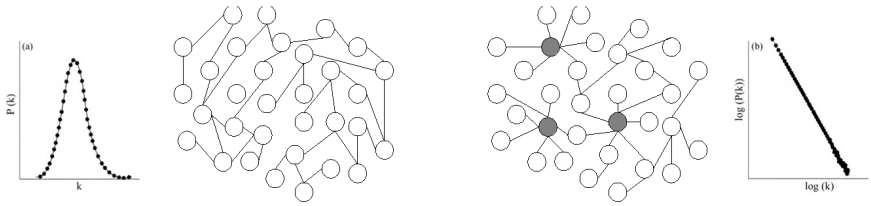


Figure 1-6 Random and scale-free networks. On the left a representation of a random network. On the right a scale-free one. In both cases the number of nodes and edges is 32. Hub nodes, in the scale free net are highlighted in grey and they allow to have, on average, shortest paths lengths. Nodes degree distributions next to each network: for the random network it is clearly a Poisson while for the scale-free

one is a power law. Probability distributions are from [136] while networks are taken from : https://en.wikipedia.org/wiki/Social_network.

The degree distribution P(k) for this type of networks follows, almost, a power-law distribution and, in general, any path between two nodes tend to be short. The small world property states that the distance between two randomly chosen nodes grows proportionally to the logarithm of the number of nodes in the network. This property is present in scale-free networks. To measure how relevant is a specific node in a network, centrality measures are adopted [137]. Three different measures of centrality, all based on shortest path computation [138], are available: degree, closeness, and betweeness. Farness is the sum of the distances from all other nodes and closeness is its reciprocal [139], [140]. Betweenes consists in the number of shortest paths that go through a certain node from any other node [140]. Bottlenecks are nodes with a high betweeness [141] and usually, because they connect hubs, these nodes have a high control on the information flow [142], eventually also on its direction, converting them in potential drug targets [141].

The small world property perfectly fits with the hierarchical connectivity structure of biological networks and accounts for the high modularity and interconnections of genes within the same cluster [126]. Network theory has studied certain graph properties that perfectly suit to biological networks and allow the identification of sub networks involved in diseases [142]. The very basic form of interaction that may be represented in a network is the one that occurs between two nodes. In addition to this there are motifs (patterns), in complex networks, that generate predictable functional consequences [143]. The simplest motif of network architecture is represented by small circuits, composed of three nodes, between TFs and target genes [144], [145]. The loops composing the circuits can be: feed-forward loops (FFL) [143], that act as filters for transient signals, single-input motifs (SIMs), the best example is a TF controlling various genes, and multiple input motifs (MIMs) [143], for example a group of genes sharing a reduced set of regulators [145].

1.3.1.2 Examples of networks in systems biology

Biology, by definition, is the science of life; it describes the processes of our environment, from the molecular to the ecosystem level. At all levels of detail networks appears to be a good representation of the connections between the respective processes. For example networks appear to be perfect models to study the interactions that control cell's biological functions because, in a single representation, it is possible to understand relationships and functions among the different biological entities into play [146]. Different type of networks can represent different biological aspects and are used to answer different biological questions: for example in PPI the nodes are proteins and the edges are physical interactions among them, in transcriptional regulatory networks nodes are genes and proteins and the edges are represented by TFs regulating a gene. Different types of biological networks, at the macroscopic level, include phylogenetic networks (aimed at analysing evolutionary processes through the interrelationships among biological entities) and ecological networks (aka food webs, they describe consumer-resource interactions detailing who is present and who affects whom, directly or indirectly, by feeding interactions, aka trophic interactions, that may also contain quantitative information); at microscopic level we find metabolic networks (in which metabolites and their interconversions by enzymes are represented), cell signalling networks [147], protein-protein interaction [148], [149], pathway cross-talk [150], transcriptional regulatory networks [151], gene-disease networks [152] among others. The core of this thesis is based on gene regulatory and protein-protein interaction networks.

1.3.2 PPI networks

PPIs can be studied from many different points of view and at very different levels of detail. Available methods, both experimental and

computational, reflect the current needs in biology, not without some problems and challenges to be solved in the near future. Depending on their scale, they can be divided in low-throughput and high-throughput. The first category addresses to a very reduced set of proteins while in the second a massive group is studied. Best quality PPI data arises from experimental methods, either in vivo (using a living organism) or *in vitro* (using organisms outside their biological context). These types of experiments are very time consuming and expensive, thus computational methods, have been developed to integrate, and/or to guide, experimental ones. A flux of information is created in which the knowledge acquired with experiments is used to study PPI *in silico*, predictions are made and validated through experiments, the new notion updates the previous knowledge and so on so forth. It appears evident that, in order to obtain relevant and accurate *in silico* results, a high quality of the starting data is fundamental.

1.3.2.1 Experimental study of protein-protein interactions

Thanks to the HUPO proteomics standard initiative [153] in supporting the open biological ontology (OBO) [154], a structured, controlled, vocabulary for PPI experiments annotation is now available. It includes around 168 different experimental methods. Each technique provides information for a specific level of detail.

Protein complementation assays (PCA) represent the largest group of experimental techniques to characterize PPIs [155]. Generally speaking these protocols involve the use of fragments of a "reporter" protein, usually a TF that activates a gene with some known and visible effect on the phenotype, to which the "bait" and "prey" (the two proteins whose interaction is object of the experiment) are covalently linked. In case of interaction between the bait and prey proteins the reporter fragments are close enough to make detectable the activity of the reporter. These techniques work *in vivo* and the

most common ones are yeast two hybrid (Y2H) [156] and tandem affinity purification (TAP) [157]. To detect weak and transient interaction in the living cell, fluorescence based methods are the way to go. Biomolecular fluorescence complementation (BiFC) [158] and green fluorescence protein (GFP) [159] are the most common ones, together with Förster/fluorescence resonance energy transfer (FRET) [160]. In this technique a transfer of energy directed from the donor to the acceptor fluorophore represents the interaction. The bioluminescence resonance energy transfer (BRET) system [161] is even more sensitive because the fluorophores are substituted by luciferases. Arrays can also be used in the study of PPIs. Tangible examples are protein binding microarrays (PBM) [162] in which many probes (proteins) are covalently attached to a surface and labelled proteins are included in the system to test their interaction. In surface plasmon resonance arrays (SPR) [163] the labelled samples are substituted with an optical biosensor that adds information on the kinetic of the interaction in real time by detecting changes in the local refractive index. High resolution methods include X-ray crystallography, the most widely used for determining the structure of large biomolecules. The macromolecule needs to be crystallized and its atoms will cause a diffraction of incident X-rays in different directions. The study of the resulting diffraction map allows understanding the position of the atoms and their chemical bonds. Problems here are represented by the fact that the crystallization may differ significantly from the *in vivo* condition and not all the contacts observed in a crystal may have some biological relevance [164]. Nuclear magnetic resonance (NMR) tries to solve these problems by keeping the macromolecule in a solution. A strong magnetic field is applied, together with radio frequency pulses; the analysis of the resulting chemical shift produced in the nucleus of the macromolecule allows to measure the distance between the atoms and to build a 3D model, at atomic resolution, of the macromolecule. The size of the complex studied is the only limit of this technique [165]. In small-angle scattering the

macromolecule is exposed to X-rays (or neutron beams) and the scattered radiation is detected. At this point the scattering curve of the X-rays (or neutron beam) is used to create a model at a low resolution of the complex (around 15 Å). No crystal is required, allowing the macromolecule to stay in more realistic fluid environment, and in a few days results are available. In cryo-EM tomography the macromolecule is observed with an electron microscope at cryogenic temperatures. This technique results in difficult to interpret density maps in the case of highly dynamic systems. With both the last two methods presented here one can retrieve useful hints about shape and size of the macromolecule and use such information in computational methods. The most common experimental and computational methods are listed in the following table.

Table 1-1 Experimental methods commonly used to gather information related with protein protein interactions.

Method	Class. Specific advantages or drawbacks
Yeast Two Hybrid (Y2H) [156] (Binary interactions)	PCA. Several variants depending on reporter used (such as GAL4-VP16 or LexA-b52). Possible to apply in high-throughput experiments.
Bimolecular Fluorescence Complementation (BiFC) [158] (Binary interactions)	PCA. Based on the reconstitution of a Fluorescent Protein to become functional. Under physiological conditions. Interaction strength based on fluorescence intensity. Spatial resolution. Can detect weak and transient interactions.
Förster/fluorescence resonance energy transfer (FRET) [160]	Proximity-based assay. A fluorophore is transferred from a donor to an acceptor, which are genetically fused to proteins of interest. Two necessary

(Binary interactions)	conditions limit their sensitivity: the donor and acceptor must fall within a specific distance range and be found in favorable orientations.
Bioluminescence Resonance Energy Transfer (BRET) [161] (Binary interactions)	Proximity-based assay. Similar to the FRET experiment, but the donor is replaced by a luciferase. Compared with FRET, BRET do not require an excitation light source, avoiding some of the problems associated with FRET. Useful in photosensitive tissues.
Protein binding microarrays (PBM) [162] (Complex/Binary interactions)	Array technology. Several proteins are printed onto a chip and probed with labeled proteins. Highthroughput experiment.
Surface Plasmon Resonance Array (SPR) (Biacore) [163] (Complex/Binary interactions)	Array technology. Based on an optical biosensor that measures changes in metal array surface refraction index upon protein binding. Not required to label proteins. Provides Kinetic data in real-time.
Tandem Affinity Purification (TAP) [157] (Complex composition)	Based on affinity chromatography. Selective purification due to two purification steps. Possible to apply in high-throughput experiments.
Protein footprinting (Interaction interface)	Binding regions of interacting proteins are protected from the effect of external agents, such as degradative enzymes or oxidative agents. It can be used to detect ligand-induced conformational changes.
Cryo-electron microscopy (Low resolution structural details)	Imaging technique. Based on electron microscopy. Supra-macromolecular structures. Resolution is on average around 10 Angstroms.

X-ray crystallography (High resolution structural details)	Biophysical experiment. Based on the diffraction patterns generated by a single crystal. Only applicable to proteins which can be crystallized.
Nuclear Magnetic Resonance (NMR) (High resolution structural details)	Biophysical experiment. Based on the hydrogen nuclei relaxation after the application of radio frequency pulses of electromagnetic radiation. It allows for the detection of multiple conformations. Molecules are in solution.

1.3.2.2 *In silico* prediction of protein-protein interactions

Although high-throughput methods for predicting PPI are very powerful tools the resulting data may be unreliable and will not cover all possible interactions between proteins. To overcome this, many computational methods have been developed to predict the full range of interactions between proteins with good accuracy. Depending on the level at which the prediction goes, these methods can be divided in two main categories: the ones for predicting binary interactions and the ones for predicting the region or interface involved in an interaction.

Predicting partners of a binary interaction

With the task of identifying pairs of proteins interacting, without specifying which regions are involved in such interaction or its atomic details, these methods can be useful not only for predicting but also for validating experiments. Under this category we find:

- Genome-scale methods (e.g. domain fusion [166], gene neighbourhood [167] and phylogenetic profiles [168])
- Experimental knowledge based methods (e.g. interologs [169], domain profiles [170], and sequence signatures [171])
- Evolution based methods (e.g. correlated mutations [172], and phylogenetic mirror trees [173])

- Chemistry based methods (e.g. prediction of kinetic rates for molecular association [174])
- Docking. Although this is usually used to gain knowledge about the interacting region, it has also been used to predict binary PPI [175])

Predicting interacting region/interface

To properly understand, at a molecular level, the mechanism involved in the docking between proteins, understanding their spatial conformation is essential. To do this, methods to unravel the regions involved in an interaction are essential and they can be divided into two categories depending if they need an *a priori* knowledge on the members of the interacting partners or not. If not information on the interaction participants is known the prediction of interacting regions can be made because it is known that evolution tends to conserve amino-acids on protein surfaces [176]. Interacting interfaces, thus, tend to have certain "known" residues and this has consequences from the chemical point of view [177]–[179]. This leads to structural characteristics by which certain areas will be more favourable than others, energetically, when involved in interactions. This characteristic can be measured, for example, using the Optimal Docking Area (ODA) [180]. Machine learning methods, by combining different sources of information, are also able to predict protein-binding sites [181], [182].

On the other hand if the interacting pairs are known it is known that co-evolution affects the amino-acids involved in an interaction [183], [184]. Of course not all the residues involved in an interaction are just as important: hot spots [185] are the residues that contribute more to the binding free energy and they have more rigid restrictions both structurally [186] end evolutionary [187] with respect to the other amino-acids of the protein.

Topology based methods may also a valid alternative for discovering interacting regions in PPIs [188].

Table 1-2: Computational methods commonly used to determine information related with protein-protein interactions. Updated from [189]**.**

Method	Principle	Output
Phylogenetic profiles	Correlation about the presence and the absence of two proteins in different genomes.	Functionally related proteins
Gene neighborhood	Genomic conservation of topological neighborhood is used to infer functionally related proteins.	Functionally related proteins
Text mining	Automated processing of scientific literature, on the search of proteins co-occurring in the same sentence.	Functionally related proteins, Complex and binary interactions
Interologs mapping	Extension of experimentally detected interactions in an organism to other organisms assuming that homologue proteins maintain their interaction properties.	Complex and binary interactions
Domain fusion	Proteins whose homologues in other organisms happen to be fused into a single protein chain are likely to interact.	Binary interactions
Domain profile pairs	The regions or domains involved in the interactions of a given organism are used to create profiles of interactions. The resulting domain profiles are then used to screen the proteome of another organism and domain-domain interactions are inferred.	Binary interactions Interacting region
Correlated mutations and conservation	Proteins having correlated mutations during evolution are likely to be interacting due to co-adapted evolution of their protein interacting interfaces.	Binary interactions Interacting region
Propensities of the residues	The general composition of the residues located in the interface of PPIs is different from the rest of the protein, and this can be used to infer interacting regions.	Interface region
Computational alanine-scanning mutagenesis	Amino acids are mutated by alanine in the protein-protein interface and its thermodynamic effect on binding free energy is studied in the complex structure.	Interface region
Docking	Looks for best tridimensional structure of the two proteins based	3D Structural model

	on shape or electrostatic complementarity between protein surfaces. It also has been used to infer new PPI binary pairs.	Binary interactions
Comparative modeling	Complex structures can be modeled by means of similarity with other complexes with known structure, assuming the same direction of each partner of the interaction. It also has been used to infer new PPI binary pairs.	3D structural model Binary interactions
Integrated Modeling Platform	Integration of multiple sources of data including a wide range of resolutions to build complex models.	3D structural model
Local structural feature	Uses data from known protein interactions and putative non-interacting proteins (co-localized, non redundant, non-interacting random protein pairs non-similar to PPIs), assigning positive and negative scores to the structural features.	Binary interactions

1.3.2.3 Databases of protein-protein interactions

All the data generated by the listed experimental methods are stored in multiple databases and publications. Among others the biomolecular interactions network database (BIND) [190], the biological general repository for interaction datasets (BioGRID) [191], the database of interacting proteins (DIP) [192], the human protein reference database (HPRD) [193], the MIntAct [194], MIPS [195] and its version for yeast, MPact [196]. Access the information from a single database, most of the times, is an easy task but when it comes to cross information coming from different sources, each with its own platform and nomenclature, the problem is not trivial. Although the biological entities may be the same, their identifiers may differ, the level of information may be different, thus obtaining a general view among all levels of knowledge of PPIs at a glance has become a challenging task. Among the computational tools that tried to solve this problem: ONDEX [197],PIANA [198] and its newer version BIANA [199]. By the way the incompleteness of interaction data (that lead to false negative predictions), the presence of noisy interactions (that lead to false positive predictions) and

the fact that the majority of the mentioned methods are not able to capture time and location dependent aspects of the cellular events, the networks created using available interaction data serve only as a hint of the real, dynamic and context dependent PPI networks.

1.3.2.4 PPI prediction and non-interacting pairs

Although not directly involved in the methods used in this thesis, a fundamental aspect, when developing a prediction method, is to have "gold standard" both for positives and negatives [200]. In the field of PPI prediction it is not a trivial to have the negative set. For example proteins in different location are unlikely to interact [201] but this adds a level of information about cellular localization, which makes the task of predicting PPI easier. Since it has been estimated that for each 1000 protein pairs only 1 of them actually interacts [202]–[204] using random pairs [200], [205]–[207] may be an option but the risk of including false positives interactions when studying certain protein families can be unacceptable or may increase when adding constrains to the randomness (for example using only proteins with similar functional annotation). The ideal would be to have a set of experimentally validated set of known non-interacting protein pairs. PCA methods have the potential to give this kind of information but only recently it has been exploited [208]. It's worth mentioning also the Negatome database [209]. It also contains information on negative PPI. It is partially derived from literature and partially from structural analysis of protein complexes but it is experimental biased and centred in its scale limitation.

1.3.3 Gene regulatory networks

Being the product of gene expression and, at the same time, playing a fundamental role in controlling it, proteins significantly contribute to linking genes to each other and forming what is called gene regulatory network (GRN). By definition a GRN represents highly interconnected processes in a cell that control how genes are expressed in time. The usual representation involves pairs of proteins/genes in which the first element regulates the activity or abundance of the second. Thus, GRN are used to link genes and their products. In this introduction we already mentioned the mechanisms controlling gene expression and the crucial role of TFs in this process. There are TFs in the cytoplasm that, after their activation, translocate to the nucleus and promote the transcription of their regulated genes. In order to accomplish their mission, these TFs have to interact with many other proteins, like membrane receptors, kinases and adaptor proteins. As a clear example of TFs interacting with other proteins we find dimers, the process of dimerization, in fact, increases the specificity and affinity of TFs for DNA and allows them to interact with different proteins, for example in some cases, different combinations of monomers can transform the dimer from one that activates gene transcription to one that represses it.

Finally, In eukaryotes, it is known that transcription factors can act cooperatively forming "enhancesomes", which are assemblies of transcription factors stabilised by protein-protein as well as protein-DNA interactions (for example the enhanceosome that is formed in the human interferon beta gene [210]). Most signalling molecules are products of gene expression and are part of multiple regulatory circuits, thus GRNs involves a huge set of systems that cover many different aspects of the very complex relationship between genes and their products. From these considerations GRNs can be thought of as a kind of qualitative framework, on which

quantitative data can further be superimposed for modelling and making simulations.

1.3.3.1 Databases on gene regulation

In the section Databases of regulatory elements we introduced a list of databases in which information on TFBSs and PWMs (or PFMs) is stored. Regulatory interactions can thus be retrieved from such databases, for example from JASPAR [84], [91] or the commercially available TRANSFAC [211]. Regulatory interactions are contained also in ORegAnno [212] and PAZAR [213]. Using tools like FIMO [214], for searching into a nucleotide or protein sequence database for each of the motifs provided, TReg comparator [215], which compares PWMs or binding sequences against a user provided collection of PWMs, or Tomtom [216], that compares one or more nucleotide motifs against a database of known motifs, it is possible to link TFs to the each of the genes regulated.

1.3.3.2 Gene regulatory network reconstruction

Several approaches use high-throughput techniques, like microarrays, to reconstruct networks of interactions. The very basic form for this procedure is to use clustering. Although this is mainly applied to infer molecular signatures, it can also be used to link elements that, for example, pass a fixed correlation threshold [217]. One of the most famous tools for reverse engineering from microarray data is based on mutual information [218] and its name is ARACNE [219]. The algorithm by Zhand and Horvath, (2005) [220], uses pairwise Pearson correlations to estimate a parameter for obtaining a scale free network. Then average linkage hierarchical clustering and a dissimilarity measure are used with the aim to identify modules. Gaussian graphical models (GGMs) [221], and partial correlations are two approaches that, in contrast to previous ones, are able to distinguish between

directed and undirected correlations. A covariance matrix contains all the pairwise covariance for each pair of nodes in the network and, on the diagonal are the variances. The inverse of the covariance matrix, aka concentration matrix, shows zeros where the nodes in the network are conditionally independent, so disconnected. In a microarray, being the number of genes much higher then the number of sample, the sample covariance matrix cannot be inverted and partial correlations cannot be computed. To solve this some statistical tricks have been thought: a Bayesian approach with sparsity inducing prior [222], graphical lasso [223] and limited-order partial correlations [224]–[226]. The qpgraph package in Bioconductor, contains an *ad* hoc method for microarray data reverse engineering. It estimates the network topology by calculating the non-rejection rate (NRR), which is based on partial correlations [225]. This represents a good estimate of the weight of a direct pairwise interaction between two genes.

Usually, when trying to model the signalling and regulatory interaction using high-throughput data, knockout (KO) experiments are performed in order to have a set of starting (or end) points. The objective of the computational methods becomes to link start and end points and to integrate the paths identified in different KO experiments. In order to do so the physical network models (PNM) technique [227] builds a PPI and protein-DNA interaction network in which it tries to connect with direct paths deleted genes and their targets. Knocked out genes are the starting point also for SPINE [228] that it focused on the positive or negative effect of edges or proteins rather then orienting the PPIs. Another method that tries to explain KO is the one developed by Peleg *et al.*, (2010) [229] in which the output is not a list of paths but a functional network. ResponseNet [230] combines genetic screens with gene expression data to return a condition specific integrated signalling and regulatory network.

1.3.3.3 Integration of PPI networks with gene regulatory networks

After these considerations, the interactions between proteins appear among the most important determinants for the translation of the genotype into the phenotype. This step is not straightforward and to understand it better one of the key aspects is to study the system in its entirety. PPI are essential to regulate gene expression, not only for the PIC formation but also for the transduction of external signals into the expression of one gene or another. All the mentioned approaches for reconstructing GRN, and others [231]–[233], are not capable of modelling redundant and parallel pathways independently of the type of input data. In addition many genes are essential (e.g.~20%of yeast genes) [234] but, although they are known to play crucial roles, cannot be used as starting points. Knockout experiments are done one gene at a time and this only allows a static picture of the situation and last, but not least, regulatory networks use backup mechanisms [235]–[237] and the majority of the methods that reconstruct the dynamic regulatory networks, reviewed in Gitter *et al.*, (2010) [238], do not explain the mechanisms of activation of the TFs involved. For all these reasons, in our study, we decided to use the signalling dynamic regulatory events miner (SDREM) [239] which requires a small set of starting points, upstream proteins that are known to initiate the response to the perturbation, and gene expression data to identify the TFs that control the differentially expressed genes. Sensory proteins are linked to the active TFs, identified with the dynamic events regulatory miner (DREM) [240] using a network orientation algorithm.

1.4 Networks and diseases

In a simplistic view anomalies in gene activity may lead to a pathological condition. Although gene expression signatures have been successfully implied to classify subtypes of cancer [31], the underlying pathway changes, are far from being entirely understood. What is known at the network level, for some diseases like type 2 diabetes, glioblastoma or coronary artery disease, is that small changes in many genes cause the disease phenotype and not very heavy changes in just a few genes [241]. This reinforces the idea that complex disease phenotypes unlikely result from the behaviour of a single disease-gene.

Inclusion and deletion of nodes (can be genes or proteins) are frequent events during evolution, with the duplication of genes for example, or in alternative splicing events. Although these events occur, it has been proved that the system (many model organisms have been tested) is perfectly capable of holding them and this, from a topological point of view, is thanks to its scale free organization [242], [243]. Hub proteins are encoded by essential genes and expressed in many different tissues [152], [244]–[246]; their mutation or modification may have severe outcomes, including death. Consequently non-hub proteins mutation just creates "variation" [133].

Disease driving genes tend to cluster together in the periphery of a complex network and to create a module, a sub network [133], [134], [152], [247]. The disease driving modules principle consists in the idea that when one or more members of the module are dysfunctional it may arise a disease phenotype [142], [248]. This is based on PPI and genome-wide network studies in which it has been demonstrated that a single-gene knock-out does not affect the phenotype, while multi-gene knock-out may lead to *"in silico"* death or sickness [249]. This disease modularity implies that, after the identification of disease pathways, with the aim to highlight disease-driving

genes and identify putative drug targets, the entire disease module should be target of the treatment, paying attention not to affect essential genes because of the severe possible side effects. Gene signatures in different diseases or biochemical experiments are available from MSigDB and these can be used to find disease modules or to figure out if a certain set of genes is involved or not in a disease-phenotype [250]. When no gene set is available for the studied disease or new disease modules want to be discovered, network derived pathways are fundamental. In Figure 1-7 are summarized basic notions related to disease-genes in the context of PPI networks.

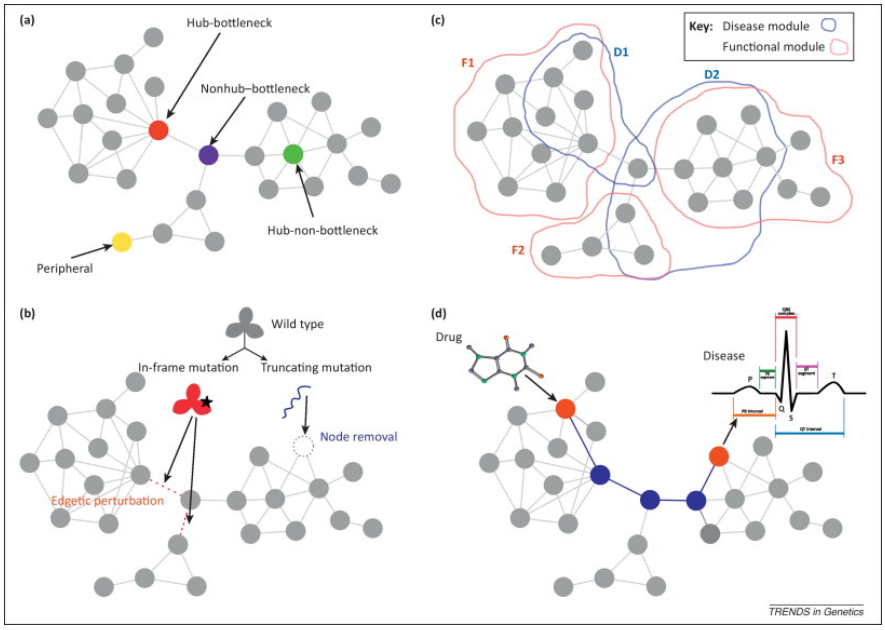


Figure 1-7 Disease-genes attributes in PPI networks. A) Scale-free PPI network is characterized by the presence of peripheral nodes, hubs (nodes with a high degree), hubs-bottleneck (nodes with a high degree and high betweeness centrality), non-hubs-bottlenecks (node with low degree but high betweenes centrality). **B)** A protein mutation may result in two different effects in the network: node removal, typical of truncating mutation, or edge rewiring, aka edgetic perturbations, typical of in-frame mutations. **C)** Nodes with common biological processes identify functional modules in a network. In the same way disease modules can be identified by a group of nodes involved in the same disease. The resulting groups may have some overlapping. **D)** The association between a drug and a potential adverse effect can be studied by examining the shortest path from the drug-target to the node associated with the adverse effect. Figure taken from [251].

1.4.1 Disease-gene prioritization and "guilt-by-association" principle

The results of experimental studies on genetic variants, usually, highlight huge genomic regions with thousands of genes associated with a certain disease. Experimentally filtering this large set of candidates to obtain only the causal disease-gene(s) can be very expensive, thus computational methods, reviewed in [252]–[254], came into play. Evidences used to link genes with each other, and with a disease phenotype, are described below.

Table 1-3 Available data repositories for genetic variants and disease-gene associations (reproduced from [Capriotti et al., 2012]).

Database	URL
Short variations	
1000 Genomes	www.1000genomes.org
dbSNP	www.ncbi.nlm.nih.gov/projects/SNP
HapMap	www.hapmap.org
Structural variations	
dbVar	www.ncbi.nlm.nih.gov/dbvar
DGV	projects.tcag.ca/variation
DGVa	www.ebi.ac.uk/dgva
General variants associated with phenotypes	
HGMD	www.hgmd.org
OMIM	www.omim.org
SwissVar	swissvar.expasy.org
GWAS and other association studies	
dbGaP	www.ncbi.nlm.nih.gov/gap
EGA	www.ebi.ac.uk/ega
GAD	geneticassociationdb.nih.gov

NHGRI GWAS Catalog	www.genome.gov/gwastudies

Cancer genes and variants

ICGC	www.icgc.org
COSMIC	sanger.ac.uk/genetics/CGP/cosmic
Cancer Gene Census	sanger.ac.uk/genetics/CGP/Census
Cancer Gene Index	ncicb.nci.nih.gov/NCICB/projects/cgdcp
TCGA	cancergenome.nih.gov

Pharmacogenomic genes and variants

DrugBank	drugbank.ca
PharmGKB	www.pharmgkb.org
CTD	ctdbase.org
KEGG	www.genome.jp/kegg

Crowdsourced genes and variants

Gene Wiki	en.wikipedia.org/wiki/Portal:Gene_Wiki
SNPedia	www.snpedia.com
WikiGenes	www.wikigenes.org

Computationally-derived / meta databases

GeneCards	www.genecards.org
PhenoGO	www.phenogo.org
PhenomicDB	www.phenomicdb.de
DisGeNet	www.disgenet.org

- **Literature** Text-mining may be of help in relate genes with diseases, for example by checking co-ocurrence of relevant terms.
- **Sequence / structure** Genes with similar functions tend to be involved in the same disease phenotype. Functional similarity derives from homology that, in turn, depends on structural similarity and common sequence. Thus components derived from the sequence and structure are used to distinguish between

disease-genes and not (for example protein subcellular location, length of the coding region, sequence motifs, chromosomal location, exon number, sequence conservation, structural domains).

- **Mutations** SNPs functional annotation tools, both using predictions or existing knowledge, may contribute in link genes and diseases. Reviewed in [252], [254], [255],
- **Ontologies** Phenotypic and functional similarity, fundamental for disease-genes identification, may also be represented by the same ontological annotation (e.g., from GO [256]).
- **Pathway involvement** Many biological pathways are affected by disease-genes. If genes belong to the same path it means that, probably, they share some similar function. Thus if a set of genes belongs to a pathway known to be implicated in a disease, they are more likely disease-related genes. Gene regulatory networks (with functional links like gene co-expression) and PPI networks are extensively used [126], together with annotated paths databases (for example MSigDB [257], KEGG [258], GenMAPP [259], Reactome [260], BioCyc [261]).
- **Orthology** The information derived from orthology studies can help in the disease-genes identification in human.

Disease-gene prioritization methods make extensive use of the "guilt-by-association" principle. This consists in considering genes linked to disease-genes more likely to be implicated in a disease. It becomes clear thus the relevance for this process of the evidences discussed above. Due to the advent of high-throughput technologies, the amount of information available on pathways is rapidly increasing, for example on gene co-expression or PPIs, thus the prioritization approaches that use paths involvement as evidence are experiencing a significant growth.

1.4.1.1 Disease-gene prioritization based on PPI networks

In order to distinguish between gene prioritization methods based on pathways one has to focus on the definition of each method of proximity between gene products in the PPI network. At an early stage those methods were trying to identify new disease-candidate genes using the direct neighbourhood idea: checking if the protein encoded by the gene of interest interacts with the product of genes already known to be associated with the disease (seeds) [262]–[267]. The direct neighbourhood approach can be extended to indirect neighbourhood, thus considering nodes neighbours of neighbours in the network and this is achieved using clustering methods [268], [269]. The newcomers in the field are global topology based approaches that try to take full advantage of the network topology. The shortest distance with respect to seeds has been used in order to rank the remaining nodes [266], [270]–[273], kernel based diffusion over network edges has also been applied, so to reduce the importance of further nodes [274]–[276], or random walks, that assign to each node a probability of ending up in the node during a random walk through the edges of the network [271], [272], [277], [278]. These last were demonstrated to outperform the local topology based methods [269], [271], [278]. Random networks have been proposed to normalize the scores of the prioritization algorithms in order to correct the bias towards highly connected known disease nodes in PPI networks [279].

To overcome the problem of data incompleteness (leading to an increase in the false negative results) and noisiness (leading to more false positives) another approach consists in data integration, thus in the addition of gene expression data or functional similarity to increase the quality of the network upon which genes are prioritized [86], [267], [270], [280], [281].

Table 1-4 Available disease-gene prioritization tools (adapted and updated from [254] **).**

Method	URL	Description*
aGeneApart	www.esat.kuleuven.be/ageneapart	L
BITOLA	ibmi.mf.uni-lj.si/Bitola	L
CAESAR	polaris.med.unc.edu/projects/Caesar	ESPNOML
CANDID	dsgweb.wustl.edu/hutz/candid.html	ESPNL
DADA	compbio.case.edu/dada	PL
DomainRBF	bioinfo.au.tsinghua.edu.cn/domainRBF/gene	SOML
ENDEAVOR	www.esat.kuleuven.be/endeavour	ESPNOL
G2D	www.ogic.ca/projects/g2d_2	ESPOL
GeneDistiller	www.genedistiller.org	ESPNOL
GeneMANIA	www.genemania.org	ESPNOL
GeneProspector	www.hugenavigator.net	SNML
GeneSeeker	www.cmbi.kun.nl/GeneSeeker	NL
GeneWanderer	compbio.charite.de/genewanderer/GeneWanderer	PNML
Genie	cbdm.mdc-berlin.de/tools/genie	ESPNL
Gentrepid	www.gentrepid.org	ESPL
GUILD/ GUILDify	http://sbi.imim.es/web/GUILDify.php	ESPNOL
MedSim	www.funsimmat.de	SPNOL
MimMiner	www.cmbi.ru.nl/MimMiner	SL
PGMapper	www.genediscovery.org/pgmapper	ESPL
PhenoPred	www.phenopred.org	SPO
PINTA	www.esat.kuleuven.be/pinta	EP
PRINCE	www.cs.tau.ac.il/~bnet/software/PrincePlugin	EP
PolySearch	wishart.biology.ualberta.ca/polysearch	L
PosMed	omicspace.riken.jp/PosMed	L
PROSPECTR	www.genetics.med.ed.ac.uk/prospectr	SNML
SNPs3D	www.snps3d.org	SPNOML
SUSPECTS	www.genetics.med.ed.ac.uk/suspects	ESPNML
ToppGene	toppgene.cchmc.org	ESPNOL
VAAST	www.yandell-lab.org/software/vaast.html	EM

*(E) Experimental observation (S) Sequence, structure, tissue specificity (P) Pathway involvement (N) Non-human data (O) Ontologies (M) Mutations (L) Literature

For the purposes of this thesis the genes underlying inheritance linked disorders (GUILD) network prioritization framework has been used [282]. It is freely available and contains four network-based gene prioritization algorithms: NetShort, NetZcore, NetScore and NetCombo. The approach used in GUILD differs from the one of other network prioritization algorithm in the way the information is spread through the network topology. NetShort "shortens" the path length between two nodes if this contains seeds. NetScore takes into account that more than one shortest path may exist from one node to another. NetZcore is capable, by randomly substitute nodes but maintaining the original network configuration, to determine the biological relevance of the neighborhood of a node. Finally, the consensus method we choose for our analysis, the one that better performs with respect to existing prioritization algorithms because it combines the previously described ones: NetCombo.

1.4.2 Host-pathogen PPIs

Bacteria and viruses are external pathogens that may cause infectious diseases. In the context of this thesis we are interested, at the molecular level, in the interaction between the pathogen and its host(s) and how this occurs in terms of PPIs. Pathogen proteins physically bind with host proteins to manipulate its biological processes and being able to thrive, grow and multiply, without host's immune system repressive intervention.

1.4.2.1 Experimental studies

In order to understand these interactions, experimental techniques can be adopted (as it has been the case for *Herpesvirus* and human cells [283]). We can split these experimental techniques in two groups:

- **Small-scale methods.** Biochemical, biophysical and genetic experiments (co-immunoprecipitation, far-western blot analysis, co-crystallization, pull down assays) that involve a small set of

proteins. These methods tend to be very time consuming but their results are very reliable.

- **Large-scale methods.** High-throughput methods (yeast two-hybrid, affinity purification, mass spectrometry, microarrays among others) to scan the entire proteomes of the studied organisms. Usually those experiments are relatively rapid in relation to the amount of data they produce, their cost is becoming reasonable but they tend to produce a much higher false positives rate.

1.4.2.2 Interactions predictions

Computational methods complement the wet-lab based ones. They take advantage of previous experimental results to make new predictions that, in turn, will need validation to weed out false positives and increase the set of real interacting proteins between the pathogen and the host. Computational methods in the field use supervised machine learning algorithms to train the data with many other features (for example protein sequences from Uniprot [284], protein families from Pfam [285], protein structure and domains from PDB[286], gene ontologies from the GO database [256], gene expression from GEO [287], interactions between protein families from iPfam [288], protein domain interactions from 3DID[289] among others). Once the training phase is over the problem becomes a classification problem in two categories: "interacting" and "non-interacting". Some problems with the machine learning may arise because of the unbalanced set of "interacting". These are just a small proportion of the total number of proteins. It is very recent a catalogue of non-interacting domains but we are still far from the "non interacting proteins database". For various reasons some of the properties mentioned before may not be available for some proteins and the available data are stored in many databases but only a few pathogens have been intensively studied (PIG [290] for example only contains data for 12

pathogens). All these aspects may heavily influence a machine learning approach intended to classify pairs of proteins in two categories.

Another family of computation tools are based on the idea that homolog proteins, preserving their functional behaviour, will also maintain their ability to interact thus will share interactions [169], [291]. The so called interlog based tools include, among others, InterlogFinder [292], PPISearch [293] and a recently developed perl module [294] but for the analysis contained in this thesis we used the BIANA interlog prediction server (BIPS) [295]. As the name suggests, BIPS is based on the integration framework BIANA [199] allowing the usage of a very large dataset of PPIs (derived from the integration of 10 databases: DIP [192], HPRD [193], IntAct [296], MINT [297], MPact [196], PHI_base [298], PIG [290], BioGRID [191], BIND [299] and VirusMINT [300]) and a fine tuning regarding the prediction parameters.

1.4.2.3 Host-pathogen PPIs databases

Specific manually curated databases for host-pathogen PPIs are PHI-base [298], PIG [290], HPIDB [301] and PHISTO[302]; they collect interactions from low and high-throughput sources.

1.4.3 The drug-target space

Networks can be used to represent interactions between drugs and the genes encoding their modulated proteins, aka targets. Doing so, researchers found out the so called drug "promiscuity": this modulation does not only occur on the proteins specifically targeted by the drug, but also to a others [303]. This discovery is crucial for the so called "drug repositioning", that consists in finding new therapeutic indications to existing drugs. Sometimes multi-targeting is intrinsic for the therapeutic efficacy [304] but the expansion of the drug-target space derived from the promiscuity discovery led to a deepening in its study. Among these studies the ones on side effects,

in which drugs with similar unexpected side effects lead to group together their target genes and these, in turn, are used to identify new targets, thus new therapeutic indications [305]. From this approach it has been possible the creation of a database containing all repositioned drugs due to their side effects, it is called SIDER [306]. Another validated approach [304], based on chemical properties, involves a previous classification of ligands according to their chemical similarity, then, to increase one drug target space, an algorithm similar to BLAST [307] is applied in order to find additional drug targets with similar affinities to its ligands [304], [308]. The high interconnection between drugs and their targets is evident but there is still a lot to explore in the area.

1.4.3.1 Integration with gene signature

The basic approach to find drug targets is scanning collections of approved compounds but thanks to the recent explosion of high-throughput technologies, gene expression profiling has fully entered in the drug finding and repositioning processes. Its ability to find the molecular changes that occur during disease progression [309] allowed a better comprehension of the relations between physiological profiles and gene expression signature of test animals [310]. Given that genomic profiles are capable of identifying all biological states [309], [311], [312], gene expression can describe disease phenotypes [313] and it can be used to measure physical reaction after exposure to a compound, thus also for inferring drug effectiveness. A collection of gene signatures in response to different compounds is represented by the connectivity map database (cmap) [314]. Thanks to this database the user can compare its own disease signature with the ones in the database and find a potentially effective compound. Other databases that contain genomic profiles obtained with drug perturbation studies are the gene expression omnibus (GEO) [287], and, specific for cancer, the cancer cell line encyclopedia [315], a collection of cancer data that includes gene expressions, sequencing data and chromosomal copy number from 947

human cancer cell lines. Good examples of usage of these databases for finding new drug disease association are the study of Sirota *et al.*, (2011) [316] and Dudley *et al.*, (2011) [317].

1.4.3.2 Drug-Drug interactions

Unpredictable clinical effects arise when drugs interact between themselves. This happens when the pharmacologic effect of a given drug is altered by the action of another drug and this can be the cause of severe (or not) adverse drug reactions. Interaction between drugs can be divided into three categories:

- **Pharmaceutical.** The cause is a physical or chemical incompatibility.
- **Pharmacokinetic.** The cause is the interference of one drug in the absorption, distribution, metabolism or excretion of another. The target sites will receive a different amount from the planned.
- **Pharmacodynamics.** This type of interaction occurs if drugs are antagonistic, additive, synergistic or with an indirect pharmacologic effect one on the other.

Although the majority of studies have been focused on pharmacokinetic, a large number of interactions can be explained only with pharmacodynamics. To address the problem, computational methods took two main approaches:

- **Similarity based.** Measures drug information and predicts interactions. The study of Gottileb *et al.,* (2012) [318] is an example. Many of the methods mentioned to identify new drug targets can be applied also to study drug interactions depending on the type of data available.
- **Knowledge based.** Scientific literature is used to predict the type of interaction, together with information available on

the FDA adverse event reporting system and electronic medical record database.

The biggest limitation of these computational methods is that for novel drugs no information is available and the fact that considering the action of a drug and its effects in the context of a complex biological network is quite unusual.

1.5 Thesis motivation

As we have shown during this introduction, every organism carries an impressive amount of functions as much in its normal state as when this is disrupted. Complex biological processes can be represented through networks, thus analysed with methods derived from maths, statistics and physics. Biology has become a much more interdisciplinary field and the advent of high-throughput technologies also accelerated this process and computer science came into the field. This lead to a massive development of *in-silico* technologies trying to study, understand and replicate natural processes. The interpenetration between *in-vivo* and *in-silico* is essential: to create and validate hypothesis in a two-way connection. This thesis contains this essential two-way link between the two worlds. Using available experimental and computational technologies we addressed the problem of understanding the modification in the normal signalling path of the cell derived from some external agent. Starting from the effect at the gene regulation level, measured with high-throughput techniques like microarrays, using specific clustering algorithms we derived sets of transcription factors that may regulate the behaviour of the clustered genes. In this sense, using gene regulatory links, we tried to answer the question: "WHO is responsible for the observed gene expression response?" We integrated this approach with the analysis of the paths, in the protein-protein interaction network, that also link the initial disruptive cause to the final response, with the aim to answer the question: "HOW does the organism react to a certain stimulus? Which paths are modulated?"

2 OBJECTIVES

This thesis aims to fulfil the following objectives:

- Identify, with currently available computational tools, biological pathways and cell networks that underlie a specific phenotype, e.g. infection process.
- Identify the transcription factors, or main regulators (MRs), from a set of genes with similar behaviour (gene signatures) by integrating DISPOM's [82] predictions, i.e. putative binding motifs, and the information on specific transcription factors collected for example from JASPAR [84] or CIS-BP[94]. The link between the predicted MRs and the regulated genes, would lead to the creation of a gene regulatory network (GRN).
- Combine predicted GRN and protein-protein interactions (PIN) to derive a combined network: GRN + PIN underlying the given phenotype.
- Apply a message-passing algorithm from the predicted MRs, using the derived GRN+PIN network, to pinpoint the regulatory elements of the genetic signature and the molecular basis of the given phenotype.
- Apply this strategy to two specific systems describing the infection process of *Salmonella spp.* in two different hosts: *Arabidopsis thaliana* and *Homo sapiens*. Implied in this objective is the use of interology relationships to infer cross-species interaction networks.
- Demonstrate the potential of the approach in the field of pharmaco-dynamic drug-drug interactions. To achieve this objective drugs-specific genetic signatures will be extracted from the cmap database [314] and common MRs identified. Drug targets from Drug Bank [319] will be used as emitters for a message-passing algorithm and the scores of the MRs will be compared in the cases of mono-drug and drug combination therapies.

3 SALMONELLA INFECTION IN ARABIDOPSIS

In this chapter I introduce a method for unveiling main regulators of sets of genes with similar behaviour during *Salmonella spp.* infection in *Arabidopsis thaliana*. I combined this analysis with the application of a message-passing algorithm on a predicted host-pathogen protein-protein interaction network and with the use of a specific software to reconstruct the dynamic and causal response pathways related to the invasion. The problem is tackled from both sides: predictions of new putative *Salmonella spp.* effectors are made explicit and, on the other hand, *Arabidopsis thaliana*, key regulators are proposed and subsequently experimentally validated.

This article is in the process of being submitted.

Supplementary Tables S1, S2, S3 and S4 are not included in this book but are available on the CD copy of this thesis.

3.1 Unravelling signaling pathways involved in *Arabidopsis* response to *Salmonella* infection using gene-expression and predicted cross-species protein-protein interactions.

Daniel Poglayen[1], Ana Garcia[2], Oriol Fornes[1], Jascha Casadio[1], Javier Garcia-Garcia[1], Guy Zinman[3,#a], Ziv Bar-Joseph[3], Heribert Hirt[2,#b], Judith Klein- Seetharaman[4,#c], Baldo Oliva[1*]

[1] Structural Bioinformatics Laboratory, Universidad Pompeu Fabra, Barcelona, Catalonia, Spain

[2] Unité de Recherche en Genomique Végétale, URGV, Evry, France

[3] System Biology Group, Carnagie Mellon University, Pittsburgh, Pennsylvania, United States of America

[4] Department of Computational Biology – School of medicine, University of Pittsburgh, Pittsburgh, Pennsylviania, United States of America

[#a] Current Address: Healthcare, SparkBeyond, Pittsburgh, Pennsylviania, United States of America

[#b] Current Address: Center for Desert Agriculture, King Abdullah University of Science and Technology, Thuwal, Kingdom of Saudi Arabia

[#c] Current Address: Metabolic and Vascular Health, University of Warwick, Coventry, United Kingdom

* Corresponding author

E-mail: baldo.oliva@upf.edu

3.1.1 Abstract

Salmonellae are gram-negative bacterial pathogens capable of infecting a wide range of organisms, including Arabidopsis among others. To ensure survival and propagation, the bacteria secretes effector proteins into the host. However, the roles of some of them and the mechanisms activated on Arabidopsis defence-response still remain unknown.

In this study, we took a system-wide approach to fully grasp the mechanisms by which *Arabidopsis* responds to *Salmonella* infection. We integrated the analysis of high-throughput data together with computationally predicted protein-protein interactions to identify the key genes/proteins involved in the response to infection. Based on time-series microarray data of *Arabidopsis* infected samples, we clustered genes with similar expression profiles and predicted potential transcription factors that could regulate each cluster. We further analysed putative signalling pathways pointing on the activation of the predicted regulators and combined the approach into a rational selection of candidates to trigger the Arabidopsis response under Salmonella infection.

As a result of this analysis, WRKY18, WRKY40 and WRKY60 were selected and knocked-out to validate their role as main-regulators. The predicted gene-regulatory network model was tested with a designed qPCR experiment of a selected group of genes. The effect was in agreement with the model, confirming the role of WRKY18, WRKY40 and WRKY60 in the defense–response of Arabidopsis and justifying the signalling networks derived from the model.

3.1.2 Author summary

Salmonella typhimurium is the causative agent of various human and animal diseases. According to the World Health Organization, salmonellosis is the most frequent foodborne disease with around 1,5 billion infections worldwide yearly. Although hygiene conditions have considerably improved, the number of *Salmonella* infections has increased over the last decade due to antimicrobial resistance, as well as the ability of Salmonella to hide inside host cells. To address this global health problem, we present the results of a novel multidisciplinary approach in which we combined biological data with computational predictions. Our results, which have been experimentally validated, suggested three proteins as key factors during *Arabidopsis* early response to *Salmonella* infection.

The identification of key genes/proteins involved in the process of infection is a key step towards a hypothetical scenario in which we will be able to fully grasp the mechanisms of host-pathogen interaction, allowing changing or modulating the infection virulence.

3.1.3 Introduction

Salmonella includes several members of the *Enterobacteriaceae* family that can be discriminated into two main species: *S. enterica* and *S. bongori*. The first one, in turn, has been divided in more than 2500 *serovars* depending on different biochemical characteristics, as for example the composition of their somatic and flagellar antigens.

Thanks to these characteristics, the mentioned pathogens can invade both cold and warm-blooded hosts. Salmonellosis, by definition, is the infectious disease of humans and animals caused by organisms of the before specified two species of *Salmonella* [320]. According to the World Health Organization *S.enterica*, subs. *enterica* and serovar Typhimurium, is the major human pathogenic serovar. Only in the United States it is responsible for almost 1 million infections and more than 350 deaths every year [321].

Thus *Salmonella enterica* subspecies enterica serovar Typhimurium, often written as *Salmonella* Typhimurium is capable of finding hosts not only in the *Animalia* but also in the *Plantae* kingdom, including *Arabidopsis*, *Medicago sativa* (alfalfa), *Solanum lycopersicum* (tomato plant) among other species with green leafs [322]–[331]. S. Typhimurium uses natural openings and sores to assault plant tissues [324]–[326] where, at different levels, it can endure and reproduce [322]–[324], [328]. The immune system of plants is able to recognize pathogenic or beneficial invasions and reacts accordingly. The inoculation of S. enterica provokes the activation of many defence mechanisms, including stomatal closure, ROS production, activation of mitogen-activated protein kinases (MAPKs) and defence gene expression [324], [330], [332], [333]. One of the plant strategies to control endophytic colonization [323]–[325], [328], [331] involves a PAMP-triggered system that consists in the detection of pathogen (or microbe)-associated molecular patterns (PAMPs) by membrane-resident receptor kinases [334].

On the other hand, one of the known mechanisms of infection employed by S. enterica consists in the use of a type III secretion system-1 (T3SS-1) expressed at the extracellular stage and a T3SS-2 that is induced after internalization into animal cells. Both systems are encoded by two, so called, Salmonella pathogenicity islands (SPI-1 and SPI-2 respectively) and secrete a set of effectors proteins that are related to the bacterial pathogenicity [335]. Studies on plant invasion using S. Typhimurium mutated in both secretion systems have highlighted an enhancement of the host immune reaction, with a consequent reduction of the bacterial proliferation, which, in turn, leads to the idea that T3SS effectors play a key role in *Salmonella* invasion of plants [330], [332]. In this regard, the Salmonella gene prgH, which is known to encode a constituent of the T3SS-1 needle complex, is expressed only under determined culture conditions. This observation suggests that prgH is activated by factors either present or secreted on the Arabidopsis surface. Interestingly a S. enterica type III secretion system (T3SS) prgH- mutant, that would cause deficiencies in animal cell invasion [336], [337], induced stronger defence gene expression than wild type (WT) bacteria in Arabidopsis, which suggests that T3SS effectors are involved in host defence suppression [338].

Although much is known about the infection strategies of the pathogen, there are still open questions regarding the specific functions of the known effectors and the variations in virulence related to the different hosts. We were interested in understanding the mechanisms of host-pathogen interaction thus in studying the dynamic of the infection process and identifying which genes are involved during its different stages. The potential of a complete understanding of such mechanisms will lead to address existing or novel therapeutic strategies to control the infection.

For this purpose the analysis of time series expression data appeared a straightforward tool but, as pointed out in Ernst et al., 2005 [125], although there have been time series experiments with as many as 80 time points [339], almost all time series are much shorter. Specifically in the case of Salmonella, infection occurs quickly and obtaining microarray data for many time points at short intervals is technically challenging and expensive.

Furthermore, differential expression between infected cells or tissues and non-treated controls identifies the main transcriptional effects of an infection. However, an infection process may not necessarily result in a significant increase or decrease of the targets′ mRNA level [340]. Also, there are cases in which genes with related functions show very different expression profiles or even exceptional cases of inverse correlation [341]. In this work we have studied the systemic actions of plant taken upon invasion by including Protein- Protein Interaction (PPI) networks in our analysis. Early studies [342], [343] on possible relations among mRNA and protein expression level pinpointed, to some degree, a correlation between expression levels and protein abundances [344] and also the association between PPI and gene expression [345], [346]. Since then, many gene prioritization algorithms (Reviewed in [347], [348] and benchmarked in [349]) , use seeds, genes known to be related to a certain disease /condition/phenotype derived from the literature, GWAS studies or other sources, for a functional association between genes. In other words disease-genes are clustered o close in the network, thus being "similar" genes. The concept of gene similarity may include, among others, text mining, pathway membership, functional annotations, protein properties, sequence, co-expression and closeness in protein–protein interaction (PPI) networks [350].

In this work, we used the wild type *Salmonellae* strain 14028s (WT) and a *prgH*- mutant (that we will call prgH-). We combined short time

microarray analysis of Arabidopsis infected with Salmonella WT or prgH-mutant with the study of predicted Protein-Protein Interaction (PPIs) networks. We used predictions of interactions based on interologs (when there are known interactions of similar proteins) by including interactions obtained by Tandem Affinity Purification (TAP) methods. We analysed the biological pathways involved in Salmonella infection of Arabidopsis using Shortest Paths (SPs) analysis, a gene-prioritization method such as GUILD [282], and a specific method to unveil signalling pathways, SDREM [32]. This last is a computational method that integrates condition specific time-series expression data with PPI and Protein-DNA interactions. In addition, based on the idea that genes with similar behaviour can have one, or more, common regulator(s), we introduced also a protocol of prediction of Main Regulators (MRs) that integrates ChIP data. A graphical representation of our set of hypothesis is shown in Figure 1 and detailed in the results section.

Finally, we confirmed our findings with two experiments: a microarray on samples harvested 30 minutes after Salmonella inoculation and a qPCR experiment performed with a mutant of Arabidopsis lacking (by knockout) the most relevant MRs predicted by our approach.

3.1.4 Materials and methods

3.1.4.1 Microarray data

We used microarray data on the response of *Arabidopsis* to *Salmonella* infection[338], (both 14028s and prgH- mutant). The data comprise two independent hybridizations, consequence of two biological repetitions, on CATMA arrays [95], [351], [352], of samples harvested at 2, 4, 6, 12 and 24 hours after inoculation of Salmonella (WT or prgH-). The data can be

accessed at CATdb (http://urgv.evry.inra.fr/CATdb/, Project: RA11-01_prgH-) and at Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/, accession no. GSE38828). For the sake of comparison, we further applied the invariant normalization method[353] contained in the DNA-Chip Analyzer software [103], [354] using as baseline the sample with median overall intensity. The resulting box-plots and MAXY-plots confirmed that the normalization step smoothed any differences among the different samples (see Supplementary Information Figure S1).

3.1.4.2 Clustering of genes with similar profiles

We clustered *Arabidopsis* genes according to the similarity of their time expression profiles upon Salmonella infection (wilde-type and prgH-mutant form) on CATMA arrays (see before). For a global view of time-series data, we applied the Short Time-series Expression Miner (STEM) algorithm [125], which is specifically designed for clustering genes looking at their expression profiles derived from microarray experiments with a few time points (~8 time points or fewer). We computed Gene Ontology (GO) enrichments of the clusters using the default STEM parameters (see details on STEM application in the Supplementary Information). We computed the correlations among the clusters of the two infections (WT and prgH-) and identified TFs within each cluster, obtained from the Cis-BP database [94]. Similar profiles of both infections were identified by a correlation higher than 0.99 and low (<0.0001) or lowest significant p-value, calculated with the hypergeometric distribution of the number of shared genes. Correlated profiles were named with the same code for further analyses (usually the same as the profile code for wild type infection, this was named the merged-code).

3.1.4.3 Prediction of main regulators

For each cluster obtained with STEM (see before), we retrieved the promoters of all of its genes from the database AGRIS AtcisDB [355]–[357]. Then, with the DISPOM program [82], we extracted a putative binding-site motif common to the maximum number of genes within the cluster. We used the promoters of the genes in the remaining clusters as background. Potential binding motives for each cluster were reported by means of position-weight matrices (PWM) if they reached a p-value smaller than 10^{-4}. Then, with TReg comparator [215], we searched for matches between the retrieved PWMs and the PWMs available for Arabidopsis TFs. We used a dissimilarity score of 0.9 to accept the TFs as potential main regulators (MRs) of the cluster. Finally, we used these MRs to build a gene regulatory network and identify MRs common to WT and prgH- (hereafter named CMRs). Some MRs were found within clusters while others were not. As expected, highly similar profiles of both infections were regulated by CMRs (we specifically named them CCMRs). MRs would regulate the expression of specific clusters (i.e. a specific profile). Thus, the probability that a TF would act as MR of two clusters was calculated using a hypergeometric distribution formulae:

$$pValue = \sum_{k=L}^{k=\min(n,R)} \frac{\binom{n}{k}\binom{M-n}{R-k}}{\binom{M}{R}}$$

using the total number of TFs of Arabidopsis as background (M), being R the total number of predicted MRs, n is the size of the set of MRs that regulate each cluster, and L is the number of MRs that we found in common (CMRs).

3.1.4.4 Cross-species network

To infer the complete Arabidopsis-Salmonella PPI network (PIN), we used the server BIPS[295]. We set the conditions of sequence similarity as follows: maximum blast e-value threshold 0.001, percentage of identical residues limited to 60%, 80% minimum coverage between Salmonella-query and Arabidopsis-template sequences. We applied the "matrix" model for co-complex methods, such as tandem affinity purification. The "resulting network containing Salmonella and Arabidopsis proteins was named "TAP" network. We then filtered those interactions retrieved using co-complex methods, obtaining a subset of interactions that we called "NOTAP". Results presented in the manuscript refer to the TAP network, and we have included the most restricted analysis (NOTAP network) in the Supplementary Information.

3.1.4.5 Pathways reconstruction

We used the Signaling Dynamic Regulatory Events Miner (SDREM) [239], [350] to reconstruct the dynamic and causal response pathways related to the infection and highlight the most relevant Arabidopsis TFs using the time-series data. We split the results, as before, in groups of potential main regulators common for the infection by WT Salmonella or prgH- mutant form (CCMRs and CMRs). SDREM integrates condition-specific time-series expression data with general PPI and protein-DNA data. Starting from a set of known source-target interactions, the algorithm tries to orient them and then applies a variant of the Dynamic Regulatory Events Miner (DREM) [240], [358] to identify the active TFs in the response at each time point. The iteration of these two methods predicts additional TFs and other proteins that can be involved in the response pathways. We ran SDREM using the default parameters, integrating the gene-regulatory network provided by DREM (associations between TFs

and their regulated genes) and the PPI data of the Salmonella-Arabidopsis network obtained before.

3.1.4.6 Gene prioritization

We predicted proteins that could be involved in salmonellosis using GUILD [282], a network-based gene prioritization tool. The method requires as input a list of genes whose implication in a disease is known, also called seeds, and a PPI network. We applied the NetCombo message-passing algorithm to transfer disease-gene association through the network and identify new putative disease-associated candidates (in this study we search for candidates associated with the infection and the response to infection). We applied default parameters: an initial score of 1 for seeds and 0.01 for non-seeds, with edge weight of 1, no more than 5 iterations and up to 100 sampled graphs for Z-score calculation (see references [33], [359], [360]). The PPI networks were the cross-species networks defined previously: TAP and NOTAP. We used different seeds depending on the search: *Salmonella* effectors to find the specific connection with potential MRs and other TFs of Arabidopsis; all *Salmonella* proteins, to find new potential effectors if they were connected with Arabidposis TFs and in particular predicted MRs; CMRs (with special attention to CCMRs), to find Salmonella proteins or transmembrane receptors that could eventually trigger the Arabidopsis response. Then, according to the seeds used, we selected the best ranked (top 20%) TFs of *Arabidopsis*, or *Salmonella* proteins, or *Arabidopsis* transmembrane proteins.

3.1.4.7 Shortest paths analysis

We obtained shortest paths (SPs) smaller than 4 steps between CMRs and membrane proteins using NetworkX[361]. We also investigated SPs, with the same characteristics as before, between CMRs, with special attention to CCMRs, and Salmonella proteins, highlighting the ones that involve known effectors.

3.1.4.8 Detection of potential *Arabidopsis* TFs among *Salmonella* proteins

We tested the hypothesis that a Salmonella protein could act directly as a host TF. We checked potential homologs between *Arabidopsis* TFs and *Salmonella* proteins on the basis that two TFs are more likely to bind (consequently promoting the transcription of the same set of genes) if their sequences are highly similar and have common PPIs [237]. For the criterion of sequence similarity we used Rost's sequence identity curve of the twilight-zone [362], and forced to share at least one DNA-binding domain from Pfam [285]. For the criterion of sharing PPIs, we used a threshold of at least one common interactor (this idea is graphically represented in the Supplementary Information Figure S2).

We used the sequences of the 1,727 *Arabidopsis* TFs and the *Salmonella* proteome from UniProt [284]. For each Arabidopsis TF, we performed a BLAST [363] search against *Salmonella* and identified all hits according to Rost's sequence identity curve [362]. For the Pfam-based orthology, the sequences of both *Arabidopsis* and *Salmonella* proteins were scanned against Pfam [285] using HMMER (version 3.0) [364]. We only considered hits over the HMMER inclusion threshold involving Pfam domains classified as DNA-binding domains.

3.1.4.9 Experimental integration with qPCR

In order to integrate our predictions on highly relevant genes during Salmonella infection, qPCR experiments were made using the knockout of three WRKY genes that had been identified with high probability as potential CMRs. These genes were selected among those involved in the mechanisms triggering the signal response of *Arabidopsis* upon infection. All knocked-out genes were selected to be in a SP<4 connecting them with Salmonella effectors. Wild-type *Salmonella* (WT) and the prgH-mutant (prgH-) infection were used for the analysis of the model describing the regulation of defence response. The expression of several genes (using qPCR) was analysed according to their role in the system as it was predicted by our approach. Therefore, the tested genes were selected among STEM's expression clusters, with explicit attention to the clusters of genes with specific profile for each type of infection (WT and prgH-) and for similar profiles under both infections. Their expression levels were checked at 2 and 24 hours *post*- infection and compared with the basal expression calculated with the exposure to a Mock solution.

3.1.4.10 Early infection stage experimental integration with microarray at 30 minutes *post* inoculation

In order to check genes expression at early stages of the infection we performed a microarray experiment on two weeks old seedlings infected with Salmonella WT. The plants were treated exactly in the same way as we did for the previous microarrays. Samples were collected after 30 minutes post-inoculation of Salmonella. Data have undergone the same processes for normalization as before (see above). Since it was not a time series experiment it was only possible to search for 3-fold differentially expressed genes.

3.1.5 Results

3.1.5.1 Strategy to unveil the response system of Arabidopsis on infection

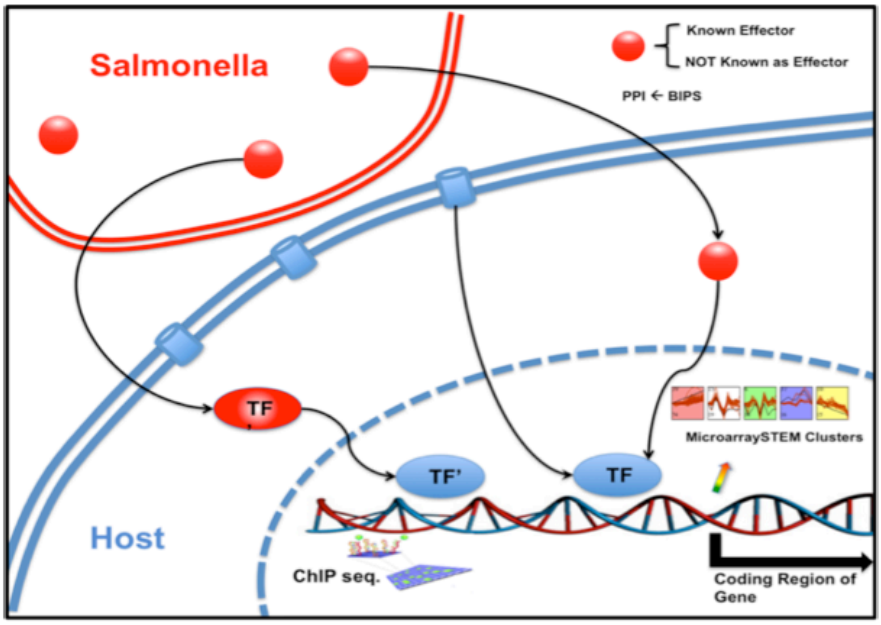


Figure 3-1:**Graphical representation of the set of hypothesis**. Salmonella proteins can be known effectors or other proteins from the Salmonella proteome (in red). Some of them can act as one of the Arabidopsis transcription factors (TF). Arabidopsis proteins (in blue) can also transfer a signal if they are located in the plasma membrane.

Our goal is to find the key players on the defence response of Arabidopsis upon infection by Salmonella. Therefore, we have pondered all possible mechanisms of signalling involved in the profiles of gene expression. We have considered the following possible sources to start the signal (see Figure 1): i) receptors in the plasma membrane that become activated by Salmonella when it approaches the cell-wall; ii) Salmonella proteins, in particular those considered effectors, cross-talking with the proteome of Arabidopsis. Once started, the signal is transferred through

interactions to the plant transcription factors (TFs) causing the specific profile expression of response, i.e. acting as main regulators (MRs).

The protein-interaction network (PIN) is constructed with proteins of both species, Salmonella and Arabidopsis (i.e cross-species network). The signalling paths are usually short, thus all transcription factors at distance shorter than 4 steps are potential candidates to receive the signal. A particular case is when the Salmonella protein can activate the gene profiles of Arabidopsis, acting as one of the TFs of Arabidopsis (i.e. the number of steps of the shortest path is zero). This implies that the Salmonella protein is similar to one or more TFs (the detection of potential activity as TFs is described in methods). In this work we have applied two methods to predict the signalling network: i) SDREM and ii) GUILD. SDREM is a method specifically addressed to solve this problem, while GUILD is a message-passing method for gene-prioritization that uses the underlying topology of the network and a set of nodes with starting non-null score (seeds) acting as source of the information. We have used GUILD to score the nodes of the network and select the top scoring receivers. The approach can use as seeds the sources of the signal as before, in which case we plan to unveil the potential MRs among the set of TFs being the receivers of information, or use the predicted MRs as seeds and then suggest potential Arabidopsis membrane receptors or other Salmonella proteins acting as effectors. Thus, top-scoring proteins identified with GUILD can also be used to reinforce MRs in both ways: 1) using as seeds Salmonella proteins (specially known effectors) and selecting TFs of Arabidopsis with highest scores as their potential targets, hence predicting potential MRs; 2) using predicted MRs as seeds and checking for Salmonella proteins within the top selected nodes of the cross-talking PIN, specially known effectors, hence reinforcing the

prediction of the MR at the same time as we predict new potential effectors of Salmonella.

Finally, we integrate all the information and select a few Arabidopsis TFs that have been predicted as MRs by several methods for further experimental validation.

3.1.5.2 Gene regulatory network

3.1.5.2.1 Time series analysis of gene expressions

We used microarray data on the response of *Arabidopsis* to *Salmonella typhymurium* infections of WT strain and prgH- mutant from CATMA arrays [95], [351], [352]. We applied the same protocol to both infections (see methods). First, genes without sufficient response were filtered out from the analysis, setting the threshold of minimum absolute expression change to 1. Second, we applied STEM to cluster genes with similar profiles and used the Arabidopsis Thaliana ontology database (TAIR/JCVI) contained in STEM [365]. We obtained 11 significant profiles for WT infection containing 732

genes (2.78% of the total number of genes in the array) (Figure 2A). Details on the genes contained in each cluster and significant GO enrichments with enough significance (p-value <0.01) are in Supplementary Information (Tables S1 and S2, respectively). Similarly, we obtained 13 significant profiles for the infection with prgH- mutant form of Salmonella, containing a total of 972 genes (3.68% of the total number of genes in the array) (Figure 2B). Also details on the genes contained in each cluster and the GO enrichments with enough significance (p-value <0.01) are in Supplementary Information (Tables S3 and S4, respectively)

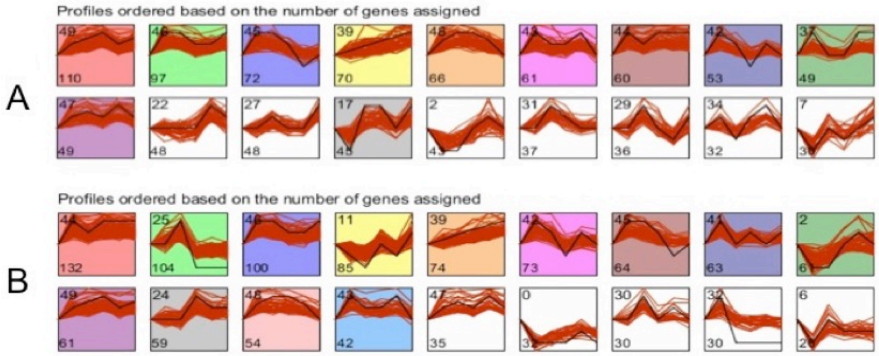


Figure 3-2: Clusters of Arabidopsis' genes producing similar time-expression profiles after Salmonella infection WT (A) and prgH- mutant (B). Significant clusters are shown in coloured background. The number on the top left corner of each coloured cell indicates the cluster/profile ID and in the bottom left corner is shown the number of genes contained in the cluster.

When comparing the profiles of the infections by WT and prgH- mutant forms we obtained 8 similar profiles as defined in methods (with correlation higher than 0.99 and a significant highest number of shared genes). In Table 1 we summarize the comparison and indicate the code for considering genes of any of the profiles from WT or prgH- Salmonella infections. Three additional pairs of clusters are compared in Table 1, but their correlations were too small to be considered similar. We also compared the enrichment of GO terms between similar profiles: a total of 1024 GO terms are the same for both infections, while 244 (18%) are specific for WT and 399 (24%) for prgH- (see details in Tables S2 and S4).

WT	prgH	C	# Shared	p-value	Common TFs	Merge
44	44	1.00	27	3e-47	AT5G49520	44
	49	0.70	6	5e-9	None	
	41	0.59	6	7e-9	AT4G01250	
	46	0.59	5	3e-6	None	
46	46	1.00	26	7e-42	AT4G36990; AT5G26920	46
	44	0.50	13	2e-15	AT3G15500; AT3G23250	
	41	0.50	9	2e-12	AT5G67450	
	49	0.60	8	6e-11	AT4G08350; AT5G13080	
	45	0.54	6	1e-7	AT4G23810	
	43	0.27	5	5e-7	None	
48	48	1.00	20	1e-39	AT5G61890	48
	47	0.59	8	2e-14	None	
	44	0.35	10	1e-12	AT3G49530	
	49	0.61	5	4e-7	None	
45	45	1.00	18	2e-32	AT5G47230; AT5G51780; AT2G14760	45
	48	0.52	10	3e-16	None	
	41	0.47	6	2e-8	AT1G51700	
	46	0.54	6	3e-7	None	
	44	0.00	5	3e-5	None	
49	49	1.00	18	4e-29	AT1G18860	49
	39	0.63	13	5e-18	AT2G47270; AT5G47370	
	24	0.66	12	1e-17	AT5G49450	
	44	0.70	13	1e-14	AT1G43160	
39	39	1.00	17	4e-29	AT3G04070; AT3G23030; AT4G31800; AT5G39610	39
	11	0.36	9	1e-12	AT1G71030	
42	42	1.00	15	5e-27	None	42
	46	0.47	8	2e-11	None	
	45	0.31	6	3e-9	None	
47	46	0.49	12	4e-19	None	-
	44	0.63	9	3e-12	None	
	49	0.51	6	2e-9	None	
17	11	0.23	9	2e-14	AT3G61890; AT5G59780	-
	2	0.20	6	9e-10	AT5G65210	
43	43	1.00	7	6e-12	None	43
	44	0.53	10	5e-13	None	
	46	0.27	7	3e-9	AT5G47220; AT2G38470	
	41	0.27	5	4e-7	None	
37	41	0.45	6	2e-9	None	
	46	0.45	6	3e-8	None	

.

Table 3-1 Summary of correlations (C) between profiles/clusters of Arabidopsis genes after response to infection with wild-type (WT) and prgH- forms of Salmonella. The number of common genes (#shared) is calculated with the p-value of significance based on a hypergeometric test (p-value). In addition, out of the common genes found, the common TFs are identified (Common TFs) and clusters with C>0.99 and sufficient common genes are renamed with the merged-code (Merge).

3.1.5.2.2 *Prediction of main regulators (MRs)*

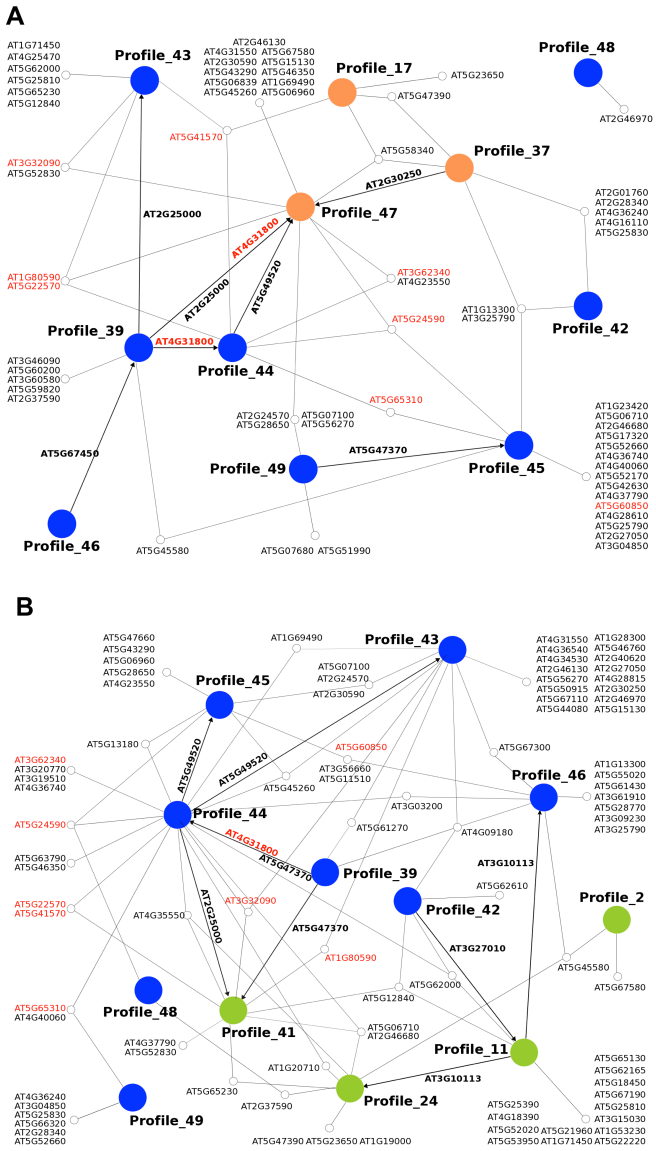


Figure 3-3: Gene regulatory network of transcription factors predicted as MRs. Nodes in the graph represent either gene clusters of Arabidopsis (profiles) obtained with STEM in response to Salmonella WT infection (A) and Salmonella prgH- infection (B) or single TF predicted as MRs. Edges between profiles indicate that the MR of a cluster was found within another cluster (the code of the MR is shown in the edge). An arrow indicates the direction of the control: the profile containing the MR towards the profile it regulates. Similar clusters of both infections are shown in blue (specific clusters for WT infection in orange and for prgH- in green) and CCMRs are highlighted in red.

We were able to predict a total of 107 putative Arabidopsis MRs for 10 out of 11 profiles obtained with the clustered data for WT infection using the approach described in method (the list of predicted MRs is shown in Table S5A in Supplementary Information). We predicted more than one potential MR for each cluster (only cluster 48 was predicted to be regulated by a single TF), but also some predicted MRs could regulate more than one profile, for example AT5G24590 could regulate genes in clusters 44, 45 and 47; AT5G22570 for clusters 43, 44 and 47; AT1G13300 and AT3G25790 for clusters 37, 42 and 45; AT2G25000 for 43 and 47; AT2G24570 (47 and 49); AT4G31800 (44 and 47), etc.. In total, only 70 Arabidopsis TFs were predicted as MRs for the response to Salmonella WT infection. In figure 3A is shown the network of regulation of Arabidopsis under Salmonella WT infection, indicating the potential MRs of each cluster. Some MRs were also clustered within a profile. This indicates the dependence between profiles and is shown in figure 3A with an arrow from the profile containing the MR towards the regulated profile, while the code of the predicted MR is in bold on top of the edge (e.g. AT4G31800 shows the regulation of profile 39 on profiles 44 and 47).

Similarly, we predicted 145 MRs for 12 out of 13 clusters obtained with the clustering of STEM of the data for Salmonella prgH- mutant form infection (Figure 3B). As before, we observed the same MR for more than one cluster: AT5G45580 (for clusters 2, 24 and 46), AT5G62000 (11, 24 and 44), AT5G12840 (11, 41 and 42) or AT5G24590 (44, 45 and 48). Thus, only 97 TFs were predicted as MRs (the list of predicted MRs is shown in Table S5B in Supplementary Information). We found 52 common MRs (CMRs) when we compared the predicted Arabidopsis MRs in response to Salmonella WT and prgH- infections. On the rest of MRs, 18 were specific for WT and 45 for prgH- infections (see the list in Table S6 of Supplementary Information). By focusing on similar profiles, we found CMRs (referred as CCMRs) only for three profiles. We found a

total of 9 CCMRs that are highlighted in red in Figure 3 and listed in Table 2.

.

Cluster WT	Cluster prgH	Common MRs	p-value
39	39	None	1.0
42	42	None	1.0
43	43	AT3G32090 AT1G80590	0.0153708865751
44	44	AT5G65310 AT5G22570 AT3G62340 AT5G41570 AT5G24590 AT4G31800	9.64153201721e-11
45	45	AT5G24590 AT5G60850	0.0166140853446
48	48	None	1.0
49	49	None	1.0

Table 3-2: **Common MRs of the merged clusters**. The significance of common MRs is calculated with an hypergeometric test (p-value).

We calculate the probability that a CMR regulates the expression of two profiles with a hypergeometric distribution (where the total number of TFs is used as background, see methods). Particular attention can be shown at AT4G31800, which is found within two similar profiles (merged-code 39) and controls two other similar profiles (merged-code 44). Besides, the two profiles identified by the merged-code 44 are controlled by a relevant number of common MRs (with a p-value $\sim 1^{-10}$).

3.1.5.3 Signalling pathways

3.1.5.3.1 Combining expression and the interaction network into oriented paths (SDREM)

We used the approach of SDREM [239], [350] on the Salmonella-Arabidopsis PIN, using Salmonella effectors as the original source of the signalling pathways and searching for Arabidopsis' TFs as potential targets. SDREM uses as input the PIN, the time-series expression of genes and the links of TFs with their regulated genes by means of TF-DNA binding promoters. We used Arabidopsis TFs from DREM 2.0 and included the potential MRs predicted in the previous steps (i.e. the links between the MRs and the genes that were predicted to regulate).

While the algorithm DREM identifies the potential TFs regulating the network, the algorithm of SDREM orients the edges of the network to generate potential signalling pathways. The approach runs 10 iterations with thousands of potential pathways. Target nodes of the network can be any node receiving the signal (being connected to the original source). Nodes' ranking arises from the percentage of pathways involving each node (i.e. running through it). Additionally, SDREM also infers what are the best potential pathways and ranks them. Thus, targets can also be inferred by selecting the nodes with more than 1% of the highest confidence oriented paths going through them.

We focus our study in candidate TFs of Arabidopsis. In Table 3 are shown the results of best-ranked nodes of the PIN, indicating those that correspond to TFs of Arabidopsis and are considered as potential MRs, being the final step of the signalling pathway that was started in Salmonella effector proteins.

A)

node	Target	Degree	SDREM score
D0ZV15	N	1	0.065
AT2G25000	Y	289	1.000
D0ZWZ8	N	543	0.028
D0ZVQ4	N	3	0.113
D0ZY43	N	8	0.709
D0ZY42	N	3	0.113

B)

node	Target	Degree	SDREM score
AT2G43140	Y	50	0.021
D0ZV15	N	1	0.078
AT1G06070	Y	32	0.020
D0ZVQ4	N	3	0.135
D0ZY42	N	3	0.135
D0ZWZ8	N	543	0.011
AT5G15850	Y	49	0.123
D0ZY43	N	8	0.652
AT1G43700	Y	51	0.019
AT4G31800	Y	289	0.734
AT4G17750	Y	25	0.083

Table 3-3: **SDREM results. In A are the results for Salmonella WT infection and in B for prgH- form**. The column "target" indicates if the node of the PIN is a TF of Arabidposis (Y) or not (N). We have included the degree of the node in the PIN (TAP network) and the score of SDREM based on the ratio of the number of oriented paths with highest confidence that go through the node (see methods).

Results were limited to those with a percentage higher than 1% of the top 1000 best-ranked paths (the same approach as in previous references[239], [350]). We analysed both Salmonella forms of infection (WT and pgrH-). Further details on a more restrictive PIN (NOTAP) are shown in the Supplementary Information (Table S7). The change to a more restricted network, with less number of edges, produces different results, accusing the high dependence on the selection of the underlying network. By using the largest network (TAP), SDREM identified two CMRs previously described: AT2G25000 and AT4G31800, both included as CCMRs. Besides these two MRs, SDREM also highlighted other potential targets when studying the infection of the prgH- mutant of

Salmonella: AT2G43140, AT1G06070, AT5G15850, AT1G43700, and AT4G17750. The satisfied paths connecting the sources (Salmonella effector proteins) and the potential CCMRs AT2G25000 and AT4G31800 are shown in Figure 4. We calculated the Gene Ontology terms enrichment of this network using BinGO [366]. The most significant enriched biological processes are shown in Figure 4 and detailed in Table S11. Still significant, but ranking in lower position, we also found the terms "defense response to bacterium" and "response to bacterium". In the graph of Figure 4, we coloured the nodes according to its functional association, using only the GO terms selected [367]. Response to stimulus (abiotic and chemical), cellular metabolic processes and small GTPase mediated signal transduction were among the top enriched processes. We neglected some very general process, such as cellular process.

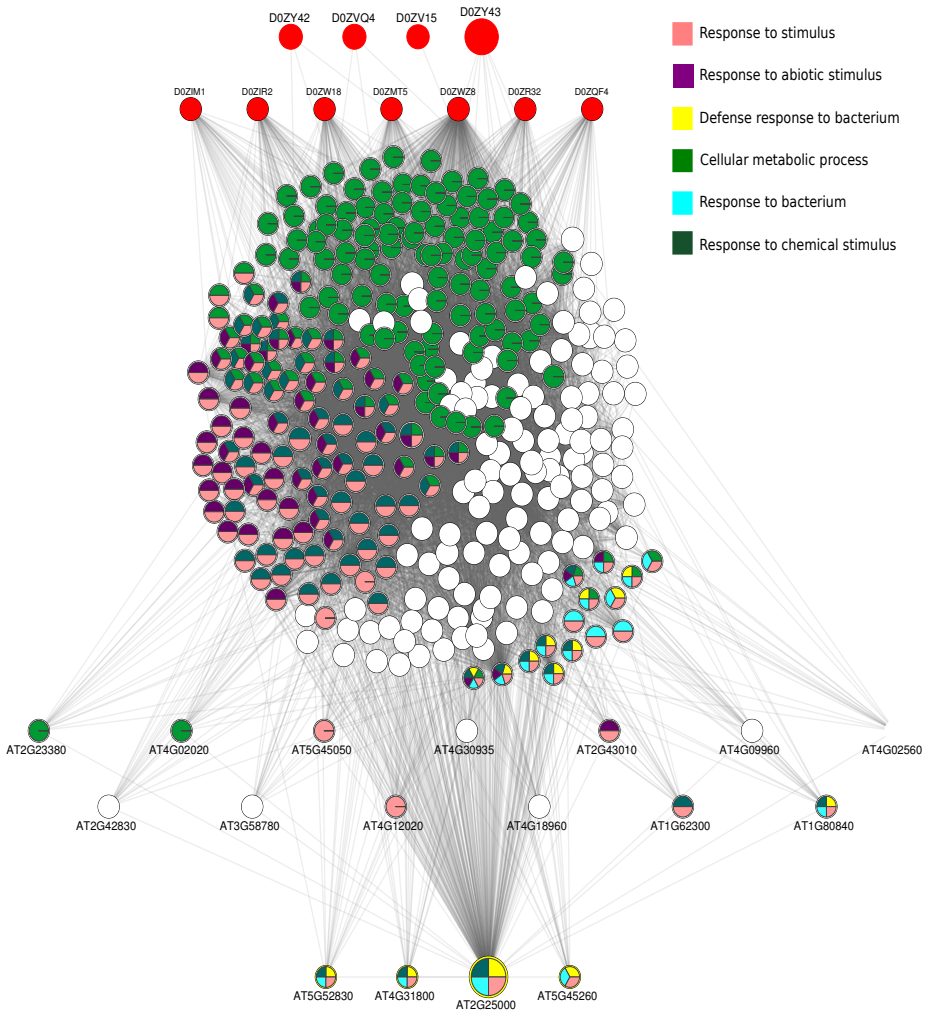


Figure 4A: Highly SDREM-scored paths between *Salmonella* effectors and *Arabidopsis* TFs. Graph representation of shortest paths selected when using SDREM for the analysis of *Samonella* WT infection. TFs are placed at the bottom (predicted MRs in the last row) and Salmonella proteins and effectors at the top. The size of nodes is proportional to the SDREM score of the protein. Functional enrichment of the network was analysed with BinGO [366] and nodes representing *Arabidopsis* proteins are coloured according to their association to the functions selected. Nodes representing *Salmonella* proteins are coloured in red. The legend shows the colours applied for the GO terms of the top enriched functions and those associated with *Arabidopsis*-defense under bacteria infection.

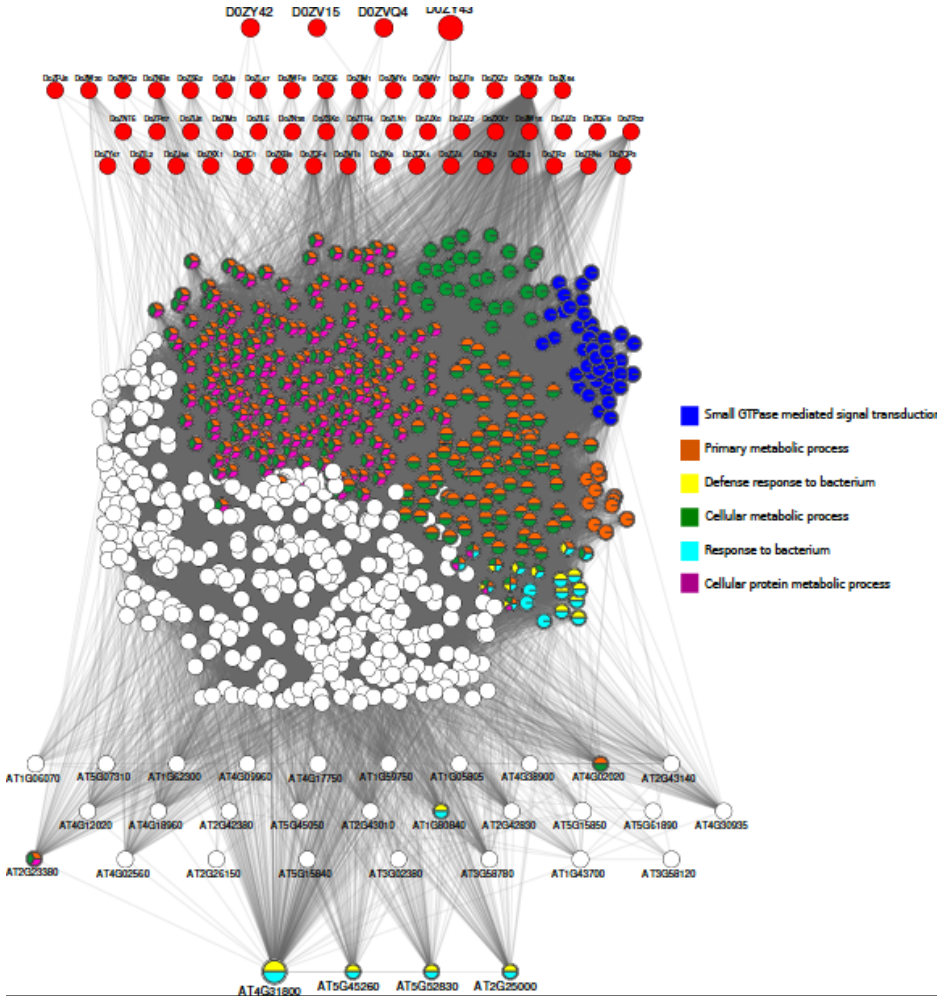


Figure 4B: Highly SDREM-scored paths between *Salmonella* effectors and *Arabidopsis* TFs. Graph representation of shortest paths selected when using SDREM for the analysis of *Samonella* prgH- mutant infection (B). TFs are placed at the bottom (predicted MRs in the last row) and *Salmonella* proteins and effectors at the top. The size of nodes is proportional to the SDREM score of the protein. Functional enrichment of the network was analysed with BinGO [366] and nodes representing *Arabidopsis* proteins are coloured according to their association to the functions selected. Nodes representing Salmonella proteins are coloured in red. The legend shows the colours applied for the GO terms of the top enriched functions and those associated with *Arabidopsis*-defense under bacteria infection.

3.1.5.3.2 *Paths shorter than 4 steps in the PIN between MRs and potential receptors or Salmonella effectors*

3.1.5.3.2.1 *Path at 0 steps: when a Salmonella protein acts as an Arabidopsis TF*

In order to cover the hypothesis of a Salmonella protein acting directly as a host TF, we tested the sequence and function similarities between Salmonella proteins and TFs of Arabidopsis (see methods).

Salmonella Uniprot entry	Sequence Identity	Coverage of target	Coverage of TF	Common Pfam domains	% Common interactions
D0ZK33	51%	95%	33%	CSD;OB_RNB	3%
D0ZKM1	52%	75%	27%	CSD;OB_RNB	3%
D0ZLC6	56%	75%	27%	CSD;OB_RNB	3%
D0ZPD2	57%	75%	27%	CSD;OB_RNB	3%

Table 3-4: **Salmonella proteins that could act as Arabidopisis TF (AT4G38680) according to the criteria of: i) sequence similarity; ii) common Pfam domains involved in DNA binding; and iii) common interactions**. Percentage of identical residues aligned (sequence identity) and coverage of the aligned region with respect to the TF (Coverage of TF) and the Salmonella protein (Coverage of target) are shown for the criteria of sequence similarity. The name of the PFAM domains in common and the percentage of common interactions over the total of interactions of AT4G38680 are shown in the last two columns, respectively.

Among the proteins that pass the twilight-zone threshold of sequence similarity, following Rost criterion [362], we searched for common interactions in the cross-species PIN. We could only find shared interactions between four Salmonella proteins and the TF AT4G38680 of Arabidopsis when using the TAP network (see Table 4 for details). According to our approach AT4G38680, also known as Cold Shock protein 2 (CSP2), not only shares a few interactions with this four proteins of Salmonella, but they also share the Pfam domains Cold-shock domain

(CSD) and the Ribonuclease B OB domain (OB_RNB), the first associate with DNA binding proteins and the second with RNA binding.

When using less restrictive criteria (i.e. without considering neither common interactions nor shared Pfam domains, and bordering the limit of the twilight-zone), D0ZQV5 from Salmonella was similar to one of the predicted CCMRs (AT4G31800) with percentage of sequence identity 40% and covering only 4% of the sequence, while other 8 proteins from Salmonella (D0ZNU3, D0ZY02, D0ZLV4, D0ZIB1, D0ZM01, D0ZV29, D0ZTT0 and SSPH2) were similar in sequence to predicted CMRs (respectively AT5G45580, AT2G27058, AT2G28340, AT4G37790, AT2G46970, AT5G23650, AT5G43290 and AT2G25000). We report in Table S8 the list of Salmonella proteins similar to predicted MRs (CMRs, CCMRs and other predicted MRs specific for the WT and prgH-Salmonella infections).

3.1.5.3.2.2 Paths between 1 and 3 steps

We calculated all paths sorter than 4 steps around the predicted MRs and retrieved all plasma-membrane proteins that could act as receptors or direct interactions with Salmonella proteins using the TAP network (details for the NOTAP PIN are shown in Supplementary Information Table S9). We found two plasma-membrane proteins: AT2G18960 and AT4G30190 within the radius of 4 CMRs (AT5G52830, AT2G25000, AT5G45260 and AT4G40060) and one CCMR (AT4G31800). Although these are not protein receptors, we could suspect some relationship between the defence response of Arabidopsis upon infection and the ATP synthesis through protons exchange.

We found several Salmonella proteins on a distance shorter than 4 steps to some of the predicted MRs. First, with the TAP network, we found a direct interaction between the Salmonella protein D0ZWZ8 and

four predicted CMRs: AT4G31800 (which is also CCMR), AT5G52830, AT5G45260 and AT2G25000. Then, at distance 3, we found 11 CMRs (AT4G31800, which is a CCMR and AT2G25000, AT5G52830, AT4G31550, AT5G06960, AT5G28650, AT5G62000, AT2G24570, AT2G46130, AT5G45260 and AT2G30590) from 4 known Salmonella effectors (D0ZY43, D0ZY42, D0ZV15 and D0ZVQ4). When using a more restrictive PIN, the NOTAP network, the number of connections diminished. First, we found a path of two steps between the Salmonella protein D0ZRU7 and the CCMR AT5G22570. Then, at three steps, we only found one path between the predicted CMR AT5G62000 and the known effector D0ZY43. Further details of other Salmonella proteins at a radius distance smaller than 4 steps in the TAP and NOTAP networks are shown in supplementary Table S10.

3.1.5.3.3 Scoring the network to connect potential MRs and Salmonella effectors

We used the gene prioritization method, GUILD, to score the nodes of the PIN connecting the predicted MRs and the known Salmonella protein effectors or other Salmonella proteins that could potentially act as effectors. We used the TAP network and analysed the network with the top 30% nodes (nodes with scores ranking among the best 30% of all nodes) to predict MRs and Salmonella neweffectors, and to unveil the most relevant functions by means of the enrichment of GO terms.

3.1.5.3.3.1 Using predicted MRs as seeds

When using the predicted MRs as seeds we search for Salmonella proteins that could act as effectors and, at the same time, if some of the known effectors are among the top ranking scores, then we can confirm the potential of the predicted MRs. Table 5 shows the scores and ranking

of the known effectors that obtained a positive score with GUILD NetCombo approach. Only 4 effectors were found when using the predicted CMRs, but they were not among the top ranking (i.e. D0ZY43 was the best ranked, 2651 out of 6162 proteins). Still, the three best-ranked effectors are among the top 50% of the total ranking of Salmonella proteins. We did also check the 10 best scores of Salmonella proteins, because they could act as potential effectors (these results are also included in Table 5, with the corresponding ranking). This hints some potential protein-effector candidates of Salmonella for further studies.

Salmonella protein	Effector	#Ranking/ #Total	#Ranking/ #Salmonella	NetCombo score
D0ZY43	Yes	2651/6162	241/1196	0.02594
D0ZVQ4	Yes	3170/6162	617/1196	0.024873
D0ZY42	Yes	3187/6162	634/1196	0.024873
D0ZV15	Yes	3807/6162	1025/1196	0.024419
D0ZWZ8	No	241/6162	1/1196	0.102867
D0ZMT5	No	399/6162	2/1196	0.034938
D0ZW18	No	435/6162	3/1196	0.034374
D0ZIR2	No	547/6162	4/1196	0.03383
D0ZR32	No	754/6162	5/1196	0.032507
D0ZQF4	No	786/6162	6/1196	0.032303
D0ZIM1	No	963/6162	7/1196	0.03128
D0ZQ18	No	967/6162	8/1196	0.031266
D0ZNB8	No	1034/6162	9/1196	0.03116
D0ZIQ6	No	1081/6162	10/1196	0.030915

Table 3-5: **GUILD scores and ranking of Salmonella proteins**. Scores of GUILD are calculated with the NetCombo approach (Netcombo scores) using the predicted CMRs as seeds. The second column indicates if the *Salmonella* proteins are known effectors (Y) or not (N). We show the results for *Salmonella* protein effectors with a positive score and the best 10 scores of Salmonella proteins. The third column shows the ranking among the total number of nodes in the PIN and the fourth column the ranking over the total number of Salmonella proteins in the PIN.

3.1.5.3.3.2 Using Salmonella known effectors

We used *Salmonella* known effectors as seeds and tested the scores of TFs of *Arabidopsis*, specifically checking those that were predicted as MRs. The best scores are obtained, as expected, by *Salmonella* proteins (details on the top 10 scored nodes are shown in supplementary Table S12). We found a subset of 17 TFs among top 30% of best-scored nodes. Four out of 17 were CMRs (AT4G31800, which is a CCMRs, AT2G25000, AT5G52830 and AT5G45260), which is a significant enrichment among all the predicted MRs for *Salmonella* WT infection (p-value=0.004) and prgH- infection (p-value=0.01).

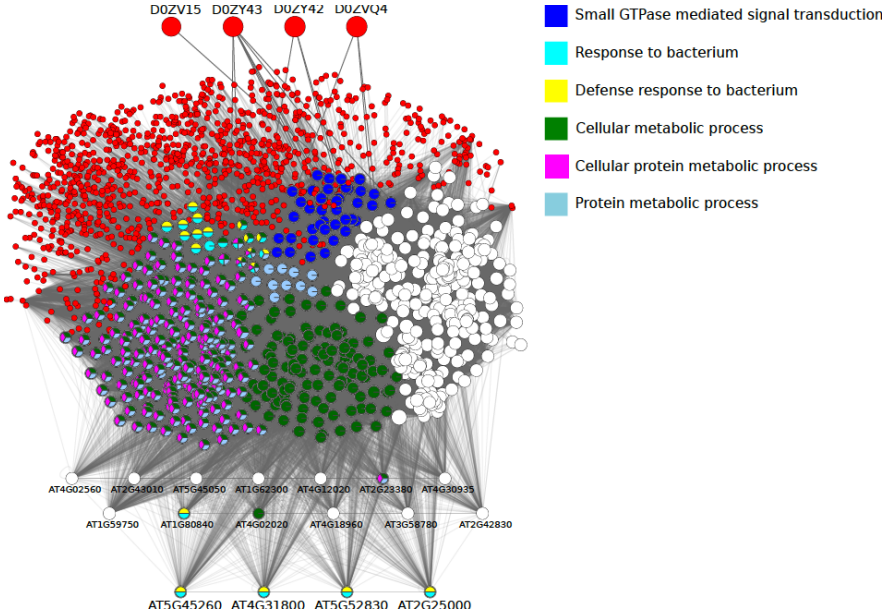


Figure 5: Top scored subnetwork of GUILD scores. Predicted CMRs participating in the network are located at the very bottom of the figure. TFs of Arabidopsis are located at the bottom rows and Salmonella proteins at the top, being at the very top the effectors used as seeds in GUILD. Nodes are coloured according to the most relevant functions as in figure 5. The size of nodes representing Arabidpopsis proteins is proportional to the scores of GUILD (this was not applied on Salmonella proteins in order to help visual inspection).

Figure 5 shows the sub-network of the top 30% nodes, indicating the seeds (known *Salmonella* effectors), the 4 predicted CMRs and other Salmonella proteins and Arabidopsis TFs. We calculated the enrichment of biological processes in this subnetwork using BinGO [366] (as above, see section 3.1), and coloured the nodes associated with the four most relevant terms plus "defense response to bacterium" and "response to bacterium" (enrichment scores are shown in supplementary Table S13). As in our previous analysis of shortest paths, cellular metabolic processes and small GTPase mediated signal transduction were on the top, after neglecting the most general terms, such as "cellular processes".

We further analysed the top 20% of best ranked Arabidopsis TFs (116 TFs) with positive GUILD scores when using the TAP network. Among them, 13 were predicted as MRs, 11 were CMRs and one of the a CCMR (AT4G31800), which was among the best top 10 ranked TFs. Table 6 shows the best scored MRs and their position in the total ranking of the PIN and within the top 20% of TFs with positive scores.

We also studied the results on a more restrictive network, NOTAP, but the best position of a CMRs was ranked 22 (AT5G28650) and only 8 of the predicted MRs were found in the top 20% subnetwork (results are included in the supplementary table S14. Probably, not only the lack of known experimental interactions between *Salmonella* and *Arabidopsis* affects the message-passing in the NOTAP network, but also restricting mostly to yeast-two hybrid known interactions entails the lost of many of the interactions of TFs and consequently the lost of score-transfer through its connections.

Node	MR	GUILD score	Ranking/Total	Ranking/TFs
AT4G31800	CCMR	0.395464	1786/6162	9/116
AT2G25000	CMR	0.395464	1787/6162	10/116
AT5G52830	CMR	0.395462	1790/6162	12/116
AT5G45260	CMR	0.395429	1832/6162	17/116
AT4G16110	WT	0.383318	2497/6162	24/116
AT3G19510	prgH-	0.380545	2665/6162	33/116
AT5G62000	CMR	0.379974	2743/6162	36/116
AT5G22220	prghH-	0.37938	2795/6162	42/116
AT5G59820	WT	0.378787	2842/6162	43/116
AT2G46130	CMR	0.376413	3106/6162	69/116
AT5G28650	CMR	0.376229	3208/6162	88/116
AT2G24570	CMR	0.376229	3214/6162	90/116
AT4G31550	CMR	0.376200	3250/6162	95/116
AT2G30590	CMR	0.376200	3259/6162	98/116
AT5G06960	CMR	0.376173	3301/6162	106/116

Table 3-6: **Predicted MRs among Arabidopsis TFs with best and positive GUILD scores (top 20% of TFs)**. We calculated the NetCombo GUILD scores using Salmonella effectors as seeds and the TAP network as the underlying PIN. In the second column (MR) we indicate if the TF was predicted for WT or prgH- infection, as CMR or CCMR. The third column shows the score and the next two columns show the ranking with respect to the total number of nodes in the network and the relative ranking with respect to the 20% of the total number of TFs in the PIN.

In Figure 6 we show the subnetwork of the shortest path between the seeds (known *Salmonella* effectors) and the predicted MRs found among the 116 TFs best scored (top 20%). As before, we calculated the enrichment of biological processes in this subnetwork and the nodes associated with the four most relevant terms plus "defense response to bacterium" and "response to bacterium" (for enrichments scores see supplementary Table S15). Interestingly, the top biological processes were the response to stimulus, in particular response to inorganic, chemical and metal substances, which were also significantly enriched in previous analysis. Furthermore, as expected, in all networks connecting *Salmonella*

effectors and selected *Arabidopsis* TFs, the response to bacterium and defence response to bacterium were significantly enriched.

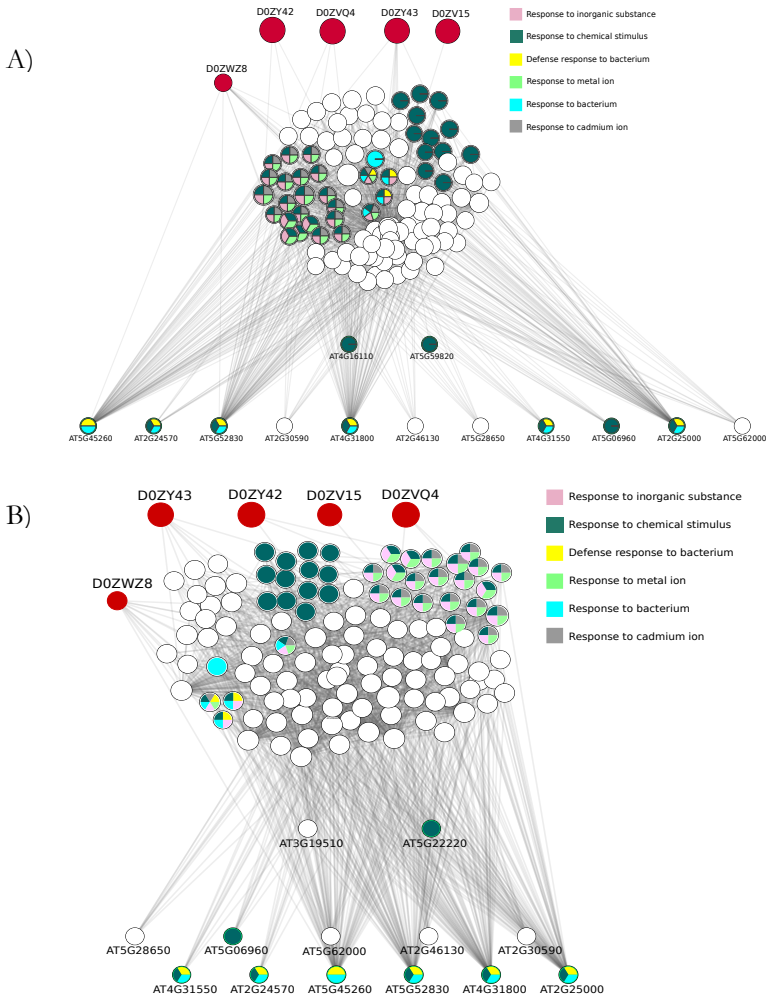


Figure 6: Network of SPs between Salmonella effectors and predicted MRs among the top 20% Arabidopsis TFs ranked with GUILD. Subnetworks of shortest paths on Salmonella WT (A) and prgH- (B) infections. Predicted MRs are shown at the bottom of the figure and Salmonella proteins and effectors at the top. Nodes are coloured according to the enriched GO term biological processes and the size of the nodes representing Arabidopsis proteins is proportional to the GUILD scores (as in figure 5).

3.1.5.4 Integration of approaches and validation of key players on the system response to infection

3.1.5.4.1 Unveiling potential key players of the defence system of Arabidopsis

TF	Clustered		SP plasma membrane		SP Effector		SP Not effector		Remote homology	Diff 30'	SDREM		GUILD	
	WT	PrgH-	WT	PrgH-	WT	PrgH-	WT	PrgH-		WT	WT	PrgH-	WT	PrgH-
AT4G31800 (WRKY18)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓
AT2G25000 (WRKY60)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗	✓	✓
AT5G59820 (ZAT12)	✗	✓	✗	✓	✗	✓	✗	✓	✓	✓	✗	✓	✗	✗
AT1G80840 (WRKY40)	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✓	✓
AT4G23810 (WRKY53)	✓	✓	✓	✓	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗
AT5G49520 (WRKY48)	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
AT5G47370 (HAT2)	✓	✓	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗

Table 3-7: **Integrated results for the prediction of CMRs of Arabidopsis**. A Tick indicates that the TF was predicted as MR, a cross indicates that either the TF was not predicted as MR or it did not fulfil the requirement of the column. Main columns are then split in results for Salmonella WT and prgH- mutant form infections. The first column (TF) shows the name and TAIR code of Arabidopsis TFs. The second column (clustered) shows if the TF belongs to some of the clusters obtained with STEM. The third column (SP Plasma membrane) indicates if a plasma membrane is found within a shortest path smaller than 4 steps to the TF. The fourth and fifth columns indicate if a Salmonella protein is found at a shortest path smaller than 4 steps when the Salmonella protein is a known effector (SP effector) or not (SP non-effector). The sixth column indicates if there is one or more similar Salmonella proteins under the less restrictive criteria (see Table S8 in Supplementary Information). The next two columns show if the TF was predicted as MR by SDREM or belonged to the top 20% best scored TFs with GUILD when using Salmonella effectors as seeds (GUILD). The last column (30') shows if there is a differential expression of the TF after 30' of infection by Salmonella WT.

In Table 7 we summarize the results of a selected set of TFs of Arabidopsis that could act as MRs, with particular attention to CCMRs and CMRs. For each of the candidates we outline the approach taken for the prediction: shortest paths connecting them with Salmonella effectors or potential actors in the plasma membrane, or the scoring obtained with SDREM and GUILD. Therefore, we focused our attention on WRKY18 (AT4G31800) and WRKY60 (AT2G25000). The two TFs are predicted as MR for both prgH- mutant and WT forms, they are in a radius shorter then 4 steps from plasma membrane proteins and also from Salmonella proteins (specifically from known effectors). Consequently, they also got a positive score with GUILD, being included in the top 30% nodes with best scores and among the top scoring TFs. Additionally, WRKY18 and WRKY60 were predicted by SDREM as MRs on the response to Salmonella prgH- mutant form and WT infections, respectively. According to our model, WRKY18 is in cluster 39 and it regulates the profiles 44 of both infections. Also, WRKY60 is in cluster 39 and it regulates profile 43 in WT infection and cluster 41 in prgH- infection. Consequently, we expect the knockout of these two TFs of Arabidopsis to affect the profiles 43 and 44 of Salmonella WT and prgH- infections. Both TFs act together with (WRKY40) AT1G80840 inhibiting the expression of AT2G36270 (ABI5) and/or AT2G40220 (ABI4). Hence they are involved in the regulation of the phytohormone abscisic acid (ABA) which is a signalling path known to respond to environmental stress and plant pathogens [368]. Besides, AT4G31800 (WRKY18) and AT1G80840 (WRKY40) play a significant role in resistance to S. littoralis herbivory[57]. WRKY40 was not predicted as MR but it was found within clusters 42 (infection by prgH-) and 41 (infection by WT) and also in shortest path distance from Salmonella effector proteins and within the top 20% best scored TFs. There is a cooperative behaviour of AT1G80840 (WRKY40), AT4G31800 (WRKY18) and AT2G25000 (WRKY60) in biological

processes associated with stress functions, that were significantly enriched in the sub-networks connecting Salmonella effectors and the predicted MRs. This cooperative behaviour was recently shown: AT4G31800 (WRKY18) and AT2G25000 (WRKY60) act as weak transcriptional activators while AT1G80840 (WRKY40) is a transcriptional repressor modulating gene expression under stress [369]. Therefore, we concluded to knock-out these three genes and test the effect on the regulation of expression of genes during Salmonella infection.

3.1.5.4.2 qPCR validation of the potential players

In order to validate the predictions we have obtained a mutant of Arabidopsis (hereafter named 3KO mutant) with the knock-out of WRKY18 (AT4G31800), WRKY40 (AT1G80840) and WRKY60 (AT2G25000) and have checked the expression of five Arabidopsis genes under infection by Salmonella WT and prgH-: WRKY33 (AT2G38470), TET8 (AT2G23810), NUDT7 (AT4G12720), TIR-NBS (AT1G66090) and NHL3 (AT5G06320). We have analysed by qPCR the levels of expression of these genes at 2 and 24 hours after infection (2h and 24h, time points, respectively). We used three replicas to estimate the noise on the comparison of expression levels (usually due to experimental deviations). After 24 hours of inoculation of Salmonella, the changes of expression of the genes under test were too small and the results were less consistent between the different replicas. We presume that the response after 24 hours of inoculation of Salmonella has already been produced and many of the genes involved in the starting defence of Arabidopsis have lost their activity. Consequently, we only use the 24h time point of this analysis to corroborate that the levels of expression are diminished and the genes affected correspond to the starting response of Arabidopsis defence.

The five genes selected for the test belong to clusters 43 and 44, but only TIR-NBS and NHL3 are shared by the profiles of infection of Salmonella WT and prgH- (TIR-NBS in profile 43 of WT infection and in profile 44 of prgH- infection, while NHL3 is in profile 44 of both WT and prgH-). For the other genes, WRKY33 and TET8 have a specific pattern of expression only for the response to Salmonella WT (i.e. WRKY33 is in profile 43 and TET8 in 44), while NUDT7 has a specific pattern of expression only for the infection by Salmonella prgH- (i.e. it belongs to cluster 44)

In Figure 7 we compare the levels of expression of these genes in Arabidopsis wild type (WT) and the 3KO mutant form after 2hours of infection. We corroborate that the expression levels of WRKY33 and TET8 are not affected under the infection by Salmonella prgH-, while after inoculation of Salmonella WT its expression in Arabidopsis WT is higher than in 3KO mutant form. This implies that the knockout has changed the levels of response, affecting the profiles of expression of clusters 43 and 44 produced on the inoculation of Salmonella WT. These results are in agreement with the our model of the regulatory network. Furthermore, the expression of NHL3, which is in similar profiles of infection by Salmonella WT and prgH- (merged-code 44), is also affected by the knockout, being its expression diminished with respect to Arabidopsis WT. The changes of expression of TIR-NBS and NUDT7 are almost in agreement with our reasoning. We expected that the expression of NUDT7 would be affected after inoculation of Salmonella prgH-, diminishing the levels of expression in Arabidopsis 3KO mutant form with respect to the WT. This effect is shown in Figure 7 after 2 hours of infection. However, we also saw a decrease in Arabidopsis 3KO mutant form with respect to the WT for the infection with Salmonella WT. Still, this was smaller than the decrease observed for prgH-. Finally, the expression of TIR-NBS at 2h after Salmonella WT infection suffers a

dramatic decrease in Arabidopsis 3KO mutant form with respect to the WT. We would expect a similar decrease for the infection with Salmonella prgH- mutant form, as TIR-NBS also belongs to one of the affected profiles of the prgH- infection, but this was not observed. For all genes, the changes of expression were not observable after 24 hours of exposition to Salmonella, which implies that the knocked-out TFs were involved in the first response of Arabidopsis, as expected.

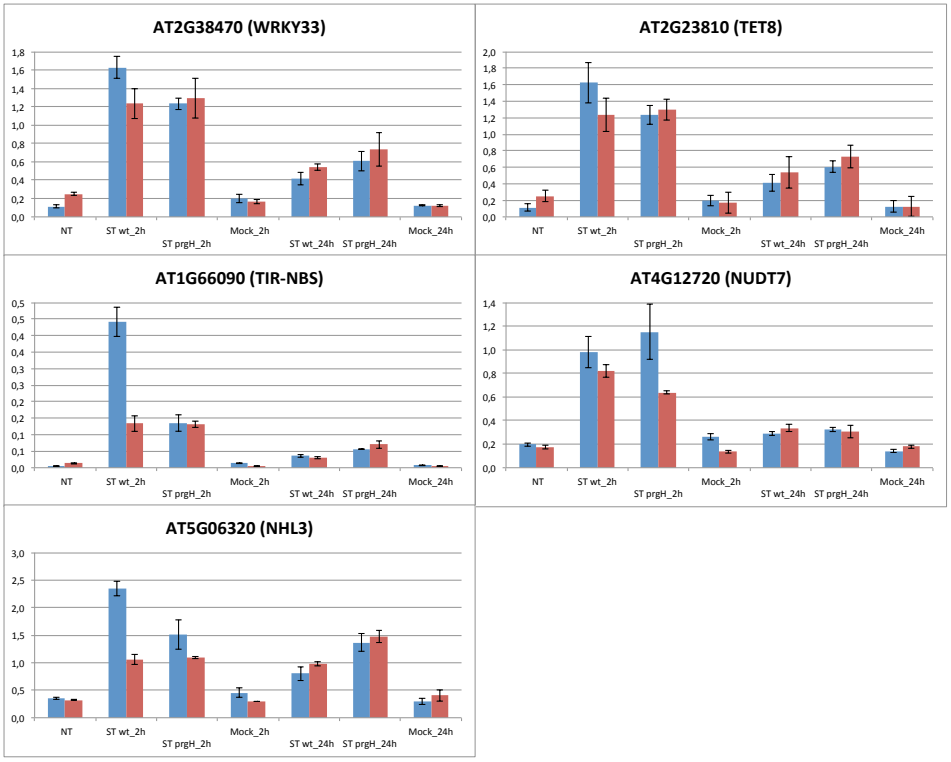


Figure 7: qPCR Experimental test of selected genes of Arabidopsis. We have tested the average levels of expression of WRKY33, TET8, TIR-NBS, NUDT7 and NHL3 in histogram bar-plots. The interval deviation of expression is shown in error-bars. Levels of expression for Arabidopsis wild-type (blue) and triple-mutant knock-out (3KO) of WRKY18, WRKY40 and WRKY60 (red) are shown after 2 hours and 24 hours time-points of infection by

Salmonella WT and prgH- mutant form. For the sake of comparison, Arabidopsis WT and triple mutant form (3KO) were also treated with Mock solution to test the basal effect of the reagent.

3.1.5.4.3 Short time expression changes

We analysed the immediate changes of expression in the first 30' after infection by *Salmonella* WT and prgH-. We found 396 genes differentially expressed: 363 are up and 33 down regulated with respect to the basal state without infection (see the total list in table S16). AT4G31800 (WRKY18) and ZAT12 are the only CMRs among the up-regulated genes, and we also observed other changes in MRs, such as AT5G59820 and AT3G46090 that are among the MRs specific for Salmonella WT infection.

3.1.6 Conclusions and discussion

We presented a multidisciplinary approach that, integrating the analysis of data coming from high-throughput technologies and the information from *in-silico* PPI prediction methods, is capable of identifying key proteins during *Salmonella* infection of its host (*Arabidopsis*). Due to the rapidity of the invasion process our approach has been specifically designed for short time series expression data. The integration step with *in-silico* predictions allowed us to explore the connections of the infection process in both directions: from *Salmonella* effectors to *Arabidopsis* transcription factors and *vice versa*. The approach helped us to unveil the potential role of transcription factors in the regulation of the defence response of Arabidopsis and the mechanisms of activation.

The overlapping among the results obtained with the gene prioritization algorithm and the ones derived from the shortest path approach can be imputable to the usage of the same networks. Besides, the consistency observed with the output of SDREM, that integrates

protein-DNA interaction and our predictions on putative main regulators, increases the potential of our extrapolations.

Finally, the results of two experiments (qPCR of some check-point genes and 30 minutes post-infection microarray) help us to conclude that AT4G31800 and AT2G25000, members of the same WRKYs family, play a crucial role in response to Salmonella infection at its early stages.

To contextualize our findings, it has been demonstrated that Arabidopsis activate the synthesis of the phytohormone jasmonate-isoleucine in response to insect herbivory infection. Jasmonate-isoleucine binds to a complex formed by the receptor COI1 and JAZ repressors. Upon proteasome-mediated JAZ degradation, AT1G32640 (MYC2), AT5G46760 (MYC3), and AT4G17880 (MYC4) become activated and this results in the expression of defence genes. [370]. In the same study it was shown that AT4G31800 (WRKY18), AT1G80840 (WRKY40) and AT5G59820 (ZAT12) play a significant role in resistance to *S. littoralis* herbivory, together with other 6 TFs. Besdies, Shen et al. [371] demonstrated that WRKYs retain the natural ability of the plant to activate stimulus-dependent, PAMP-triggered, defence response genes.

From Brotman et. al [372] we know that an enhancement in the expression of AT4G31800 (WRKY18) and AT1G80840 (WRKY40) suppresses JAZ repressors, thus provoking the negative regulation of the expression of the defence genes AT1G19250 (FMO1), AT3G26830 (PAD3) and AT2G30770 (CYP71A13). This allows the non-pathogenic Arabidopsis root colonization by the *Trichoderma* fungi. Jasmonic Acid (JA) is just one of the phytohormones that is known to play a role during plant response to adverse environmental conditions, together with salicylic acid (SA), ethylene (ET) and abscisic acid (ABA). Nowadays we are aware of the key role of plant WRKY DNA-binding transcription factors for defence responses. Specifically in Arabidopsis a majority of its WRKY

genes are induced by pathogen infection or SA treatment [373]. This happens because most of plant defence, or defence related genes, contain W box sequences in their promoter regions and this allows them to be recognized by the WRKY proteins [374].

As mentioned in results (section 4.1), AT4G31800 (WRKY18) and AT2G25000 (WRKY60), together with AT1G80840 (WRKY40), cooperate in biological processes associated with stress. They inhibit the expression of AT2G36270 (ABI5) and/or AT2G40220 (ABI4) genes, which is consistent with their negative role in ABA signalling [368]. The three WRKY transcription factors antagonize or aid each other in a highly complex manner and they play roles in both plant biotic and abiotic stress responses. These functions were significantly enriched in the subnetworks connecting the predicted MRs and Salmonella effectors studied by SDREM (see results section 3.1). Studies on their expression, DNA binding and transcription-regulating activities have lead to understand that they interact physically with themselves and with each other through a leucine-zipper motif[375]. Specifically AT4G31800 (WRKY18) and AT2G25000 (WRKY60) act as weak transcriptional activators and AT1G80840 (WRKY40) is a transcriptional repressor. Such complex pattern of DNA binding and transcription regulatory activities is the key aspect by which the three WRKYs are known to modulate gene expression in both plant defence and stress responses [369] and this makes the results of our analysis consistent with previous studies.

About Salmonella proteome, we highlighted a subset of known effectors: D0ZY43, D0ZV15, D0ZVQ4 and D0ZY42, and predicted the potential role of some other proteins. Not much is known about the mechanisms by which Salmonella is able to colonize its hosts. We suggest further analysis on the Salmonella proteins: D0ZWZ8 (fumC), involved in fumarate metabolic process; D0ZIR2 (dnaK), involved in protein folding and ATP binding; atpD (D0ZMT5), involved in plasma membrane ATP

hydrolysis coupled proton transport; and gapA (D0ZW18) which takes part in the glucose metabolic process. The functions associated with these proteins were also enriched in the subnetworks analysed that connected the Salmonella effectors with the predicted MRs of Arabidopsis. DnaK is known to be essential for cell survival inside macrophages, thus leading to systemic mechanism of infection [376]. FumC is known to be activated by soxS, thus it is somehow related to oxidative stress response [377]. These proteins were not in the list of effectors, but our analyses suggest they could be implicated in the mechanisms of infection. In conclusion, our multidisciplinary study, focused on the interaction between Arabidopsis and Salmonella, represents a systematic approach to understand the mechanisms by which Salmonella is capable of invading a multitude of different hosts and in particular the triggering response of Arabidopsis.

3.1.7 Bibliography

The Bibliography for this article is at the end of this thesis.

3.1.8 Supplementary Information

3.1.8.1 STEM

The program is intended to apply the mentioned algorithm and to visualize and compare the behavior the genes also across multiple conditions. Each gene is assigned to a model profile, from a set of previously computed outlines to be representative of any possible behavior. To test the null hypothesis that observing a certain value at any time point does not dependent on past and future ones a Bonferroni's corrected permutation test is applied. As distance metric we decided to use the correlation coefficient because the authors proved that, when working with this type of data, two very different profiles cannot be both similar to a third one. Once significant groups have been found, in case they have a minimum correlation of 0.75, they are grouped together. Such method is proved by its authors to perform better, in terms of Gene Ontology (GO) enrichment of the clusters, with respect to k-means and CAGED (Ramoni et al., 2002) algorithms. In our analysis we included only GO terms at maximum level 3 in the hierarchy and to have a corrected p-value for the multiple hypothesis testing for the actual size base enrichment using 500 randomized samples.

3.1.8.2 Common main regulators (CMRs) and correlated clusters common main regulators (CCCMRs)

After calculating the predicted main regulators of the clusters we found a set of 52 proteins in common between the two infections and we called them common main regulators (CMRs). Here we list the set of 52 CMRs:

AT5G49520, AT5G52830, AT2G37590, AT5G65310, AT5G06710, AT5G65230, AT3G62340, AT4G37790, AT4G23550, AT2G46130, AT5G46350, AT5G45260, AT5G56270, AT1G13300, AT2G28340, AT3G25790, AT2G30590, AT2G27050, AT2G46970, AT4G36240, AT4G40060, AT1G80590, AT5G07100, AT4G31550, AT5G45580, AT2G30250, AT5G25810, AT3G32090, AT5G47370, AT2G46680, AT5G22570, AT3G04850, AT5G60850, AT5G62000, AT5G25830, AT5G41570, AT5G43290, AT5G47390, AT5G12840, AT1G69490, AT5G28650, AT4G31800, AT5G06960, AT2G25000, AT4G36740, AT5G23650, AT5G52660, AT5G67580, AT2G24570, AT5G24590, AT5G15130, AT1G71450

Among this set of CMRs there are 9 proteins that have been predicted to be main regulators of the correlated clusters between the two infection. We called them correlated clusters common main regulators (CCCMRs). Here the list of those 9 CCCMRs:

AT5G65310, AT4G31800, AT5G22570, AT3G62340, AT5G41570, AT5G24590, AT5G60850, AT3G32090, AT1G80590

3.1.8.3 Shortest Paths results

We report here the SPs found to connect the predicted common main regulators of the two infections, these include also correlated clusters common main regulators with plasma membrane, using the TAP network.

- WT:

['AT2G25000', 'AT2G18960']
['AT2G25000', 'AT4G30190']
['AT5G45260', 'AT2G18960']
['AT5G45260', 'AT4G30190']
['AT4G40060', 'AT5G57050', 'AT2G18960']
['AT4G40060', 'AT5G57050', 'AT4G30190']
['AT4G31800', 'AT2G18960']
['AT4G31800', 'AT4G30190']

- PrgH:

['AT2G25000', 'AT2G18960']
['AT2G25000', 'AT4G30190']
['AT5G45260', 'AT2G18960']
['AT5G45260', 'AT4G30190']
['AT4G40060', 'AT5G57050', 'AT2G18960']
['AT4G40060', 'AT5G57050', 'AT4G30190']
['AT4G31800', 'AT2G18960']
['AT4G31800', 'AT4G30190']

We report here the SPs found to connect the predicted common main regulators of the two infections, these include also correlated clusters common main regulators with plasma membrane, using the NOTAP network.

- WT:

['AT4G31550', 'AT3G22930', 'AT1G26480', 'AT2G18960']
['AT4G31550', 'AT3G22930', 'AT1G26480', 'AT4G30190']
['AT2G24570', 'AT3G22930', 'AT1G26480', 'AT2G18960']
['AT2G24570', 'AT3G22930', 'AT1G26480', 'AT4G30190']
['AT2G30590', 'AT3G22930', 'AT1G26480', 'AT2G18960']
['AT2G30590', 'AT3G22930', 'AT1G26480', 'AT4G30190']
['AT5G06960', ' AT2G41110', 'AT2G42590', 'AT2G18960']
['AT5G06960', ' AT2G41110', 'AT2G42590', 'AT4G30190']
['AT2G46130', 'AT3G22930', 'AT1G26480', 'AT2G18960']
['AT2G46130', 'AT3G22930', 'AT1G26480', 'AT4G30190']
['AT5G28650', 'AT3G22930', 'AT1G26480', 'AT2G18960']
['AT5G28650', 'AT3G22930', 'AT1G26480', 'AT4G30190']

- PrgH:

['AT4G31550', 'AT3G22930', 'AT1G26480', 'AT2G18960']
['AT4G31550', 'AT3G22930', 'AT1G26480', 'AT4G30190']
['AT2G24570', 'AT3G22930', 'AT1G26480', 'AT2G18960']
['AT2G24570', 'AT3G22930', 'AT1G26480', 'AT4G30190']
['AT2G30590', 'AT3G22930', 'AT1G26480', 'AT2G18960']
['AT2G30590', 'AT3G22930', 'AT1G26480', 'AT4G30190']
['AT5G06960', ' AT2G41110', 'AT2G42590', 'AT2G18960']
['AT5G06960', ' AT2G41110', 'AT2G42590', 'AT4G30190']
['AT2G46130', 'AT3G22930', 'AT1G26480', 'AT2G18960']
['AT2G46130', 'AT3G22930', 'AT1G26480', 'AT4G30190']
['AT5G28650', 'AT3G22930', 'AT1G26480', 'AT2G18960']
['AT5G28650', 'AT3G22930', 'AT1G26480', 'AT4G30190']

We report here the SPs found to connect the predicted common main regulators of the two infections, these include also correlated clusters common main regulators, with Salmonella known effectors, using the TAP network.

Effectors found: D0ZY43, D0ZY42, D0ZV15, D0ZVQ4

- WT:

['AT4G31550', 'AT2G27030', 'AT1G10430', 'D0ZY43']
['AT4G31550', 'AT2G27030', 'AT4G37910', 'D0ZY42']
['AT4G31550', 'AT2G27030', 'AT4G35020', 'D0ZV15']
['AT4G31550', 'AT2G27030', 'AT4G37910', 'D0ZVQ4']
['AT4G31800', 'AT5G59370', 'AT1G10430', 'D0ZY43']
['AT4G31800', 'AT5G66280', 'AT4G20360', 'D0ZY42']
['AT4G31800', 'AT3G12580', 'AT4G35020', 'D0ZV15']
['AT4G31800', 'AT5G66280', 'AT4G20360', 'D0ZVQ4']
['AT2G24570', 'AT2G27030', 'AT1G10430', 'D0ZY43']
['AT2G24570', 'AT2G27030', 'AT4G37910', 'D0ZY42']
['AT2G24570', 'AT2G27030', 'AT4G35020', 'D0ZV15']
['AT2G24570', 'AT2G27030', 'AT4G37910', 'D0ZVQ4']
['AT2G30590', 'AT2G27030', 'AT1G10430', 'D0ZY43']
['AT2G30590', 'AT2G27030', 'AT4G37910', 'D0ZY42']
['AT2G30590', 'AT2G27030', 'AT4G35020', 'D0ZV15']
['AT2G30590', 'AT2G27030', 'AT4G37910', 'D0ZVQ4']
['AT5G22570', 'AT4G38130', 'AT1G69960', 'D0ZY43']
['AT5G22570', 'AT4G38130', 'AT4G37910', 'D0ZY42']
['AT5G22570', 'AT4G38130', 'AT4G37910', 'D0ZVQ4']
['AT5G28650', 'AT2G27030', 'AT1G10430', 'D0ZY43']
['AT5G28650', 'AT2G27030', 'AT4G37910', 'D0ZY42']
['AT5G28650', 'AT2G27030', 'AT4G35020', 'D0ZV15']
['AT5G28650', 'AT2G27030', 'AT4G37910', 'D0ZVQ4']
['AT5G62000', 'AT3G21860', 'AT1G10430', 'D0ZY43']

['AT5G62000', 'AT3G21860', 'AT4G37910', 'D0ZY42']
['AT5G62000', 'AT3G21860', 'AT4G35020', 'D0ZV15']
['AT5G62000', 'AT3G21860', 'AT4G37910', 'D0ZVQ4']
['AT5G06960', 'AT2G27030', 'AT1G10430', 'D0ZY43']
['AT5G06960', 'AT2G27030', 'AT4G37910', 'D0ZY42']
['AT5G06960', 'AT2G27030', 'AT4G35020', 'D0ZV15']
['AT5G06960', 'AT2G27030', 'AT4G37910', 'D0ZVQ4']
['AT2G46130', 'AT2G27030', 'AT1G10430', 'D0ZY43']
['AT2G46130', 'AT2G27030', 'AT4G37910', 'D0ZY42']
['AT2G46130', 'AT2G27030', 'AT4G35020', 'D0ZV15']
['AT2G46130', 'AT2G27030', 'AT4G37910', 'D0ZVQ4']
['AT5G45260', 'AT5G59370', 'AT1G10430', 'D0ZY43']
['AT5G45260', 'AT5G66280', 'AT4G20360', 'D0ZY42']
['AT5G45260', 'AT3G12580', 'AT4G35020', 'D0ZV15']
['AT5G45260', 'AT5G66280', 'AT4G20360', 'D0ZVQ4']
['AT2G25000', 'AT5G59370', 'AT1G10430', 'D0ZY43']
['AT2G25000', 'AT5G66280', 'AT4G20360', 'D0ZY42']
['AT2G25000', 'AT3G12580', 'AT4G35020', 'D0ZV15']
['AT2G25000', 'AT5G66280', 'AT4G20360', 'D0ZVQ4']

- PrgH:

['AT4G31550', 'AT2G27030', 'AT1G10430', 'D0ZY43']
['AT4G31550', 'AT2G27030', 'AT4G37910', 'D0ZY42']
['AT4G31550', 'AT2G27030', 'AT4G35020', 'D0ZV15']
['AT4G31550', 'AT2G27030', 'AT4G37910', 'D0ZVQ4']
['AT4G31800', 'AT5G59370', 'AT1G10430', 'D0ZY43']
['AT4G31800', 'AT5G66280', 'AT4G20360', 'D0ZY42']
['AT4G31800', 'AT3G12580', 'AT4G35020', 'D0ZV15']
['AT4G31800', 'AT5G66280', 'AT4G20360', 'D0ZVQ4']
['AT2G24570', 'AT2G27030', 'AT1G10430', 'D0ZY43']
['AT2G24570', 'AT2G27030', 'AT4G37910', 'D0ZY42']
['AT2G24570', 'AT2G27030', 'AT4G35020', 'D0ZV15']
['AT2G24570', 'AT2G27030', 'AT4G37910', 'D0ZVQ4']
['AT2G30590', 'AT2G27030', 'AT1G10430', 'D0ZY43']
['AT2G30590', 'AT2G27030', 'AT4G37910', 'D0ZY42']
['AT2G30590', 'AT2G27030', 'AT4G35020', 'D0ZV15']

['AT2G30590', 'AT2G27030', 'AT4G37910', 'D0ZVQ4']

['AT5G22570', 'AT4G38130', 'AT1G69960', 'D0ZY43']

['AT5G22570', 'AT4G38130', 'AT4G37910', 'D0ZY42']

['AT5G22570', 'AT4G38130', 'AT4G37910', 'D0ZVQ4']

['AT5G28650', 'AT2G27030', 'AT1G10430', 'D0ZY43']

['AT5G28650', 'AT2G27030', 'AT4G37910', 'D0ZY42']

['AT5G28650', 'AT2G27030', 'AT4G35020', 'D0ZV15']

['AT5G28650', 'AT2G27030', 'AT4G37910', 'D0ZVQ4']

['AT5G62000', 'AT3G21860', 'AT1G10430', 'D0ZY43']

['AT5G62000', 'AT3G21860', 'AT4G37910', 'D0ZY42']

['AT5G62000', 'AT3G21860', 'AT4G35020', 'D0ZV15']

['AT5G62000', 'AT3G21860', 'AT4G37910', 'D0ZVQ4']

['AT5G06960', 'AT2G27030', 'AT1G10430', 'D0ZY43']

['AT5G06960', 'AT2G27030', 'AT4G37910', 'D0ZY42']

['AT5G06960', 'AT2G27030', 'AT4G35020', 'D0ZV15']

['AT5G06960', 'AT2G27030', 'AT4G37910', 'D0ZVQ4']

['AT2G46130', 'AT2G27030', 'AT1G10430', 'D0ZY43']

['AT2G46130', 'AT2G27030', 'AT4G37910', 'D0ZY42']

['AT2G46130', 'AT2G27030', 'AT4G35020', 'D0ZV15']

['AT2G46130', 'AT2G27030', 'AT4G37910', 'D0ZVQ4']

['AT5G45260', 'AT5G59370', 'AT1G10430', 'D0ZY43']

['AT5G45260', 'AT5G66280', 'AT4G20360', 'D0ZY42']

['AT5G45260', 'AT3G12580', 'AT4G35020', 'D0ZV15']

['AT5G45260', 'AT5G66280', 'AT4G20360', 'D0ZVQ4']

['AT2G25000', 'AT5G59370', 'AT1G10430', 'D0ZY43']

['AT2G25000', 'AT5G66280', 'AT4G20360', 'D0ZY42']

['AT2G25000', 'AT3G12580', 'AT4G35020', 'D0ZV15']

['AT2G25000', 'AT5G66280', 'AT4G20360', 'D0ZVQ4']

We report here the SPs found to connect the predicted common main regulators of the two infections, these include also correlated clusters common main regulators, with Salmonella known effectors, using the NOTAP network.

Effectors found: D0ZY43, D0ZV15

- WT:

['AT5G62000', 'AT3G21860', 'AT1G10430', 'D0ZY43']

- PrgH:

['AT5G62000', 'AT3G21860', 'AT1G10430', 'D0ZY43']

We report here the SPs found to connect the predicted common main regulators of the two infections, these include also correlated clusters common main regulators, with Salmonella proteins not known to be effectors, using the TAP network.

- WT:

['D0ZWZ8', 'AT5G45260']
['D0ZWZ8', 'AT2G25000']
['D0ZWZ8', 'AT4G31800']

- PrgH:

['D0ZWZ8', 'AT5G45260']
['D0ZWZ8', 'AT2G25000']
['D0ZWZ8', 'AT4G31800']

We report here the SPs found to connect the predicted common main regulators of the two infections, these include also correlated clusters common main regulators, with Salmonella proteins not known to be effectors, using the NOTAP network.

- WT:

['D0ZWE1', 'AT3G46520', 'AT5G63110', 'AT5G22570']
['D0ZL47', 'AT5G28540', 'AT4G38130', 'AT5G22570']
['D0ZL47', 'AT3G12580', 'AT3G21860', 'AT5G62000']
['D0ZXJ7', 'AT3G46520', 'AT5G63110', 'AT5G22570']
['D0ZQW8', 'AT5G59160', 'AT3G21860', 'AT5G62000']
['D0ZXH6', 'AT3G46520', 'AT5G63110', 'AT5G22570']
['D0ZW18', 'AT5G56030', 'AT3G22930', 'AT5G28650']
['D0ZW18', 'AT5G56030', ' AT2G41110', 'AT5G06960']
['D0ZW18', 'AT5G56030', 'AT3G22930', 'AT2G30590']
['D0ZW18', 'AT1G07820', 'AT4G38130', 'AT5G22570']
['D0ZW18', 'AT5G55260', 'AT3G21860', 'AT5G62000']
['D0ZW18', 'AT5G56030', 'AT3G22930', 'AT2G24570']
['D0ZW18', 'AT5G56000', 'AT3G51920', 'AT2G46130']
['D0ZW18', 'AT5G56030', 'AT3G22930', 'AT4G31550']
['D0ZS62', 'AT3G46520', 'AT5G63110', 'AT5G22570']
['MGTC', 'AT3G46520', 'AT5G63110', 'AT5G22570']
['D0ZWJ3', 'AT1G75780', 'AT3G21860', 'AT5G62000']
['D0ZWJ6', 'AT4G38780', 'AT5G63110', 'AT5G22570']
['D0ZQN7', 'AT1G75780', 'AT3G21860', 'AT5G62000']

['D0ZNS4', 'AT3G46520', 'AT5G63110', 'AT5G22570']
['D0ZMT5', 'AT1G26480', 'AT3G22930', 'AT5G28650']
['D0ZMT5', 'AT1G26480', ' AT2G41110', 'AT5G06960']
['D0ZMT5', 'AT1G26480', 'AT3G22930', 'AT2G30590']
['D0ZMT5', 'AT3G46520', 'AT5G63110', 'AT5G22570']
['D0ZMT5', 'AT1G04820', 'AT3G21860', 'AT5G62000']
['D0ZMT5', 'AT1G26480', 'AT3G22930', 'AT2G24570']
['D0ZMT5', 'AT1G26480', 'AT3G22930', 'AT2G46130']
['D0ZMT5', 'AT1G26480', 'AT3G22930', 'AT4G31550']
['D0ZSK0', 'AT1G26480', 'AT3G22930', 'AT5G28650']
['D0ZSK0', 'AT5G65430', 'AT3G22930', 'AT5G06960']
['D0ZSK0', 'AT1G26480', 'AT3G22930', 'AT2G30590']
['D0ZSK0', 'AT1G26480', 'AT3G22930', 'AT2G24570']
['D0ZSK0', 'AT5G65430', 'AT3G22930', 'AT2G46130']
['D0ZSK0', 'AT1G26480', 'AT3G22930', 'AT4G31550']
['D0ZXX7', 'AT5G28540', 'AT4G38130', 'AT5G22570']
['D0ZXX7', 'AT1G09080', 'AT3G21860', 'AT5G62000']
['D0ZR32', 'AT3G50000', 'AT5G37780', 'AT5G28650']
['D0ZR32', 'AT3G50000', 'AT5G37780', 'AT5G06960']
['D0ZR32', 'AT3G50000', 'AT5G37780', 'AT2G30590']
['D0ZR32', 'AT3G50000', 'AT5G63110', 'AT5G22570']
['D0ZR32', 'AT3G50000', 'AT5G37780', 'AT2G24570']
['D0ZR32', 'AT3G50000', 'AT5G37780', 'AT2G46130']
['D0ZR32', 'AT3G50000', 'AT5G37780', 'AT4G31550']
['D0ZWF9', 'AT3G46520', 'AT5G63110', 'AT5G22570']
['D0ZMY4', 'AT3G46520', 'AT5G63110', 'AT5G22570']
['D0ZSF6', 'AT5G59160', 'AT3G21860', 'AT5G62000']
['D0ZX84', 'AT5G28540', 'AT4G38130', 'AT5G22570']
['D0ZX84', 'AT3G12580', 'AT3G21860', 'AT5G62000']
['D0ZUE5', 'AT5G28540', 'AT4G38130', 'AT5G22570']
['D0ZUE5', 'AT3G12580', 'AT3G21860', 'AT5G62000']
['D0ZIR2', 'AT5G56030', 'AT3G22930', 'AT5G28650']
['D0ZIR2', 'AT5G56030', ' AT2G41110', 'AT5G06960']
['D0ZIR2', 'AT5G56030', 'AT3G22930', 'AT2G30590']
['D0ZIR2', 'AT5G28540', 'AT4G38130', 'AT5G22570']
['D0ZIR2', 'AT5G56030', 'AT3G21860', 'AT5G62000']
['D0ZIR2', 'AT5G56030', 'AT3G22930', 'AT2G24570']

['D0ZIR2', 'AT5G56000', 'AT3G51920', 'AT2G46130']

['D0ZIR2', 'AT5G56030', 'AT3G22930', 'AT4G31550']

['D0ZRU7', 'AT4G38130', 'AT5G22570']

['D0ZMH8', 'AT3G46520', 'AT5G63110', 'AT5G22570']

['D0ZSC9', 'AT1G75780', 'AT3G21860', 'AT5G62000']

['D0ZJT9', 'AT5G65430', 'AT3G22930', 'AT5G28650']

['D0ZJT9', 'AT5G65430', 'AT3G22930', 'AT5G06960']

['D0ZJT9', 'AT5G65430', 'AT3G22930', 'AT2G30590']

['D0ZJT9', 'AT5G65430', 'AT3G22930', 'AT2G24570']

['D0ZJT9', 'AT5G65430', 'AT3G22930', 'AT2G46130']

['D0ZJT9', 'AT5G65430', 'AT3G22930', 'AT4G31550']

['D0ZMW7', 'AT3G51260', 'AT3G21860', 'AT5G62000']

['D0ZIM1', 'AT3G19980', 'AT3G21860', 'AT5G62000']

['D0ZVP7', 'AT1G35160', 'AT3G22930', 'AT5G28650']

['D0ZVP7', 'AT1G35160', 'AT3G22930', 'AT5G06960']

['D0ZVP7', 'AT1G35160', 'AT3G22930', 'AT2G30590']

['D0ZVP7', 'AT1G35160', 'AT3G22930', 'AT2G24570']

['D0ZVP7', 'AT1G35160', 'AT3G22930', 'AT2G46130']

['D0ZVP7', 'AT1G35160', 'AT3G22930', 'AT4G31550']

- Prgh:

['D0ZWE1', 'AT3G46520', 'AT5G63110', 'AT5G22570']

['D0ZL47', 'AT5G28540', 'AT4G38130', 'AT5G22570']

['D0ZL47', 'AT3G12580', 'AT3G21860', 'AT5G62000']

['D0ZXJ7', 'AT3G46520', 'AT5G63110', 'AT5G22570']

['D0ZQW8', 'AT5G59160', 'AT3G21860', 'AT5G62000']

['D0ZXH6', 'AT3G46520', 'AT5G63110', 'AT5G22570']

['D0ZW18', 'AT5G56030', 'AT3G22930', 'AT5G28650']

['D0ZW18', 'AT5G56030', ' AT2G41110', 'AT5G06960']

['D0ZW18', 'AT5G56030', 'AT3G22930', 'AT2G30590']

['D0ZW18', 'AT1G07820', 'AT4G38130', 'AT5G22570']

['D0ZW18', 'AT5G55260', 'AT3G21860', 'AT5G62000']

['D0ZW18', 'AT5G56030', 'AT3G22930', 'AT2G24570']

['D0ZW18', 'AT5G56000', 'AT3G51920', 'AT2G46130']

['D0ZW18', 'AT5G56030', 'AT3G22930', 'AT4G31550']

['D0ZS62', 'AT3G46520', 'AT5G63110', 'AT5G22570']
['MGTC', 'AT3G46520', 'AT5G63110', 'AT5G22570']
['D0ZWJ3', 'AT1G75780', 'AT3G21860', 'AT5G62000']
['D0ZWJ6', 'AT4G38780', 'AT5G63110', 'AT5G22570']
['D0ZQN7', 'AT1G75780', 'AT3G21860', 'AT5G62000']
['D0ZNS4', 'AT3G46520', 'AT5G63110', 'AT5G22570']
['D0ZMT5', 'AT1G26480', 'AT3G22930', 'AT5G28650']
['D0ZMT5', 'AT1G26480', ' AT2G41110', 'AT5G06960']
['D0ZMT5', 'AT1G26480', 'AT3G22930', 'AT2G30590']
['D0ZMT5', 'AT3G46520', 'AT5G63110', 'AT5G22570']
['D0ZMT5', 'AT1G04820', 'AT3G21860', 'AT5G62000']
['D0ZMT5', 'AT1G26480', 'AT3G22930', 'AT2G24570']
['D0ZMT5', 'AT1G26480', 'AT3G22930', 'AT2G46130']
['D0ZMT5', 'AT1G26480', 'AT3G22930', 'AT4G31550']
['D0ZSK0', 'AT1G26480', 'AT3G22930', 'AT5G28650']
['D0ZSK0', 'AT5G65430', 'AT3G22930', 'AT5G06960']
['D0ZSK0', 'AT1G26480', 'AT3G22930', 'AT2G30590']
['D0ZSK0', 'AT1G26480', 'AT3G22930', 'AT2G24570']
['D0ZSK0', 'AT5G65430', 'AT3G22930', 'AT2G46130']
['D0ZSK0', 'AT1G26480', 'AT3G22930', 'AT4G31550']
['D0ZXX7', 'AT5G28540', 'AT4G38130', 'AT5G22570']
['D0ZXX7', 'AT1G09080', 'AT3G21860', 'AT5G62000']
['D0ZR32', 'AT3G50000', 'AT5G37780', 'AT5G28650']
['D0ZR32', 'AT3G50000', 'AT5G37780', 'AT5G06960']
['D0ZR32', 'AT3G50000', 'AT5G37780', 'AT2G30590']
['D0ZR32', 'AT3G50000', 'AT5G63110', 'AT5G22570']
['D0ZR32', 'AT3G50000', 'AT5G37780', 'AT2G24570']
['D0ZR32', 'AT3G50000', 'AT5G37780', 'AT2G46130']
['D0ZR32', 'AT3G50000', 'AT5G37780', 'AT4G31550']
['D0ZWF9', 'AT3G46520', 'AT5G63110', 'AT5G22570']
['D0ZMY4', 'AT3G46520', 'AT5G63110', 'AT5G22570']
['D0ZSF6', 'AT5G59160', 'AT3G21860', 'AT5G62000']
['D0ZX84', 'AT5G28540', 'AT4G38130', 'AT5G22570']
['D0ZX84', 'AT3G12580', 'AT3G21860', 'AT5G62000']
['D0ZUE5', 'AT5G28540', 'AT4G38130', 'AT5G22570']
['D0ZUE5', 'AT3G12580', 'AT3G21860', 'AT5G62000']
['D0ZIR2', 'AT5G56030', 'AT3G22930', 'AT5G28650']

['D0ZIR2', 'AT5G56030', ' AT2G41110', 'AT5G06960']

['D0ZIR2', 'AT5G56030', 'AT3G22930', 'AT2G30590']

['D0ZIR2', 'AT5G28540', 'AT4G38130', 'AT5G22570']

['D0ZIR2', 'AT5G56030', 'AT3G21860', 'AT5G62000']

['D0ZIR2', 'AT5G56030', 'AT3G22930', 'AT2G24570']

['D0ZIR2', 'AT5G56000', 'AT3G51920', 'AT2G46130']

['D0ZIR2', 'AT5G56030', 'AT3G22930', 'AT4G31550']

['D0ZRU7', 'AT4G38130', 'AT5G22570']

['D0ZMH8', 'AT3G46520', 'AT5G63110', 'AT5G22570']

['D0ZSC9', 'AT1G75780', 'AT3G21860', 'AT5G62000']

['D0ZJT9', 'AT5G65430', 'AT3G22930', 'AT5G28650']

['D0ZJT9', 'AT5G65430', 'AT3G22930', 'AT5G06960']

['D0ZJT9', 'AT5G65430', 'AT3G22930', 'AT2G30590']

['D0ZJT9', 'AT5G65430', 'AT3G22930', 'AT2G24570']

['D0ZJT9', 'AT5G65430', 'AT3G22930', 'AT2G46130']

['D0ZJT9', 'AT5G65430', 'AT3G22930', 'AT4G31550']

['D0ZMW7', 'AT3G51260', 'AT3G21860', 'AT5G62000']

['D0ZIM1', 'AT3G19980', 'AT3G21860', 'AT5G62000']

['D0ZVP7', 'AT1G35160', 'AT3G22930', 'AT5G28650']

['D0ZVP7', 'AT1G35160', 'AT3G22930', 'AT5G06960']

['D0ZVP7', 'AT1G35160', 'AT3G22930', 'AT2G30590']

['D0ZVP7', 'AT1G35160', 'AT3G22930', 'AT2G24570']

['D0ZVP7', 'AT1G35160', 'AT3G22930', 'AT2G46130']

['D0ZVP7', 'AT1G35160', 'AT3G22930', 'AT4G31550']

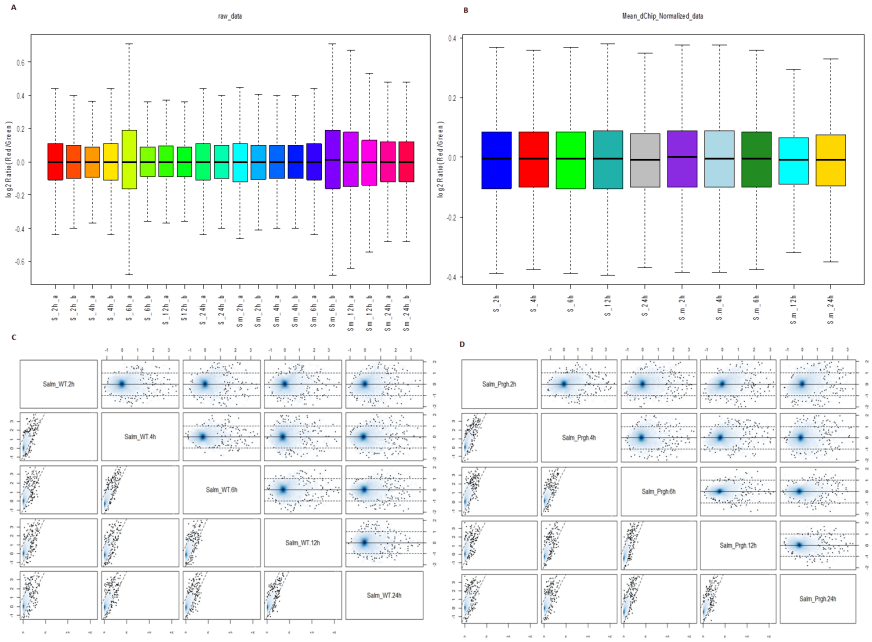


Figure S 3-1: **A) Box plot distribution of raw data of the Microarray CATMA experiment**. Each array is represented indicating the type of infection (S means Salmonella WT, Sm stays for Salmonella prgH mutant), the time point (2h, 4h, 6h, 12h and 24h) and the replica (a and b). **B) Box plot of the Microarray data after dChip normalization (see methods)**. For each time point the mean value between replicas has been calculated. C and D) MAXY plots for WT and prgH- Salmonella infection after dChip normalization, respectively.

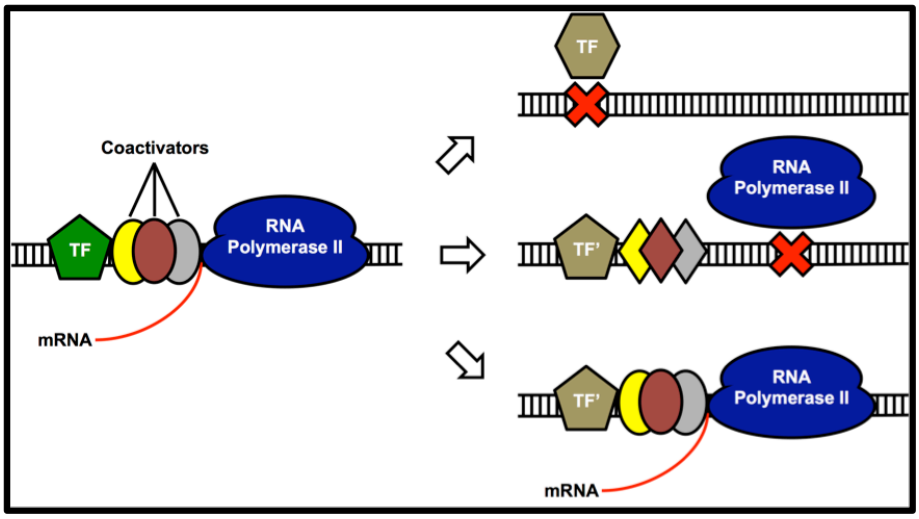


Figure S 3-2: **Representation of the transcription factor (TF) complex formation and substitution by similar proteins**. A protein (TF') may be similar to a known transcription factor (TF), but to produce the translation and bring the RNA polymerase still requires a minimum number of similar interactions, usually with the necessary co-activators. Therefore, similarities between potential transcription factors depend on: global sequence similarity, similar specific DNA binding domains, and a minimum percentage of common interactors.

Table S 3-1: **STEM clustered genes for Salmonella WT infection.**

This table is not included on this book but is available in the CD of this thesis.

Table S 3-2: **Significant GO terms enrichment (p-value < 0.01) of clusters obtained with STEM on Salmonella WT infection.**

This table is not included on this book but is available in the CD of this thesis.

Table S 3-3: **STEM clustered genes for Salmonella prgH- mutant form infection**

This table is not included on this book but is available in the CD of this thesis.

Table S 3-4: **Significant GO terms enrichment (p-value<0.01) of clusters obtained with STEM on Salmonella prgH- infection.**

This table is not included on this book but is available in the CD of this thesis.

Table S 3-5 A: **Predicted Putative MRs of the clusters obtained applying the STEM clustering algorithm on Salmonella WT infections. and prgH-mutant form (B) infections.**

Cluster ID	Putative Main Regulators (MRs)
Profile_17	AT5G23650, AT5G47390, AT5G41570, AT5G58340
Profile_37	AT4G36240, AT4G16110, AT2G01760, AT4G36240, AT2G28340, AT5G25830, AT1G13300, AT3G25790, AT5G47390, AT5G58340
Profile_39	AT5G59820, AT2G37590, AT3G60580, AT5G67450, AT5G60200, AT3G46090, AT5G45580
Profile_42	AT4G16110, AT2G01760, AT2G28340, AT4G36240, AT5G25830, AT1G13300, AT3G25790
Profile_43	AT5G12840, AT5G41570, AT5G52830, AT5G22570, AT5G65230, AT1G71450, AT5G62000, AT3G32090, AT1G80590, AT2G25000, AT5G25810, AT4G25470
Profile_44	AT5G24590, AT3G62340, AT5G41570, AT5G22570, AT4G23550, AT5G65310, AT1G80590, AT4G31800
Profile_45	AT4G28610, AT3G04850, AT5G25790, AT5G60850, AT2G46680, AT5G17320, AT4G37790, AT5G06710, AT5G47370, AT4G40060, AT5G52170, AT3G25790, AT5G45580, AT1G13300, AT5G42630, AT4G36740, AT5G52660, AT5G65310, AT5G24590, AT1G23420, AT2G27050
Profile_47	AT5G43290, AT3G62340, AT5G28650, AT2G24570, AT5G46350, AT5G41570, AT5G45260, AT2G30590, AT5G22570, AT5G07100, AT5G49520, AT5G56270, AT1G69490, AT5G52830, AT4G31550, AT2G30250, AT4G23550, AT3G32090, AT5G67580, AT4G31800, AT1G80590, AT2G46130, AT5G15130, AT5G06960, AT2G25000, AT5G06839, AT5G24590, AT5G24590, AT5G58340
Profile_48	AT2G46970
Profile_49	AT5G51990, AT5G28650, AT5G07100, AT5G56270, AT5G07680, AT2G24570

Table S 3-5 B: **Predicted Putative MRs of the clusters obtained applying the STEM clustering algorithm on *Salmonella* prgH- mutant form infections.**

Cluster ID	Putative Main Regulators (MRs)
Profile_2	AT5G67580, AT5G45580
Profile_11	AT5G18450, AT5G52020, AT1G53230, AT5G62000, AT5G65130, AT5G25810, AT5G21960, AT5G25390, AT5G12840, AT3G15030, AT1G71450, AT5G53950, AT5G67190, AT3G27010, AT5G22220, AT5G62165, AT4G18390
Profile_24	AT5G47390, AT5G65230, AT2G37590, AT5G23650, AT3G10113, AT1G19000, AT1G20710, AT5G45580, AT4G35550, AT5G06710, AT2G46680
Profile_39	AT4G09180, AT5G61270
Profile_41	AT5G65230, AT5G12840, AT4G35550, AT5G52830, AT4G37790, AT5G41570, AT5G22570, AT2G46680, AT3G32090, AT1G80590, AT2G25000, AT5G06710, AT5G47370
Profile_42	AT4G09180, AT5G62610, AT5G12840, AT5G62000
Profile_43	AT5G50915, AT1G28300, AT2G27050, AT2G40620, AT5G67300, AT5G61270, AT5G07100, AT4G36540, AT1G69490, AT2G46130, AT4G34530, AT1G80590, AT5G67110, AT2G46970, AT3G32090, AT5G15130, AT5G56270, AT5G49520, AT2G30250, AT5G45260, AT4G31550, AT4G28815, AT2G24570, AT2G30590, AT5G44080, AT4G09180, AT5G46760
Profile_44	AT2G46680, AT5G06710, AT1G20710, AT5G62000, AT5G47370, AT5G65310, AT4G35550, AT4G31800, AT5G45260, AT4G36740, AT3G62340, AT5G46350, AT5G24590, AT1G69490, AT3G62340, AT5G63790, AT4G40060, AT3G20770, AT5G41570, AT3G32090, AT3G03200, AT5G13180, AT5G22570, AT3G19510
Profile_45	AT5G24590, AT5G47660, AT3G56660, AT5G11510, AT2G24570, AT5G06960, AT5G13180, AT5G28650, AT5G60850, AT5G43290, AT2G30590, AT4G23550, AT3G60030, AT5G07100, AT5G49520, AT5G45260
Profile_46	AT5G67300, AT5G55020, AT3G03200, AT1G13300, AT3G25790, AT3G56660, AT5G45580, AT4G09180, AT5G28770, AT3G61910, AT5G60850, AT5G11510, AT3G10113, AT3G09230, AT5G61430
Profile_48	AT2G37590, AT5G24590
Profile_49	AT2G28340, AT4G36240, AT5G65310, AT5G66320, AT4G40060, AT3G04850, AT5G52660, AT5G25830

Table S 3-6: **List of MRs divided by their type.** The ones predicted to regulate only STEM clusters from Salmonella wt infection are called WT specific. The ones found to regulate only clusters of the prgH- mutant are called prgH- specific. Common MRs (CMRs) are the one predicted to regulate clusters from the two infection and correlated common MRs (CCMRs) are CMRs that regulate similar clusters of the two Salmonella infection.

Type of MRs	List of main regualtors
WT specific	AT5G52170; AT3G60580; AT5G58340; AT5G59820; AT5G51990; AT4G28610; AT2G01760; AT4G25470; AT1G23420; AT5G07680; AT5G67450; AT5G17320; AT5G25790; AT5G60200; AT3G46090; AT4G16110; AT5G42630; AT5G06839
PrgH- specific	AT5G61270; AT1G53230; AT3G09230; AT5G46760; AT3G19510; AT5G63790; AT5G66320; AT2G40620; AT3G61910; AT4G09180; AT5G67300; AT5G52020; AT3G10113; AT4G36540; AT1G28300; AT5G67110; AT3G27010; AT5G22220; AT5G28770; AT5G65130; AT4G35550; AT5G25390; AT5G62610; AT5G67190; AT5G53950; AT5G11510; AT5G62165; AT5G47660; AT5G21960; AT5G50915; AT5G18450; AT5G44080; AT5G13180; AT4G18390; AT4G34530; AT5G61430; AT3G20770; AT4G28815; AT3G56660; AT3G03200; AT5G55020; AT3G15030; AT1G20710; AT3G60030; AT1G19000
Common MRs (CMRs)	AT5G49520; AT5G52830; AT2G37590; AT5G65310; AT5G06710; AT5G65230; AT3G62340; AT4G37790; AT4G23550; AT2G46130; AT5G46350; AT5G45260; AT5G56270; AT1G13300; AT2G28340; AT3G25790; AT2G30590; AT2G27050; AT2G46970; AT4G36240; AT4G40060; AT1G80590; AT5G07100; AT4G31550; AT5G45580; AT2G30250; AT5G25810; AT3G32090; AT5G47370; AT2G46680; AT5G22570; AT3G04850; AT5G60850; AT5G62000; AT5G25830; AT5G41570; AT5G43290; AT5G47390; AT5G12840; AT1G69490; AT5G28650; AT4G31800; AT5G06960; AT2G25000; AT4G36740; AT5G23650; AT5G52660; AT5G67580; AT2G24570; AT5G24590; AT5G15130; AT1G71450;
Correleated Common MRs (CCRMs)	AT5G65310; AT4G31800; AT5G22570; AT3G62340; AT5G41570; AT5G24590; AT5G60850; AT3G32090; AT1G80590

Table S 3-7: **SDREM results:** In **A)** are the results for *Salmonella* **WT infection** and in **B)** *Salmonella* **prgH-** , both using the NOTAP PIN. The column target indicates if the node of the PIN is a TF of Arabidposis (Y) or not (N). We have included the degree of the node in the PIN and the score of SDREM based on the ratio of oriented paths with the highest confidence passing through the node.

A)

Name	Target	Degree	SDREM score
AT5G15850	Y	38	1.000
D0ZY43	N	7	0.895
D0ZV15	N	1	0.105

B)

Name	Target	Degree	SDREM score
AT3G56400	Y	4	0.031
AT5G15840	Y	38	0.079
D0ZV15	N	1	0.072
AT5G59820	Y	97	0.022
D0ZY43	N	7	0.928
AT5G15850	Y	38	0.083
AT4G17750	Y	24	0.683
AT3G46070	Y	97	0.023
AT4G14540	Y	6	0.022
AT3G02380	Y	38	0.082
AT3G02990	Y	4	0.048
AT3G46080	Y	97	0.020
AT5G16820	Y	4	0.048

Table S 3-8: **Low similarity between Arabidopsis TFs and Salmonella proteins.** We limited the results to those CCMRs, CMRs, and WT or prgH specific MRs are similar with any protein of the pathogen according to the sequence criteria (see methods). For each pair we indentify the common Pfam domains, if there are common interactions, the percentage of Sequence identity of the aligned residues (SI) and the percentage of sequence coverage (COV) of aligned residues versus the Arabidopsis TF and the Salmonella query.

	Arabidopsis proteins	Salmonella proteins	Shared Pfam domains	Shared PPI	Homology
CCCMRs	AT4G31800	D0ZQV5	NO	NO	SI:40% COV:4%
CMRs	AT5G45580	D0ZNU3	NO	NO	SI:42% COV:16%
	AT2G27050	D0ZY02	NO	NO	SI:34% COV:18%
	AT2G28340	D0ZLV4	NO	NO	SI:24% COV:14%
	AT4G37790	D0ZIB1	NO	NO	SI:32% COV:41%
	AT2G46970	D0ZM01	NO	NO	SI:38% COV:15%
	AT5G23650	D0ZV29	NO	NO	SI:32% COV:33%
	AT5G43290	D0ZTT0	NO	NO	SI:37% COV:37%
	AT2G25000	SSPH2	NO	NO	SI:34% COV:10%
Prgh MRs	AT5G50915	D0ZMQ3	NO	NO	SI:34% COV:10%
	AT5G50915	D0ZW61	NO	NO	SI:31% COV:31%
	AT5G50915	D0ZW61	NO	NO	SI:31% COV:31%
	AT5G66320	D0ZUX2	NO	NO	SI:33% COV:12%
	AT1G28300	D0ZV29	NO	NO	SI:30% COV:27%
	AT3G61910	D0ZQQ4	NO	NO	SI:32%

					COV:33%
	AT5G25390	D0ZS23	NO	NO	SI:35% COV:26%
	AT3G27010	D0ZJ78	NO	NO	SI:44% COV:8%
	AT4G18390	D0ZX97	NO	NO	SI:45% COV:12%
	AT1G53230	D0ZRV3	NO	NO	SI:39% COV:9%
WT MRs	AT2G01760	D0ZJD2	Response_reg	NO	SI:31% COV:49%
	AT2G01760	D0ZL42	Response_reg	NO	SI:28% COV:87%
	AT2G01760	D0ZL43	Response_reg	NO	SI:34% COV:27%
	AT2G01760	D0ZM78	Response_reg	NO	SI:34% COV:51%
	AT2G01760	D0ZMP4	Response_reg	NO	SI:32% COV:13%
	AT2G01760	D0ZNE3	Response_reg	NO	SI:30% COV:56%
	AT2G01760	D0ZNY7	Response_reg	NO	SI:35% COV:21%
	AT2G01760	D0ZPC8	Response_reg	NO	SI:28% COV:59%
	AT2G01760	D0ZPL1	Response_reg	NO	SI:34% COV:11%
	AT2G01760	D0ZPV9	Response_reg	NO	SI:30% COV:62%
	AT2G01760	D0ZQX8	Response_reg	NO	SI:34% COV:29%
	AT2G01760	D0ZS20	Response_reg	NO	SI:27% COV:52%
	AT2G01760	D0ZS32	Response_reg	NO	SI:27% COV:53%
	AT2G01760	D0ZSP8	Response_reg	NO	SI:30%

					COV:22%
	AT2G01760	D0ZU07	Response_reg	NO	SI:29% COV:53%
	AT2G01760	D0ZV78	Response_reg	NO	SI:33% COV:43%
	AT2G01760	D0ZV90	Response_reg	NO	SI:27% COV:52%
	AT2G01760	D0ZVQ4	Response_reg	NO	SI:28% COV:12%
	AT2G01760	D0ZWR1	Response_reg	NO	SI:33% COV:47%
	AT2G01760	D0ZWR7	Response_reg	NO	SI:34% COV:12%
	AT2G01760	D0ZX87	Response_reg	NO	SI:27% COV:48%
	AT2G01760	D0ZY26	Response_reg	NO	SI:25% COV:18%
	AT4G16110	D0ZJD2	Response_reg	NO	SI:30% COV:51%
	AT4G16110	D0ZL42	Response_reg	NO	SI:32% COV:90%
	AT4G16110	D0ZL43	Response_reg	NO	SI:32% COV:29%
	AT4G16110	D0ZMP4	Response_reg	NO	SI:29% COV:17%
	AT4G16110	D0ZNE3	Response_reg	NO	SI:30% COV:41%
	AT4G16110	D0ZNH4	Response_reg	NO	SI:28% COV:50%
	AT4G16110	D0ZNY7	Response_reg	NO	SI:41% COV:21%
	AT4G16110	D0ZPL1	Response_reg	NO	SI:35% COV:12%
	AT4G16110	D0ZPV9	Response_reg	NO	SI:29% COV:53%
	AT4G16110	D0ZQB7	Response_reg	NO	SI:28%

					COV:51%
	AT4G16110	D0ZQQ5	Response_reg	NO	SI:29% COV:24%
	AT4G16110	D0ZQX8	Response_reg	NO	SI:33% COV:26%
	AT4G16110	D0ZS32	Response_reg	NO	SI:30% COV:48%
	AT4G16110	D0ZSP8	Response_reg	NO	SI:35% COV:24%
	AT4G16110	D0ZU07	Response_reg	NO	SI:25% COV:58%
	AT4G16110	D0ZV68	Response_reg	NO	SI:36% COV:33%
	AT4G16110	D0ZV90	Response_reg	NO	SI:30% COV:52%
	AT4G16110	D0ZVQ4	Response_reg	NO	SI:31% COV:12%
	AT4G16110	D0ZWR7	Response_reg	NO	SI:30% COV:11%
	AT4G16110	D0ZY26	Response_reg	NO	SI:31% COV:14%

Table S 3-9: **Shortest path lengths between CMRs and plasma membrane proteins in the two cross-species PINs (TAP and NOTAP).** "Not in net" means that the CMR is not in the network and "NO Paths" means that there is no connection between the plasma membrane and the CMRs.

		Plasma membrane			
		AT2G18960		AT4G30190	
		TAP	NOTAP	TAP	NOTAP
CCMRs	AT4G31800	1	NO Paths	1	NO Paths
CMRs	AT2G25000	1	NO Paths	1	NO Paths
	AT5G52830	1	Not in net	1	Not in net
	AT5G45260	1	Not in net	1	Not in net
	AT4G40060	2	4	2	4
	AT5G06960	3	3	3	3
	AT2G30590	3	3	3	3
	AT5G28650	3	3	3	3
	AT2G24570	3	3	3	3
	AT2G46130	3	3	3	3
	AT4G31550	3	3	3	3

Table S 3-10: Shortest path lengths between CMRs (two of them are CCMR) and Salmonella proteins in the two cross-species PINs studied (TAP and NOTAP). "Not in net" means that the CMRs is not in the network and "NO Paths" means that there is no connection between the plasma membrane and the CMRs. Not all Salmonella proteins in the table were in the NOTAP network.

		Salmonella Proteins								
		Effectors						Not Effectors		
		D0ZY43		D0ZY42	D0ZV15		D0ZVQ4	D0ZWZ8	D0ZRU7	
		TAP	NOTAP	TAP	TAP	NOTAP	TAP	TAP	TAP	NOTAP
CC MRs	AT4G31800	3	NO Paths	3	3	NO Paths	3	1	3	NO Paths
	AT5G22570	3	4	3	4	4	3	3	2	2
CMRs	AT2G25000	3	NO Paths	3	3	NO Paths	3	1	3	NO Paths
	AT5G52830	3	Not in net	3	3	Not in net	3	1	3	Not in net
	AT5G45260	3	Not in net	3	3	Not in net	3	1	3	Not in net
	AT4G40060	4	6	4	4	6	4	3	4	5
	AT5G06960	3	4	3	3	4	3	3	3	4
	AT2G30590	3	4	3	3	5	3	3	3	4
	AT5G28650	3	4	3	3	5	3	3	3	4
	AT2G24570	3	4	3	3	5	3	3	3	4
	AT2G46130	3	4	3	3	5	3	3	3	4
	AT4G31550	3	4	3	3	5	3	3	3	4
	AT5G62000	3	3	3	3	4	3	3	3	4

Table S 3-11: **Most significant Gene Ontology enrichments of the Arabidopsis genes found in the top scoring paths* after the last iteration of SDREM**. We included the GO terms "response to bacterium" and "defense response to bacterium", which are still significant. We reported the corresponding GO term identifier, its description, and the significance of the hypergeometric test (p-value) after Benjamini-Hochberg false discovery rate correction for multiple testing (corrected p-val). We provide the total number of genes associated with the term with respect to the total number of nodes in the network (Net frequency), the same proportion with respect to the entire Arabidopsis genome (Total genome frequency) and the GO term ranking with respect to the total enriched biological processes.

*Note: The top biological process enriched was neglected (cellular process), because it was a too general description.

SDREM WT

GO-ID	Description	P-val	Corrected p-val	Net frequency	Total genome frequency	Ranking
50896	Response to stimulus	3.7702E-27	1.6476E-24	122/313 (38.9%)	3207/22304 (14.3%)	2/362
44237	Cellular metabolic process	5.3528E-21	1.5595E-18	152/313 (48.6%)	5407/22304 (24.2%)	3/362
42221	Response to chemical stimulus	2.7664E-19	6.0445E-17	75/313 (24.0%)	1710/22304 (7.6%)	4/362
9628	Response to abiotic stimulus	2.7877E-17	4.8729E-15	58/313 (18.5%)	1168/22304 (5.2%)	5/362
9617	Response to bacterium	2.3381E-10	6.3860E-9	20/313 (6.4%)	241/22304 (1.0%)	32/362
42742	Defense response to bacterium	6.3672E-7	6.9562E-6	14/313 (4.5%)	193/22304 (0.8%)	80/362

SDREM PrgH-

GO-ID	Description	P-val	Corrected p-val	Net frequency	Total genome frequancy	Ranking
44237	Cellular metabolic process	5.7675E-58	3.3653E-55	365/704 (51.8%)	5407/22304 (24.2%)	2/461
44267	Cellular protein metabolic process	1.5517E-46	6.0360E-44	230/704 (32.6%)	2767/22304 (12.4%)	3/461
7264	Small GTPase mediated signal transduction	2.4770E-46	7.2266E-44	40/704 (5.6%)	59/22304 (0.2%)	4/461
44238	Primary metabolic process	3.8780E-43	9.0513E-41	348/704 (49.4%)	5719/22304 (25.6%)	5/461
9617	Response to bacterium	2.0472E-7	2.1142E-6	25/704 (3.5%)	241/22304 (1.0%)	113/461
42742	Defense response to bacterium	1.2417E-5	8.5743E-5	19/704 (2.6%)	193/22304 (0.8%)	169/461

Table S 3-12: **Ten best-scored TFs of Arabidopsis with positive GUILD score and top ten scored proteins of the whole PIN (TAP and NOTAP).**

Salmonella total NOTAP		
Gene name	Score	pos/total genes
D0ZW18	1,000000	1/3833
D0ZY00	0.929911	2/3833
D0ZMT5	0.923493	3/3833
D0ZSP7	0.913460	4/3833
D0ZXX9	0.910222	5/3833
D0ZT42	0.910222	6/3833
D0ZQV0	0.910222	7/3833
D0ZLL1	0.910222	8/3833
D0ZIL6	0.910222	9/3833
D0ZIL1	0.910222	10/3833
ATCG00780	0.618335	136/3833
ATCG00065	0.600495	137/3833
AT4G18440	0.520317	139/3833
AT1G36280	0.520317	140/3833
ATCG00820	0.421072	141/3833
AT4G01900	0.417658	142/3833
AT3G57560	0.417658	143/3833
AT4G35830	0.412489	144/3833
AT4G26970	0.412489	145/3833
AT2G05710	0.412489	146/3833

Salmonella total TAP		
Gene name	Score	pos/total genes
D0ZXZ4	1,000000	1/6162
D0ZMT5	0.997038	2/6162
DTPD	0.991079	3/6162
D0ZXT3	0.991079	4/6162
D0ZW73	0.991079	5/6162
D0ZVX3	0.991079	6/6162
D0ZT28	0.991079	7/6162
D0ZT24	0.991079	8/6162
D0ZMQ3	0.991079	9/6162
D0ZW75	0.991078	10/6162
AT4G20360	0.819646	1156/6162
AT4G02930	0.765853	1187/6162
ATCG00065	0.750892	1197/6162
AT4G37910	0.739776	1198/6162
ATCG00820	0.732302	1199/6162
ATCG00780	0.720396	1200/6162
AT4G18440	0.626171	1201/6162
AT1G36280	0.626171	1202/6162
AT5G43940	0.625525	1203/6162
AT3G04120	0.489252	1206/6162

Effectors_NOTAP		
Gene name	Score	pos/total genes
D0ZY43	1,000000	1/3833
D0ZV15	0.978725	2/3833
D0ZPZ6	0.629487	3/3833
AT5G66760	0.603949	4/3833
AT2G18450	0.603949	5/3833
AT1G47420	0.603949	6/3833
AT1G08480	0.603949	7/3833
D0ZQ18	0.543037	8/3833
AT5G40650	0.474168	9/3833
AT3G27380	0.457688	10/3833
AT1G69960	0.425645	11/3833
AT1G59830	0.424713	12/3833
AT1G10430	0.424713	13/3833
AT2G42500	0.422210	14/3833

Effectors_TAP		
Gene name	Score	pos/total genes
D0ZY42	1,000000	1/6162
D0ZVQ4	1,000000	2/6162
D0ZY43	0.953970	3/6162
D0ZV15	0.894847	4/6162
AT4G20360	0.624522	5/6162
AT4G02930	0.602221	6/6162
AT4G37910	0.564660	7/6162
ATCG00065	0.521198	8/6162
AT5G43940	0.504357	9/6162
ATCG00820	0.492953	10/6162
ATCG00780	0.473732	11/6162
AT5G35390	0.470838	14/6162
AT1G79860	0.464640	105/6162
AT3G04120	0.452439	264/6162

Table S 3-13: Ontology terms enrichment of Arabidopsis genes in the subnetwork of best GUILD scores (top 30%) with GUILD using as seeds Salmonella effectors and the TAP PIN. We report the GO term identifier of the top biological processes and the selected GO terms: "response to bacterium" and "defense response to bacterium". We include the functional description and the significance of the hypergeometric test (p-value) after Benjamini-Hochberg false discovery rate correction for multiple testing (corrected p-value). We provide the total number of genes associated with the term with respect to the total number of nodes in the network (Net frequency), the same proportion with respect to the entire Arabidopsis proteome (Total genome frequency) and the GO term ranking with respect to the total of enriched processes.

GO-ID	Description	P-val	Corr p-val	Freq in net	Freq in genome	Ranking
44237	Cellular metabolic process	2.6001E-63	1.4405E-60	364/676 (53.8%)	5407/22304 (24.2%)	2/402
44267	Cellular protein metabolic process	1.4987E-53	5.5351E-51	236/676 (34.9%)	2767/223 (12.4%)	3/402
19538	Protein metabolic process	1.0925E-49	3.0263E-47	247/676 (36.5%)	3147/22304 (14.1%)	4/402
7264	Small GTPase mediated signal transduction	6.4997E-49	1.4403E-46	41/676 (6.0%)	59/22304 (0.2%)	5/402
9617	Response to bacterium	3.6551E-7	3.6160E-6	24/676 (3.5%)	241/22304 (1.0%)	112/402
42742	Defense response to bacterium	1.8358E-6	1.5294E-5	20/676 (2.9%)	193/22304 (0.8%)	133/402

Table S 3-14: **Predicted MRs among the best scored Arabidopsis TFs with GUILD**. GUILD scores were calculated with NetCombo approach using Salmonella effectors as seeds and the NOTAP network. In the second column we indicate if the TF was predicted as WT or prgH- specific MR, CMR or CCMR. The third column shows the GUILD score and the next two columns the ranking with respect to the total number of nodes in the NOTAP network and the relative ranking with respect to the top 20% of TFs.

Node	MR	GUILD score	Ranking/Total	Ranking/ TFs
AT4G16110	WT	0.324083	482/3833	3/106
AT5G59820	WT	0.304402	932/3833	12/106
AT5G28650	CMR	0.302121	1054/3833	22/106
AT4G31550	CMR	0.302121	1064/3833	25/106
AT2G30590	CMR	0.302121	1095/3833	28/106
AT2G24570	CMR	0.302121	1097/3833	29/106
AT5G06960	CMR	0.301971	1113/3833	33/106
AT2G46130	CMR	0.301479	1256/3833	56/106
AT5G62000	CMR	0.297816	1500/3833	102/106
AT5G22570	CCMR	0.296976	1509/3833	103/106

Table S 3-15: **Gene Ontology terms enrichment in the subnetwork of shortest paths between the best-scored (top 20%) Arabidopsis TFs and Salmonella effectors**. Scores were calculated with the Netcombo algorithm using Salmonella effectors as seeds in the underlying TAP network. We selected the top GO terms of biological processes and included the "response to bacterium" and "defense response to bacterium". We report the GO term identifier, its description, the significance of the hypergeometric test (p-value) after Benjamini-Hochberg false discovery rate correction for multiple testing (corrected p-value). We also provide the total number of genes associated with the term with respect to the total number of nodes in the network (Net frequency), the same proportion with respect to the entire Arabidopsis proteome (Total genome frequency) and the GO term ranking with respect to the total enriched processes.

WT

GO-ID	Description	P-val	Corrected p-val	Net frequency	Total genome frequency	Ranking
50896	Response to stimulus	6.3158E-23	2.6084E-20	60/110 (54.5%)	3207/22304 (14.3%)	1/168
42221	Response to chemical stimulus	4.3113E-22	8.9029E-20	45/110 (40.9%)	1710/22304 (7.6%)	2/168
10035	Response to inorganic substance	6.7623E-20	9.3094E-18	25/110 (22.7%)	434/22304 (1.9%)	3/168
10038	Response to metal ion	1.2996E-19	1.3419E-17	23/110 (20.9%)	350/22304 (1.5%)	4/168
9617	Response to bacterium	3.1063E-8	6.1090E-7	11/110 (10.0%)	241/22304 (1.0%)	21/168
42742	Defense response to bacterium	4.9372E-7	7.5521E-6	9/110 (8.1%)	193/22304 (0.8%)	27/168

PrgH-

GO-ID	Description	P-val	Corrected p-val	Net frequency	Total genome frequency	Ranking
50896	Response to stimulus	7.3196E-22	3.1987E-19	60/114 (52.6%)	3207/22304 (14.3%)	1/164
42221	Response to chemical stimulus	2.4554E-21	5.3651E-19	45/114 (39.4%)	1710/22304 (7.6%)	2/164
10035	Response to inorganic substance	1.7209E-19	2.5067E-17	25/114 (21.9%)	434/22304 (1.9%)	3/164
10038	Response to metal ion	3.0658E-19	3.3494E-17	23/114 (20.1%)	350/22304 (1.5%)	4/164
9617	Response to bacterium	4.5139E-8	9.3932E-7	11/114 (9.6%)	241/22304 (1.0%)	21/164
42742	Defense response to bacterium	6.6903E-7	1.0525E-5	9/114 (7.8%)	193/22304 (0.8%)	27/164

Table S 3-16: **Arabidopsis genes 3-fold differentially expressed 30 minutes after inoculation of Salmonella *flg22* mutant**. The comparison has been performed with respect to mock treated plants.

UP regulated

AT3G48640	AT2G05050	AT5G13190	AT4G24310	AT2G32190	AT3G21150	AT1G13480
AT5G46080	AT3G52520	AT3G46080	AT2G22290	AT2G15390	AT1G71520	AT5G41750
AT2G20142	AT2G41640	AT5G52750	AT1G50740	AT3G10930	AT5G11210	AT5G66675
AT1G18740	AT5G01100	AT4G36500	AT3G15518	AT5G43420	AT1G26380	AT3G12910
AT4G21920	AT1G73540	AT3G03030	AT5G25930	AT3G57740	AT5G11140	AT4G23215
AT1G24140	AT4G37780	AT5G26920	AT2G44500	AT4G31950	AT4G16820	AT3G23250
AT1G72920	AT3G23230	AT4G11170	AT1G13340	AT5G41680	AT4G23810	AT3G54150
AT4G13395	AT5G59820	AT5G52760	AT3G23630	AT1G27890	AT5E46700	AT4G39670
AT2E21980	AT1G69900	AT1G32928	AT1G61470	AT4G34150	AT3G60420	AT4G14365
AT1G20510	AT1G66090	AT1G32920	AT3G50800	AT4G23610	AT5G28610	AT1G17147
AT2G46940	AT5G47960	AT5G58680	AT5G52050	AT2G40140	AT3G28340	AT1G30370
AT5G27420	AT5G41740	AT1G69930	AT4G31800	AT4G23220	AT4G11070	AT4G24380
AT5G43620	AT4G23160	AT3G48650	AT1G19020	AT4G23180	AT5G01540	AT1G27730
AT5G46295	AT3G14225	AT2G32020	AT1G23830	AT5G59730	AT5G57510	AT5G38310
AT2G31865	AT5G56960	AT5G47850	AT3G09830	AT5G57010	AT1G18570	AT1G63720
AT3G56710	AT4G14370	AT4G39320	AT1G14480	AT3G08720	AT2G31345	AT2G20562
AT4G11280	AT3G52800	AT5G64870	AT2G37430	AT2G38790	AT3G46620	AT5E08990
AT1G07000	AT4G39640	AT1G68765	AT5G61900	AT2G26190	AT4G34380	AT3G25610
AT1G07160	AT2G40180	AT5G66650	AT5G01550	AT5G64890	AT1G51915	AT3G13600
AT1G59590	AT1G79680	AT2G01180	AT5G64905	AT3G25600	AT2G37820	AT4G19520
AT1G02400	AT3G61190	AT3G56400	AT4G15417	AT1G56250	AT5G35735	AT1G13210
AT2G29110	AT2G35658	AT5G26030	AT3G59080	AT3G18710	AT1G61360	AT1G17240
AT2G20960	AT4G02200	AT3G54420	AT4G22780	AT2G38470	AT4G22030	AT5G17350
AT2G32030	AT3G50930	AT1G70170	AT3G11080	AT1G74360	AT1G27770	AT1G51920
AT2G33710	AT1G57990	AT2G22880	AT1G32720	AT1G58420	AT3G21070	AT1G02360
AT4G37370	AT4G27280	AT2G30020	AT4G18250	AT5G22250	AT3G55840	AT5G66210
AT1G69890	AT2G22500	AT3G09870	AT1G72950	AT3G45640	AT1G20823	AT4G24110
AT2G25460	AT3G50060	AT5G45340	AT2G31945	AT3G62260	AT4G23030	AT1G74450
AT5G46090	AT1G80840	AT2G35980	AT5G46910	AT4G29780	AT3G57530	AT4G08260
AT4G18540	AT4G26090	AT1G56240	AT4G28460	AT3G09520	AT5G18470	AT4G30430
AT5G42380	AT3G17690	AT1G18300	AT5G54490	AT1G17420	AT5G39020	AT3G19615
AT1G22810	AT2G34930	AT1G67880	AT1G28480	AT5G25250	AT3G09020	AT4G39570
AT2G37940	AT1G70740	AT5G36925	AT2G35930	AT1G72940	AT5G57220	AT5G6431
AT2G18210	AT4G14450	AT4G18197	AT1G01560	AT1G74440	AT1G17750	
AT4G38560	AT2G32200	AT4G18195	AT3G27140	AT5G14700	AT5G22690	
AT4G19515	AT5G62150	AT2G04495	AT1G36640	AT1G64610	AT3G11840	
AT1G49000	AT1G61340	AT4G39580	AT5G36920	AT4G18880	AT4G20000	
AT4G28085	AT4G12720	AT1G59865	AT5G44070	AT5G39670	AT1G61560	
AT2G36440	AT4G01010	AT3G29000	AT1G80820	AT1G35230	AT3G07195	
AT2G44840	AT5G12880	AT2G17040	AT1G68450	AT1G78410	AT2G23270	
AT2G47140	AT2G29720	AT1G24147	AT3G25250	AT5G60900	AT2G46620	
AT1G26410	AT3G46930	AT1G24145	AT5G59550	AT2G18670	AT4G01360	
AT5G24110	AT4G40020	AT5G61600	AT1G72900	AT3G01830	AT4G28350	
AT2G39650	AT3G44720	AT2G47550	AT3G10114	AT2G26530	AT1E38580	
AT2G46400	AT5G41550	AT4G20780	AT4G02410	AT5G37490	AT4G37290	
AT1G53080	AT5G04340	AT2G36780	AT1G11050	AT3G02800	AT5G22520	
AT3E13410	AT1G68340	AT5G15870	AT5G47910	AT1G06137	AT3G52430	
AT1G35210	AT3G45960	AT2G24600	AT3G25780	AT1G21326	AT4G21390	
AT1G09940	AT2G03540	AT2G25735	AT4G18170	AT1G06135	AT1G71400	
AT5G22530	AT1G64065	AT3G44260	AT1G43000	AT3G16860	AT2G31990	
AT5G51190	AT1G72910	AT1G28370	AT1G14540	AT3G49530	AT5G47230	
AT5G64660	AT1G72520	AT1G56060	AT1G23710	AT4G11470	AT2G18680	
AT4G24570	AT1G16420	AT5G63130	AT3G02840	AT1G51700	AT1G08105	
AT4G33050	AT1G05575	AT1G75000	AT1G09932	AT5G65600	AT3G26980	
AT1G42990	AT3G57640	AT3G52450	AT5G58120	AT4G01950	AT2G36770	

Down- regulated

AT2G40230, AT5G57780, AT4G38825, AT1G78170, AT4G38860, AT1G29440, AT2G21210, AT3G59940, AT2G25200, AT1G49200, AT4G38840, AT1G49220, AT4G34770, AT1G26920, AT1G50040, AT3G25717, AT1G31173, AT5G18080, AT4G34760, AT5G61590, AT5G54145, AT5G56550, AT2G21220, AT1G76220, AT2G44130, AT3G10120, AT1G29490, AT5G01740, AT4G10910, AT2G42870, AT4G34750, AT1G15670, AT5G67480

4 SALMONELLA INFECTION IN HUMAN

In this section is presented an analysis of salmonella-infected human data. As in the previous chapter, from the clustering of time series microarray data a set of MRs is computationally derived. Human- salmonella cross-species protein interaction network is inferred and used for the analysis of those shortest paths, between known salmonella effectors and the predicted MRs, that contain any plasma membrane protein. The hypothesis of a therapy targeting the predicted regulators based on a drug-specific genetic signature is then investigated and its results are shown to corroborate the MRs predictions.

Results presented in this chapter will be published together with the software implemented. This manuscript is in preparation.

4.1 Unravelling signalling pathways involved in human response to salmonella infection leads to gene-specific drug targeting

Daniel Poglayen[1], Oriol Fornes[1], Jascha Casadio[1], Javier Garcia-Garcia[1], Guy Zinman[2,#a], Ziv Bar-Joseph[2], Judith Klein- Seetharaman[3,#c], Baldo Oliva[1*]

[1] Structural Bioinformatics Laboratory, Universidad Pompeu Fabra, Barcelona, Catalonia, Spain

[2] System Biology Group, Carnagie Mellon University, Pittsburgh, Pennsylvania, United States of America

[3] Department of Computational Biology – School of medicine, University of Pittsburgh, Pittsburgh, Pennsylviania, United States of America

[#a] Current Address: Healthcare, SparkBeyond, Pittsburgh, Pennsylviania, United States of America

[#c] Current Address: Metabolic and Vascular Health, University of Warwick, Coventry, United Kingdom

* Corresponding author

E-mail: baldo.oliva@upf.edu

4.1.1 Abstract

Salmonellae are gram-negative bacterial pathogens capable of infecting a wide range of organisms, including human. To ensure its survival and proliferation within the host, the bacterium secretes effector proteins. However the roles of some of them and a complete picture of which mechanisms are activated in the human immune system still are unknown. This capacity to infect a wide range of hosts, together with its exceptional antimicrobial resistance have meant that, despite the overall improvement of health and sanitary conditions, Salmonellosis, nowadays, is the most frequent food-borne diseases.

In this study, we took a system-wide approach to grasp which mechanisms are activated upon salmonella infection in human. We integrated the analysis of high-throughput data with computationally predicted cross-specie protein-protein interactions to identify the key genes/proteins involved in the response to the infection. Based on time-series microarray data of human infected samples, we clustered genes with similar expression profiles and predicted potential transcription factors that could regulate the expression of the genes in each cluster. Subsequently we analysed by means of shortest topological distance between known effectors and our predicted regulators, which paths can be involved during the infectious process. Finally we tested the hypothesis of a pharmaceutical therapy of the infection using gene-specific drug signatures.

The results we retrieved show, among the gene-specific top ranking drugs derived from our predictions on the clusters' regulators, antimicrobials. This corroborates the potential of our approach to be combined with gene-specific drug targeting.

4.1.2 Introduction

According to the world health organisation, Salmonellosis is the most frequent food-borne disease with around 1,5 billion infections world-wide yearly [320]. Disease in mammals usually occurs by oral ingestion of contaminated food or water. Systemic infection of animals and humans depends on the ability of the bacteria to survive the harsh conditions of the gastric tract before entering intestinal epithelial and subsequently other host cells. After entering the small intestine, Salmonella traverses the intestinal mucous layer and can invade non-phagocytic enterocytes of the intestinal epithelium by bacterial-mediated endocytosis. Once the epithelial barrier has been breached, Salmonella can enter intestinal macrophages, sensing the phagosomal environment and activating various virulence mechanisms in order to survive in the microbicidal environment of the host cells.

Salmonella replicates within host cells in a membrane-bound compartment, the Salmonella-containing vacuoles (SCVs). Intravacuolar bacterial replication depends on tightly controlled interactions with host cell vesicular compartments. Salmonella type III secretion (T3SS) effector proteins subvert trafficking events and alter vacuole positioning by acting on host cell actin filaments, microtubule motors and components of the Golgi complex [378]. Salmonella replicates in SCVs in both nonphagocytic epithelial cells and macrophages by recruiting actin filaments (F-actin) and microtubule-dependent motors to migrate to the perinuclear region, where they intercept secretory traffic from the Golgi apparatus [378]. Once positioned, maturation is stalled and bacterial replication is initiated. Salmonella encodes two distinct T3SS on chromosomal pathogenicity islands 1 and 2 (SPI1 and SPI2). Among 13 identified SPI1 factors, at least six coordinately trigger actin cytoskeletal rearrangements to force bacterial internalization into nonphagocytic cells [379]. The other SPI1 factors are

mostly involved in modifying signalling processes with indirect consequences on the host cell cytoskeleton. SPI2 effectors act subsequently in both epithelial cells and macrophages to promote intracellular replication and systemic spread [380]. Among the 19 currently known SPI-2 effectors, several interact with microtubules and microtubule-associated motors such as kinesin and dynein [381]. However, taking a system-wide view and determining the network of interactions between these proteins and the host proteins, is critical to grasp the mechanisms of host-pathogen response, is in its infancy.

The identification of global networks of protein-protein interactions has been accelerated by the development of new high throughput technologies such as two-hybrid assays [382] and affinity purifications followed by mass spectrometry [383]. Thus, a vast amount of protein-protein interaction data has been collected for a number of different organisms, deposited in multiple repositories and codified using various nomenclature. Protein interaction networks are a useful tool for better understanding the biology of the cell, [126], [384] and characterizing diseases [152], [264]. The topology of networks and the neighbourhood of a of a given protein within a network [385] has been used to functionally characterize proteins [386] and their role in human diseases [387]. It is therefore expected that the use of system-wide approaches to study infectious diseases, and thus the protein interaction networks mediating the communication between pathogen and host, will yield new approaches to design target the pathogens. Thus, the relationship between a pathogen and its host has been studied by means of the common proteins in their signal transduction and metabolic pathways [283], [388]–[390].

In this work we focus on Salmonella infection in human, we take a system wide approach using predictions of interactions based on interologs (when there are known interactions of similar proteins) by including interactions obtained by Tandem Affinity Purification (TAP) methods. We

analysed the biological pathways involved in Salmonella infection of human using Shortest Paths (SPs) analysis and, based on the idea that genes with similar behaviour can have one, or more, common regulator(s), we used a protocol for prediction of Main Regulators (MRs) that integrates ChIP data. With this approach we address the inter-specie mechanisms that allow the bacteria to hide and thrive inside human cells. On the other side, although we are aware that the increase of Salmonella infections in the last decade is due its antimicrobial resistance, we address the hypothesis of a pharmaceutical therapy of the infection using gene specific drug signatures.

4.1.3 Materials and methods

4.1.3.1 Microarray data

We used microarray data on the response of 21 days monolayer cultured Human HT-29 cells to *Salmonella* wild-type infection. The data comprises three replicates, on Illumina Multiplex BeadArray Assays [98] harvested at 15', 30', 1, *2*, 4, 6, 8 and 24 hours after *Salmonella* infection (at OD600 nM= 0,2). Additionally,the assays were set up on different days, using different cells grown from different stocks. For the sake of comparison, we further applied the invariant normalization method [353] contained in the DNA-Chip Analyzer software [103], [354] using as baseline the sample with median overall intensity. The resulting box-plots confirmed that the normalization step smoothed any differences among the different samples (Figure 1).

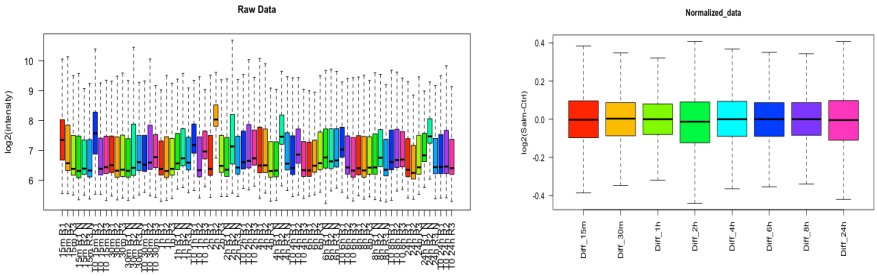


Figure 4-1: Box plots pre (left) and post (right) invariant normalization of Salmonella infected human microarray data.

4.1.3.2 Clustering of genes with similar profiles

We clustered human genes according to the similarity of their time expression profiles upon Salmonella infection on Illumina Multiplex BeadArray Assays (see before). For a global view of time-series data, we applied the Short Time-series Expression Miner (STEM) algorithm [125], which is specifically designed for clustering genes looking at their expression profiles derived from microarray experiments with a few time points (~8 time points or fewer). We computed Gene Ontology (GO) enrichments of the clusters using the default STEM parameters.

4.1.3.3 Prediction of Main Regulators

For each cluster obtained with STEM (see before), we retrieved the promoters of all of its genes from the Eukaryotic Promoter Database [391]. Then, with the DISPOM program [82], we extracted a putative binding-site motif common to the maximum number of genes within the cluster. We used the promoters of the genes in the remaining clusters as background. Potential binding motives for each cluster were reported by means of position-weight matrices (PWM) if they reached a p-value smaller than 10^{-4}. Then, with TReg comparator [215], we searched for matches between the

retrieved PWMs and the PWMs available for human TFs. We used a dissimilarity score of 0.9 to accept the TFs as potential main regulators (MRs) of the cluster. Finally, we used these MRs to find existing chemical compounds that are known to affect their expression from the cmap database (v.02) [314]. This database provides a total ranking of almost 22,000 Uniprot accession of genes, according to their differential expression profiles when treated with different bioactive small molecules.

4.1.3.4 Cross-species network

To infer the complete human-Salmonella PPI network (PIN), we used the server BIPS[295]. We set the conditions of sequence similarity as follows: maximum blast e-value threshold 0.001, percentage of identical residues limited to 60%, 80% minimum coverage between Salmonella-query and human-template sequences. We applied the "matrix" model for co-complex methods, such as tandem affinity purification.

4.1.3.5 Detection of potential Human TFs among Salmonella proteins.

We tested the hypothesis that a Salmonella protein could act directly as a host TF. We checked potential homologs between Human TFs and *Salmonella* proteins on the basis that two TFs are more likely to bind (consequently promoting the transcription of the same set of genes) if their sequences are highly similar and have common PPIs [237]. For the criterion of sequence similarity we used Rost's sequence identity curve of the twilight-zone [362]. Additionally, we tested a second filter if they shared at least one DNA-binding domain from Pfam [285]. Finally, we used a third filter if they shared at least one common interactor.

We used the sequences of the 1,624 human TFs and the *Salmonella* proteome from UniProt [284]. For each host TF, we performed a BLAST

[363] search against *Salmonella* and identified all hits according to Rost's sequence identity curve [362]. For the Pfam-based orthology, the sequences of both human and *Salmonella* proteins were scanned against Pfam [285] using HMMER (version 3.0) [364]. We only considered hits over the HMMER inclusion threshold involving Pfam domains classified as DNA-binding domains.

4.1.3.6 Shortest paths analysis

We obtained shortest paths (SPs) smaller than 4 steps between effectors and MRs using NetworkX[361]. Among these paths we focused on those containing human plasma membrane proteins in order to restrict the analysis to a small amount of shortest paths involving the first contact with the Salmonella-containing vacuoles (SCVs).

4.1.3.7 Gene-specific drug signature Shortest paths analysis

Given the ability of Salmonella to hide inside host cells, we addressed this problem by searching any pharmaceutical compound with the potential to affect the observed gene expression. We tackled this by searching in the connectivity map database (cmap v.02) [314] which drugs are reported to affect the predicted MRs. This database provides a total ranking of almost 22,000 Uniprot accession of genes, according to their differential expression profiles when treated with different bioactive small molecules. Our approach, depicted in Figure 2, searches for the MRs in the cmap gene ranking and extracts the drugs that are reported to have more effect on the expression of the MRs. We thus obtain a ranking of drugs according to their capacity to cause changes in the expression of the TFs that, in turn, we predicted to control the expression of the genes in the clusters.

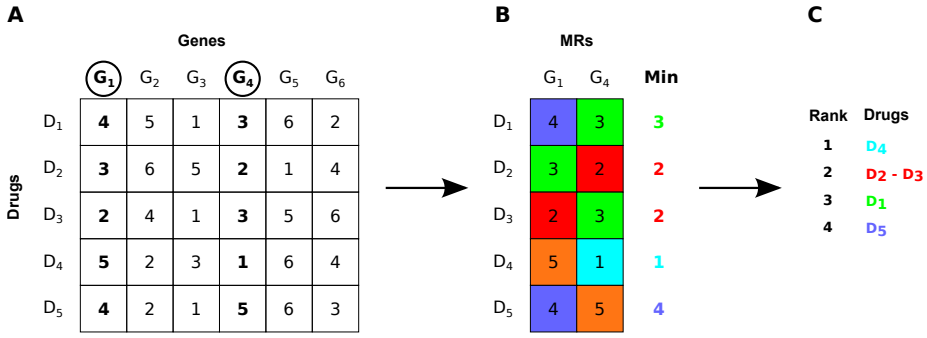


Figure 4-2: Toy example of MRs specific drugs ranking. A) In the cmap ranking of the genes according to their drug specific differential expression profiles we search for our predicted MRs. B) We extract the minimum ranking value for each drug, in other words we search which drugs are reported to affect more the expression of each MR. C) We rank drugs in descending order according to the minimum value found earlier.

4.1.4 Results

4.1.4.1 Time series analysis of gene expression

We used microarray data on the human response to *Salmonella typhymurium* WT strain infections from Multiplex BeadArray Assays. First, genes without sufficient response were neglected, setting the threshold of minimum absolute expression change to 1. Second, we applied STEM to cluster genes with similar profiles. We obtained 4 significant profiles containing 121 genes (Table 1), this is 0.35% of the total number of genes in the array.

Profile 6	ANXA4, C12ORF23, C13ORF7, FLJ43681, FNIP2, G3BP1, KRT20, LOC100129657, LOC100131261, LOC100131572, LOC100132863, LOC646909, LOC647673, MRPL1, PTGES3, RPL7
Profile 10	CDCA5, CTPS, DSCC1, FUT4, HMGCS2, HSPE1, KHK, KIFC1, LOC731314, METTL1, MOSC1, MYB, RAB7B, SIGMAR1, SRM
Profile 37	CXCL1, CYP51A1, EPS8, FOSB, FOXA3, IL8, LOC100134504, LOC220433, LOC286512, LOC641848, LOC642989, LOC646527, LOC730255, RN7SK, TSC22D1, ZFP36L1
Profile 39	AARS, ADM, ADM2, ASNS, ATF3, AXUD1, BCL3, BIRC3, C6ORF223, CCL20, CEBPG, CHAC1, CLK1, CTH, DDIT3, DDIT4, DUSP1, DUSP5, EFNA1, ENO2, ERN1, ERO1L, FAM129A, FTH1, GDF15, GTPBP2, HBEGF, IFITM3, IFRD1, IL15, IL1RAP, IRAK2, IRF1, IRF7, JMJD1A, LARP6, LCN2, LINCR, LOC730256, LTB, MT1X, MUC1, NFIL3, NFKB2, NFKBIA, NUPR1, P8, PCK2, PHGDH, PI3, PIM1, PPP1R15A, PSAT1, PTGS2, RBCK1, RELB, S100A3, S100P, SARS, SAT1, SBNO2, SDC4, SERPINA3, SLC3A2, SLC7A5, SLCO4A1, SMOX, SNORD48, SPIRE1, TAP1, TGM2, TRIB3, ULBP1, VEGFA

Table 4-1: Genes clustered according to the similarity of their expression profiles.

4.1.4.2 Prediction of Main-Regulators (MRs)

We were able to predict a total of 43 putative MRs for 3 clusters (encoded 6, 39 and 10) when using the approach described in methods (see Table 4-2). For cluster "37" the prediction failed and we were unable to assign any predicted TF acting as MR.

Profile 39	ZIC2,ZBTB7C,ZIC4,ZIC3,PLAGL1,KLF7,EGR4,KLF15 ,ZIC5,SP5,ZNF202,MZF1,ALX1
Profile 6	MTL5,ARID3C,ARID5A,HOXB6,RAX2,LHX3,NANO GP1,HLX,PDX1,HOXC13,HOXB13
Profile 10	ZBTB7C,KLF7,EGR4,ZIC5,GLIS2,ZIC3,ZNF202,ZIC2, ZIC4,ZNF148,SP5,SP9,PLAGL1,ZNF740,MZF1,ZNF26 3,KLF15,MLL, DNMT1

Table 4-2: Putative MRs retrieved for the gene expression clusters retrieved with STEM.

Overall, we predicted more than one potential MR for each cluster, while some predicted MRs could regulate more than one profile. We also found a direct connection between clusters that were regulated by the same TFs: for example, all the predicted putative MRs of profile 39, with the exception of ALX1, are also predicted putative MRs of the profile 10. In total, only 31 TFs were univocally predicted as MRs of one single cluster.

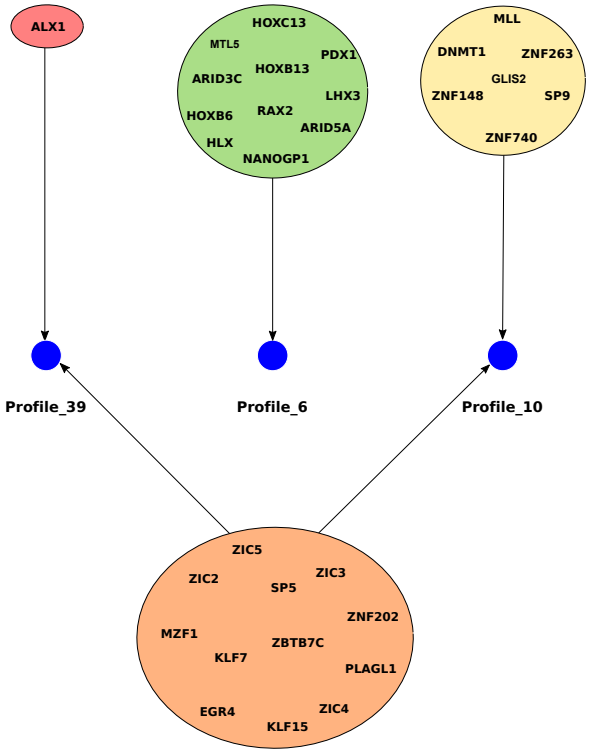


Figure 4-3: Predicted MRs and clusters regulated. In the figure we highlighted, by separating them, the common MRs between clusters 39 and 10.

4.1.4.3 Paths shorter than 4 steps in the PIN between MRs and potential receptors or Salmonella effectors

4.1.4.3.1 Path at 0 steps: when a Salmonella protein acts as a human TF

In order to cover the hypothesis of a Salmonella protein acting directly as a host TF, we tested sequence and function similarities between Salmonella proteins and human TFs (see methods).

We obtained several proteins in agreement with the criterion of sequence similarity (see methods) and searched for common interactions in the Human-Salmonella PIN. Interestingly we found 4 Salmonella effectors and 6 human TFs with sequence similarity. However, none of them shared interactions or a Pfam domain specific of DNA binding (

Table 4-3).

Human Uniprot entry	Salmonella Uniprot entry	Sequence Identity	Coverage of Target	Coverage of TF
RFX1_HUMAN	SSPH1_SALT1	31%	12%	9%
FOXI2_HUMAN	SOPA_SALT1	36%	9%	28%
MEF2B_HUMAN	SOPA_SALT1	34%	9%	22%
GBX2_HUMAN	SSPH2_SALT1	41%	7%	18%
PRRX1_HUMAN	SSPH2_SALT1	42%	6%	25%
ZHX1_HUMAN	SLRP_SALT1	36%	10%	10%

Table 4-3: Salmonella effectors that could act as Human TFs according to the criteria of sequence similarity. Percentage of identical residues aligned (sequence identity) and coverage of the aligned region with respect to the TF (Coverage of TF) and the Salmonella protein (Coverage of target) are shown for the criteria of sequence similarity.

We also studied the similarity between any Salmonella protein and the predicted MRs (

Human Uniprot entry	Salmonella Uniprot entry	Common Pfam domains	Sequence Identity	Coverage of Target	Coverage of TF
ZN202_HUMAN	D0ZNY7_SALT1	NA	31%	33%	26%
SP5_HUMAN	D0ZIU5_SALT1	NA	31%	22%	29%
KLF7_HUMAN	D0ZM42_SALT1	NA	28%	26%	65%
MZF1_HUMAN	D0ZUX2_SALT1	NA	40%	9%	12%
ZN263_HUMAN	D0ZRV5_SALT1	NA	33%	5%	14%
DNMT1_HUMAN	D0ZLC0_SALT1	DNA methylase	29%	36%	12%
PLAL1_HUMAN	D0ZLN1_SALT1	NA	62%	3%	7%
	D0ZXB6_SALT1	NA	36%	15%	17%
ODPX_HUMAN	D0ZJZ1_SALT1	NA	33%	35%	47%
	D0ZKY0_SALT1	NA	34%	13%	29%
	D0ZQF2_SALT1	NA	32%	96%	90%

Table 4-4). We found 8 putative MRs similar to Salmonella proteins according to the sequence criterion. None of these sequences shared common interactions, and only for 1 sequence (D0ZLC0_SALT1) we found a common Pfam domain with the MR (DNMT1_HUMAN) that could be indirectly associated with DNA-binding (a DNA-methylase Pfam domain).

Human Uniprot entry	Salmonella Uniprot entry	Common Pfam domains	Sequence Identity	Coverage of Target	Coverage of TF
ZN202_HUMAN	D0ZNY7_SALT1	NA	31%	33%	26%
SP5_HUMAN	D0ZIU5_SALT1	NA	31%	22%	29%
KLF7_HUMAN	D0ZM42_SALT1	NA	28%	26%	65%
MZF1_HUMAN	D0ZUX2_SALT1	NA	40%	9%	12%
ZN263_HUMAN	D0ZRV5_SALT1	NA	33%	5%	14%
DNMT1_HUMAN	D0ZLC0_SALT1	DNA methylase	29%	36%	12%
PLAL1_HUMAN	D0ZLN1_SALT1	NA	62%	3%	7%
	D0ZXB6_SALT1	NA	36%	15%	17%
ODPX_HUMAN	D0ZJZ1_SALT1	NA	33%	35%	47%
	D0ZKY0_SALT1	NA	34%	13%	29%
	D0ZQF2_SALT1	NA	32%	96%	90%

Table 4-4: Salmonella proteins that could act as Human MRs of the STEM clusters according to the criteria of i) sequence similarity; ii) common Pfam domains. Percentage of identical residues aligned (sequence identity), coverage of the aligned region with respect to the TF (Coverage of TF) and the Salmonella protein (Coverage of target) are shown for the criteria of sequence similarity. In the third column we mention, eventually, the names of the Pfam domains in common.

4.1.4.3.2 Paths between 1 and 3 steps

We calculated all paths shorter than 4 around the predicted MRs connecting them with Salmonella effectors and focused only on those involving human plasma membrane proteins. The implication of plasma-membrane proteins would be relevant for acting as receptors of a direct-interaction with the bacterial effectors in the SCVs entering the human cell. A large number of SPs were 3-step connections between a set of 6 MRs (ZIC2_HUMAN, GLIS2_HUMAN, PDX1_HUMAN, ZIC5_HUMAN, ODPX_HUMAN and A1YLA_HUMAN) and two Salmonella effectors: SPVB_SALT1 and SOPA_SALT1. We additionally found some SPs, also of 3-steps, between SPVB_SALT1 and four predicted MRs (ZN148_HUMAN, ZN740_HUMAN, DNMT1_HUMAN, and Q4FD37_HUMAN,) and two

more between SOPA_SALT1 and the predicted MRs B2RG49_HUMAN and ZBT7C_HUMAN.

4.1.4.4 Prediction of drugs for Salmonella infection based on MR-targets

We searched in cmap [314] for any drug with the potential to affect the expresison of the predicted putative MRs (see methods). Among the best ranked TF we found ARI5A_HUMAN , which is affected by Trichostatin A. Trichostatin A can be used to alter the profile of gene expression by interfering with the removal of acetyl groups of histones (histone deacetylases, HDAC) and the ability of DNA transcription. Ranks third and fourth are for ornidazole and despiramine also acting on ARI5A_HUMAN and affecting the regulation of the genes in profile 6. It's noteworthy that, we found tetracycline ranking five. Tetracycline affects the predicted main regulator that produces ZIC4_HUMAN, an inhibitor of the protein synthesis indicated for use against many bacterial infections. Reinforcing this point, in rank seven we found another known broad-spectrum antimicrobial drug, berberine, also affecting the expression of ZIC4_HUMAN and consequently the expression of genes in profiles 39 and 10.

4.1.5 Conclusions and discussions

This is a multidisciplinary approach that integrates the analyses of high-throughput sources of data, *in-silico* PPI predictions and protein-DNA binding to predict potential transcription factors implicated *Salmonella* infection of its host (human), and the underlying mechanisms of activation. Due to the rapidity of the invasion process of Samonella, our approach was specifically designed for short time series expression data. The integration step with *in-silico* predictions allowed us to hypothesize which transcription factors may be more involved in the regulation of the response of human cells upon invasion. Finally, we explored the potential role of drugs, some of them already applied in the therapy of Salmonellosis, that are directed to the predicted regulation and TFs of the cell response. The response of human cells can be modified by the drug, preventing Salmonella from hiding and evading the immune response.

We are aware of the spectacular and to date unique tolerance of Salmonella to extreme divergence of host species from plants to animals. This, combined with its antimicrobial resistance properties, leads to the necessity of the development of novel approaches to fight this growing global health problem.

A more detailed picture from the system biology approach may arise by the consideration that the networks we studied are static. Furthermore, it is known that Salmonella proteins can have different functions depending on their location [392]. Therefore, dealing with network time and location dependant may add an extra level of information. The prediction of main regulators involved in the response of the human cell to the invasion by Salmonella can be very useful to address this dynamic behaviour.

Finally, our approach can help to understand the mechanisms of action of some drugs used for therapy of Salmonellosis. Although antimicrobial therapy is not recommended for uncomplicated Salmonella gastroenteritis,

the determination of antimicrobial resistance patterns is often valuable for surveillance purposes [393]. Our approach based on drug specific gene signature retrieves coherent results with the actual therapies. Our results are encouraging, despite of suffering from the lack of genetic signatures for all drugs stored in Drug Bank.

4.1.6 Bibliography

The Bibliography for this chapter is at the end of this thesis.

5 PHARMACO-DYNAMIC DRUG-DRUG INTERACTIONS

In this chapter is presented a novel approach for the study of pharmaco-dynamic drug-drug interactions. Form public databases direct and indirect drug-targets are derived. In this context indirect are considered the most affected genes by the consumption of the chemical compound. The previously described methodology to reveal putative regulators is then applied to synthetically identify which transcription factors are affected by the drug. The combination of experimentally validated TF-gene and protein-protein interactions into a single human "signalling network" allows the description of the mechanisms of signal transduction leading from direct to indirect drug targets. Based on the hypothesis that interacting drugs should act on the same paths we modelled computationally the signal transduction by means of a message-passing algorithm. The targets of both drugs are used as signal emitters and their gene profiles (through the proposed MRs) as receivers. Then we compared and analysed the scores retrieved by the transcription factors and genes differentially expressed by both drugs for a few selected examples of interacting drugs.

This approach demonstrated promising results. I plan to benchmark it (see Discussion of this thesis) before its publication.

5.1 On the use of protein interaction networks and message passing algorithms to study potential mechanisms of drug-drug interactions

Daniel Poglayen[1] and Baldo Oliva[1*]

[1] Structural Bioinformatics Laboratory, Universidad Pompeu Fabra, Barcelona, Catalonia, Spain

* Corresponding author

E-mail: baldo.oliva@upf.edu

5.1.1 Abstract

For many decades the drug discovery field was dominated by the paradigm "single-drug, single target". Experience taught us about drug un-specificity, leading to side effects and toxicity. Recent advances in network pharmacology have enabled a system-biology view by which the chemical compound not only affects its targets but also their interactions. Thus, considering the cellular context of therapeutic targets has the potential to reduce toxicity and drug resistance while improving their clinical efficacy.

The study reported here perfectly fits into this frame. Through the integration of protein-protein interactions (PIN) and gene regulatory networks (GRN), we propose the application of a two-phase approach in order to identify the type of pharmaco-dynamic interaction occurring between two drugs. The first step consists in the identification of the transcription factors that are more likely to be affected by the consumption of a given drug, that we will call main regulators (MRs). We base this search on drug-genetic signatures, high-throughput derived and publicly available. In the second stage a message-passing algorithm is adopted to simulate the signal transduction from the known drug targets to the previously identified MRs. The results retrieved from the comparison of the scores obtained by the MRs and the regulated genes when the drugs are used alone or in combination, show that this approach allows distinguishing between different types of pharmaco-dynamic interactions.

5.1.2 Introduction

Adverse effects of drugs are one of the major risks of patient care and has become one of the major expenses of wellbeing in developed countries [394]. Several researches have addressed this problem by ranging from 3D pharmacophoric similarity studies [395] to mined databases and networks of chemicals and bio-targets (i.e. SIDER [306] and STITCH [396]). For many decades the drug discovery field was dominated by the paradigm "single-drug, single target" but experience taught us about drug "promiscuity": drugs that were theoretically single-target designed turned out not to be so specific. This leads to drug side effects and toxicity. The most widely known case is Sildenafil (known as Viagra), initially designed to relax coronary arteries to increase blood flow, turned out to have a side effect more profitable: penile erection [397]. Another example is the one of efalizumab, approved to treat autoimmune diseases, like psoriasis, five years later was removed from the market because it was responsible for the reactivation of the latent polyomavirus JVC [398]. The recent advent of network pharmacology [247] exploits the current knowledge of systems biology to study how one or more drugs affect not just their molecular targets, but also their network. This approach accommodates for the presence of multiple functions, alternative paths and backup circuits that lead to an increase of the robustness of biological systems to perturbations [399], such as a drug. Considering the cellular context of therapeutic targets has the potential to improve clinical efficacies through different strategies and to reduce toxicity and drug resistance. The most promising strategies include drug repositioning, finding new uses of existing drugs, and the identification of interacting drugs that can be used in combination to treat a certain phenotype in a synergistic fashion. Drug-drug interaction (DDI) occurs when the pharmacologic effect of a given drug is altered by the action of another drug leading to different clinical outcome than with individual drugs alone. Two types of DDIs are possible: pharmaco-kinetic and pharmaco-dynamic ones. The first type

consists in one drug changing the systemic concentration of another, thus altering its effect. The second ones occur when interacting drugs have either additive or synergistic effects, in which case the overall effect is increased, or opposing/antagonistic effects, in which case the overall effect is decreased or even 'cancelled out'.

Among the advantages of studying DDI, apart from the mentioned reduction in the risk of undesired side effects and drug resistance, is the fact that the single compounds have already been human approved, allowing an eventual experimentation on the combination to enter directly in Phase II, reducing cost and time of the study.

In the context of network pharmacology, an increasing number of computational methods has been developed, or adapted, to predict DDI, which allows for high-throughput *in-silico* screening and predictions, thus further lowering cost and time. Such computational methods can be similarity based or knowledge based. The first ones include methods based, for example, on the chemical structure of the compounds, two drugs are linked if they share structural properties [304], on the targets, two drugs are connected if they share at least one target protein [400], on indications, the drugs are connected if they share a common therapeutic indication [401], on side effects, the link derives from the similarity of the drug's side effects [305], on gene expression profiles, that connect drugs according to the correlation of the resulting gene expression data [314] and on clinical effects [402]. Most of them were previously applied for drug discovery but can be employed also for studying DDI. Another tool for predicting DDI is INDI, whose predictions are based on chemical and side effect similarity to known interactions [318]. Knowledge-based methods predict DDI based on scientific literature, for example STITCH [403] that links two drugs according to a literature co-occurrence scheme, electronic medical record database [404] and the Food and Drug Administration Adverse Event Reporting System. In the latter case drugs are connected if they are

associated at least to one reported adverse event not attributable to the individual drugs alone [405].

In this work we propose the integration of protein-protein interaction networks (PIN) and gene regulatory networks (GRN) for studying DDIs. After the creation of a joint PIN-GRN network and the identification of putative main regulators (MRs) derived from the gene expression signatures specific for each drug, we apply a message passing algorithm to the paths starting from the drug targets, first individually for each drug and then in combination. We compare, for each type of DDI, the scores obtained by the predicted MRs and the most affected genes derived from cmap (see methods) in the individual and combined cases.

5.1.3 Materials and methods

5.1.3.1 Direct and indirect targets of drugs retrieved from public databases

Given a drug of interest the first step of our approach is to search for its targets in Drug Bank (v.4.3) [319]. We then investigate which genes are most affected by the consumption of the pharmaceutical compound. To do this we build drug-genes connections derived from the connectivity map database (cmap v.02) [314]. This database provides a total ranking of almost 22,000 uniprot accession of genes, according to their differential expression profiles when treated with different bioactive small molecules.

5.1.3.2 Prediction of transcription factors affected by drugs

We then retrieve the promoter regions of the top 15 most affected genes (profiled-genes), by using the Eukaryotic Promoter Database [391], and, subsequently, we look for putative transcription factor binding motifs at their promoter shared by the majority of the 15 genes, by using DISPOM

[82]. The promoters are then scanned for occurrences of these motifs using FIMO [214] in a combined database of Position Weight Matrices of TFs collected from CisBP [94] and Jaspar [84]. Thanks to this we will be able to create paths that go from the drug to its most affected genes, passing from drug targets first and then from the TFs responsible for such difference in expression, that we call putative main regulators (MRs).

5.1.3.3 Signalling network of drugs: from direct targets to genes expression

First, we expanded the gene-regulatory network (GRN) linking TFs and the respective regulated genes, with the data of PAZAR [213], a public repository that contains transcription factors and regulatory sequence annotations. Second, we used the framework BIANA (release 2013.1)[199] to integrate several sources of protein-protein interactions (biogrid [191], dip [192], hprd [193], intact [296], mint [297] and, from the UniProt consortium, Swiss-Prot and Trembl [406]) and obtain the an experimentally validated human PPI network (PIN) by yeast two hybrid experiments. However, there is a low number of interactions for TFs that can be trusted by yeast-two-hybrid experiments. Therefore, we added the protein-protein interactions associated with TFs obtained by affinity purification methods. Finally, we merged both networks (PIN+GRN) to describe mechanisms of signal-transduction into a single "signalling network" (SN) and mapped our drug-target knowledge: i) identifying the protein drug-targets from Drug Bank; and ii) selecting the most affected genes by the drug from the cmap database, together with their putative MRs previously predicted.

5.1.3.4 Computational modelling and analysis of the signalling-network (SN)

Our objective was to compare the connections linking drug-targets with its cellular expression consequences between two drugs. We were interested in those paths implicated in the expression signature of a drug and identify if

two drugs could be related or associated if they were acting on the same paths. Therefore, we modelled the signal transduction computationally by means of a message-passing algorithm for gene prioritization, using drug targets as sources of the signal (seeds) and the gene profiles of the drug (throught the proposed MRs) the receivers of the signal. We applied a modified version of the NetScore algorithm of GUILD [282], to obtain the raw scores, running 5 iterations. First, we scored the nodes of the SN for each drug, using their corresponding seeds. Second, we used the seeds of both drugs to score again the nodes of the network. Then, we compared and analysed the scores of all nodes and in particular the set of common transcription factors and genes differentially expressed by both drugs.

5.1.3.5 Analyses of drug-drug interactions

We tested our comparison on a few selected examples of interacting drugs from the drug combination database (DCDB v.2.0) [407], which contains and organizes 1363 known examples of drug combinations with their activity/indications, possible mechanisms and drug interactions between its components. The database classifies efficacious or non-efficacious combinations: all the ones approved from the Federal and Drug Administration (FDA) are classified as efficacious, for drug combinations in Phase I trial are considered efficacious if the overall outcome is a "pass" and there is evidence of improved benefits. In case it is a Phase II/III/IV trial the efficiency depends on the absence of unacceptable toxicity and on the increase of effectiveness compared to current first-line or single mono-drug therapies. In case of a pre-clinical trial only this last aspect is taken into consideration to categorize the combination. The interactions of two drugs toward a specific phenotype can be classified as synergistic, additive or antagonistic: synergistic when both drugs address the same phenotype and the measure of their action is higher than the simple addition of both; additive when the result in efficiency of two drugs can be interpreted as the

sum of both; and antagonistic, when the effect of one drug is diminished by the other.

Consequently, we compared the scores of predicted MRs and drug-profiled genes. For each predicted MR and gene of the drug-profile, we calculated the score-difference of drugs A and B, and named Δ_{AB}, between the scores obtained when using the seeds of both drugs (score of the combination, Sc_{AB}) and the sum of scores obtained when using each single drug (Sc_A and Sc_B).

5.1.4 Results

From DCDB (v.2.0) [407] we extracted a few examples for all the types of pharmaco-dynamic interactions mentioned above for which we could find data relative to their targets in Drug Bank, most affected genes in cmap and consequent prediction of putative MRs. We ran our algorithm and here we present our first results both in terms of MRs and most affected genes from cmap. The predicted genetic interactions for each drug in this study are listed in Supplementary Material.

5.1.4.1 Additive drug-drug interactions

In the case of additive interaction, the effect of the combination of two drugs should be reflected by the "sum" of the effects of each individual drug. Thus, we expect the scores obtained by drug-profiled genes and their putative MRs with the seeds of the combination of the two drugs to be equal to the sum of scores obtained when using individual drugs. We tested our results for the combinations involving dorzolamide and timolol on one side and, on the other hydrochlorotiazide and metoprolol (Table 1). The results are in agreement with our hypothesis, implying that for additive drug-drug

interactions the pathways connecting the targets with the main-regulators of the expression profile of each drug are the same or equally distant.

A)

	Uniprot entry	Dorzolamide Sc_A	Timolol Sc_B	Combination Sc_{AB}	Δ_{AB}
Putative MR Timolol	SIX3_HUMAN	0.0128	0.02515	0.03796	0.0000
	Q8TBA2_HUMAN	0.0128	0.2515	0.03796	0.0000
	SIX4_HUMAN	0.00001	0.0247	0.02471	0.0000
	SIX1_HUMAN	0.02469	0.02561	0.0503	0.0000

B)

	Uniprot entry	Hydrochlorothiazide Sc_A	Metoprolol Sc_B	Combination Sc_{AB}	Δ_{AB}
Putative MR Metoprolol	GLIS3_HUMAN	0.00047	0.00093	0.0014	0.0000
	ZN281_HUMAN	0.00137	0.00184	0.00321	0.0000
	D3GC14_HUMAN	0.01326	0.02607	0.03933	0.0000
	ZFY_HUMAN	0.00047	0.00093	0.0014	0.0000
	NFKB1_HUMAN	0.02515	0.03796	0.06311	0.0000

Table 5-1: GUILD scores of putative MRs in the case of additive DDI. In the table are shown the scores obtained by the predicted putative MRs, derived from the cmap database (in the first column), using the targets of each drug individually as seeds (second and third columns) and for the drug combination (fourth column). Scores are calculated using the NetScore algorithm in GUILD. Section **A)** refers to the combination **dorzolamide-timolol** and section **B)** to **hydrochlorothiazide-metoprolol**.

Our approach did not find any putative MR (see methods) for dorzolamide and hydrochlorothiazide, thus we checked, for each combination, the scores of their most affected genes according to cmap and we confirm the same observation (Table S1). Only exceptions, in the case of dorzolamide and timolol, are EFNB3_HUMAN and CASL_HUMAN, with a non-significant lower (Δ=-0.0357) and higher score (Δ= 0.00045), respectively,.

5.1.4.2 Antagonistic drug-drug interactions

For antagonistic interactions, we considered the examples of the interactions diphenhydramine-theophylline and aminophylline-theophylline. The difference Δ calculated on MRs were always negative (Table 2). In fact, we corroborated that the score of the combination of two drugs was always the best score of one of them, probing that the connection between the targets of one drug and its MRs were shorter than for the other drug. We also checked the drug-profiled genes and we observed the same feature (Table S2). In particular for theophylline, it was not possible to predict a MR, therefore we could only perform this analysis on the drug-profiled genes.

A)

	Uniprot entry	Diphenhydramine Sc_A	Theophylline Sc_B	Combination Sc_{AB}	Δ_{AB}
Putative MR diphenhy-dramine	SP1_HUMAN	0.05075	0.51989	0.5583	-0.01234
	A0PJI1_HUMAN	0.00275	0.03979	0.07865	0.03611

B)

	Uniprot entry	Aminophylline Sc_A	Theophylline Sc_B	Combination Sc_{AB}	Δ_{AB}
Putative MR aminophylline	D3GC14_HUMAN	0.15135	0.01372	0.15135	-0.01372
	AP2C_HUMAN	0.1989	0.02561	0.1989	-0.02561
	KLF8_HUMAN	0.0279	0.01372	0.0279	-0.01372

Table 5-2: GUILD scores of putative MRs in the case of antagonistic DDI. In the table are shown the scores obtained by the predicted putative MRs, derived from the cmap database (in the first column), using the targets of each drug individually as seeds (second and third columns) and for the drug combination (fourth column). Scores are calculated using the NetScore algorithm in GUILD. Section **A)** refers to the combination **diphenhydramine-theophylline** and section **B)** to **aminophylline-theophylline.**

5.1.4.3 Synergistic drug-drug interactions

For synergistic interactions we selected the pairs: enalapril-hydrochlorothiazide, imatinib-vorinostat and glipizide-metformin.

The difference Δ calculated on MRs were always 0, as in the case of additive interactions, the only exception we found is for enalapril-hydrochlorothiazide (Table 3A). We could not retrieve putative MRs for hydrochlorothiazide and imatinib but in the case of glipizide and metformin we identified 9 common putative MRs:

C9JXZ2_HUMAN,	Q96SH1_HUMAN,	C1K3N0_HUMAN,
C9J6N8_HUMAN,	AP2E_HUMAN,	F8WDC8_HUMAN
F8WEX2_HUMAN,	H7C5E5_HUMAN,	H7C4N4_HUMAN.

Any of them was present in the network thus we could not check their scores. We then compared the differences Δ_{AB} calculated on the top ranking nodes derived from the cmap database (Table S3). T22D4_HUMAN is the only case in which we observe an increased Δ_{AB} combining the seeds of enalapril and hydrochlorothiazide. The other scores Δ_{AB},retrieved for the synergistic combinations studied show diminished values or 0s.

A)

	Uniprot entry	Enalapril Sc_A	Hydrochlorothiazide Sc_B	Combination Sc_{AB}	Δ_{AB}
Putative MR Enalapril	AP2C_HUMAN	0.02607	0.0375	0.06311	-0.0005

B)

	Uniprot entry	Imatinib Sc_A	Vorinostat Sc_B	Combination Sc_{AB}	Δ_{AB}
Putative MR Vorinostat	MEF2A_HUMAN	0.02561	0.06585	0.09146	0.0000
	HES4_HUMAN	0.03749	0.37403	0.41153	0.0000

C)

	Uniprot entry	Glipizide Sc_A	Metformin Sc_B	Combination Sc_{AB}	Δ_{AB}
Putative MR Glipizide	SP1_HUMAN	0.1358	0.02515	0.16096	0.0000
	Q5T6X2_HUMAN	0.00092	0.00092	0.00184	0.0000
	AP2C_HUMAN	0.03704	0.01327	0.0503	0.0000
	SP5_HUMAN	0.00047	0.00001	0.00048	0.0000
	AP2B_HUMAN	0.03704	0.01327	0.0503	0.0000
Putative MR Metformin	I6L9H2_HUMAN	0.03704	0.01281	0.04985	0.0000

Table 5-3: GUILD scores of putative MRs in the case of synergistic DDI. In the table are shown the scores obtained by the predicted putative MRs, derived from the cmap database (in the first column), using the targets of each drug individually as seeds (second and third columns) and for the drug combination (fourth column). Scores are calculated using the NetScore algorithm in GUILD. Section **A)** refers to the combination **enalapril-hydrochlorothiazide,** section **B)** to **imatinib-vorinostat** and section **C)** to **glipizide- metformin.**

5.1.5 Conclusions and discussions

Our network pharmacology approach for studying pharmaco-dynamic interactions integrates gene regulatory and protein-protein interaction networks for a better traceability of drug effects in the cell. It aims to find key players for defining the type of DDI. To do this a central role is played by a gene prioritization algorithm and we entrust the task to the comparison of the scores obtained by previously predicted main regulators in the genetic response to the specific drugs in the case of individual and combined drug consumption. The different topological combinations and how they can

affect the specific Δ_{AB} score calculations depending on the message passing algorithm are depicted and explained in Figure 1.

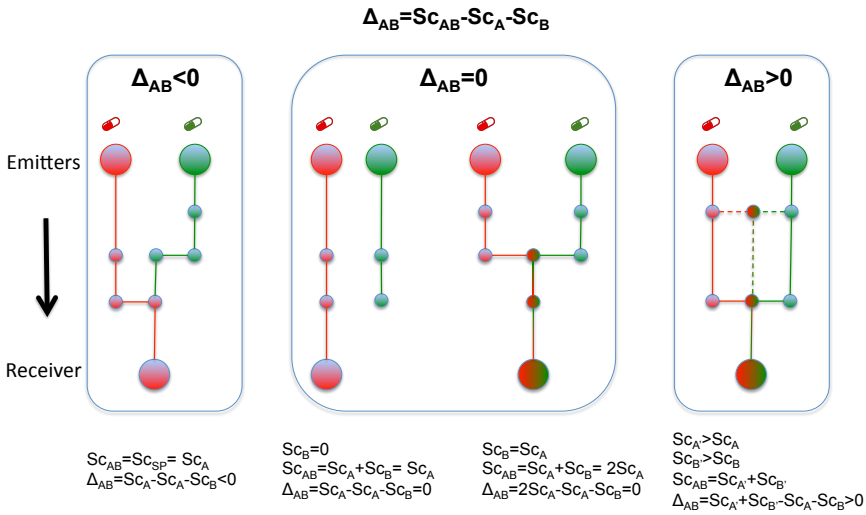


Figure 5-1: Topology effects on the Δ_{AB} score. The figure shows the different topological combinations that our Δ_{AB} score reflects. It can assume negative values when the shortest path (SP) connecting the emitter of one drug to the receiver is shorter then the other. Thus the score assigned to the receiver will depend only on the message sent by the closer emitter. Δ_{AB} can assume a value of 0 when i) the emitter of one drug is not connected with the receiver or ii) the SP connecting the emitters from the two drugs and the receiver has exactly the same length. Finally we can observe positives scores when, using as seeds the combination of the two drugs targets, a node that was previously getting a lower score, now, being connected with both emitters, exhibits a higher score. The SP from this node to the receiver must of equal length than the previously found SPs.

With some examples we proved that this approach reflects the expectations, in terms of scoring, for additive and antagonistic DDIs. Limitation to this type of approaches is the inability to handle novel drugs and the scarse availability of drug specific gene expression signatures,

together with limited knowledge in the field of gene regulation. Notwithstanding this, it can be definitely considered a high potential approach as this type of information is collected in databases of increasing completeness. Further improvements to our approach involve the possibility to tune the initial scores of the seeds representing individual drug dosage and study the impact on the scores of the predicted putative MRs and gene expression signature. This better reflects that synergism is not merely a property of two drugs. It also depends on the doses of each in the combination. Thus, to determine synergism, point where our approach seems to be most deficient, a quantitative approach should begin with the individual dose-effect curves from which the combined additive effect is calculated. If the combined effect is significantly greater than the expected (additive) effect, there is synergism.

Despite this, gene prioritization algorithms, together with the integration of PINs and GRNs appears to be a promising line for the pharmaco-dynamic study of drug interactions.

5.1.6 Bibliography

The Bibliography for this article is at the end of this thesis.

5.1.7 Supplementary Information

5.1.7.1 Predicted regulators of the top ranking genes found in cmap for each drug studied.

Drug	TF	Gene
Aminophylline	ZIC5	klhl25
	KLF8	dusp9
	TFAP2A	klhl25
	TFAP2E	dusp9
	KLF6	dusp9
	KLF7	dusp9
	KLF7	klhl25
	KLF6	klhl25
	KLF8	klhl25
	TFAP2E	klhl25
	TFAP2C	klhl25
Diphenhydramine	TFAP2A	ttll1
	SP1	s100a8
	HIC1	miip
	SP1	ndrg2
	MZF1	ndrg2
	MZF1	s100a8
Enalapriil	TFAP2C	hps4
	TFAP2A	hps4
	MECP2	hps4
	TFAP2E	hps4
	ZIC5	hps4
Glimepiride	DNMT1	map4
	MLL	plscr3
	ZIC5	map4

	ZIC5	cat
Glipizide	KLF4	scaf4
	Klf4	scaf4
	TFAP2B	klf6
	TFAP2C	klf6
	KLF5	scaf4
	SP9	msh3
	KLF7	msh3
	TFAP2A	klf6
	SP7	scaf4
	KLF5	msh3
	SP9	scaf4
	KLF7	scaf4
	TFAP2E	klf6
	SP1	msh3
	SP1	scaf4
	SP6	scaf4
	SP7	msh3
	KLF4	msh3
	SP5	scaf4
	SP6	msh3
	SP5	msh3
	Klf4	msh3
Metformin	DNMT1	flnc
	DNMT1	ap1s1
	TFAP2E	ap1s1
	TFAP2A	ap1s1
	TFAP2E	rpl10
	ZIC5	ap1s1
Metoprolol	NFKB1	fgfr2

	TFAP2A	srpk3
	KLF6	irf9
	ZFY	fgfr2
	GLIS3	irf9
	TFAP2A	tp53i11
	Sox3	fgfr2
	KLF6	fgfr2
	ZNF281	irf9
	KLF7	irf9
	TFAP2A	fgfr2
	ZIC5	irf9
	KLF7	srpk3
	KLF7	fgfr2
	KLF6	srpk3
Vorinostat	MEF2A	dnajc6
	HES4	st3gal5
	TCFL5	st3gal5
	MEF2A	st3gal5
	HES4	dnajc6

Table S 5-1: GUILD scores of top ranking cmap genes in the case of additive DDI. In the table are shown the scores obtained by the top ranking genes from the cmap database (in the first column), obtained using the targets of each drug individually as seeds (second and third columns) and for the drug combination (fourth column). Scores are calculated using the NetScore algorithm in GUILD. Section A) refers to the combination dorzolamide-timolol and section B) to hydrochlorothiazide-metoprolol.

A)

	Uniprot entry	Dorzolamide Sc_A)	Timolol Sc_B	Combination Sc_{AB}	Δ_{AB}
Cmap Timolol	IFIT1_HUMAN	0.00091	0.013172	0.01463	0.0005
	TTC38_HUMAN	0.00091	0.00183	0.00274	0.0000
	Q86V38_HUMAN	0.12346	0.04938	0.17284	0.0000
	CXCR4_HUMAN	0.00046	0.01372	0.01418	0.0000
	STX7_HUMAN	0.00091	0.11248	0.1134	0.0000
	EFNB3_HUMAN	0.02469	0.02561	0.01463	-0.03567
Cmap Dorzolamide	PLS3_HUMAN	0.0128	0.01372	0.02652	0.0000
	MASP1_HUMAN	0.02469	0.13626	0.16095	0.0000
	CFLAR_HUMAN	0.00091	0.00183	0.00274	0.0000
	ZN609_HUMAN	0.00091	0.00183	0.00274	0.0000
	ZN639_HUMAN	0.00046	0.00093	0.00139	0.0000
	PGFRA_HUMAN	0.0046	0.00002	0.00049	0.0000
	FGF1_HUMAN	0.00091	0.01372	0.01463	0.0000
	ENOX1_HUMAN	0.11157	0.00183	0.1134	0.0000
	CASL_HUMAN	0.0128	0.00138	0.01463	0.0005

B)

	Uniprot entry	Hydrochloro-thiazide Sc_A	Metoprolol Sc_B	Combination Sc_{AB}	Δ_{AB}
map Metoprolol	MX1_HUMAN	0.00137	0.12483	0.12621	0.0000
	Q8TCE5_HUMAN	0.02515	0.00229	0.02744	0.0000
	IFIT1_HUMAN	0.00092	0.00229	0.00321	0.0000
	SRPK3_HUMAN	0.01236	0.00093	0.01374	0.0000
	IRF9_HUMAN	0.02515	0.00229	0.02744	0.0000
	Q96E98_HUMAN	2e-05	2e-05	3e-05	0.0000
	OAS1_HUMAN	0.00092	0.00184	0.00276	0.0000
	IFI6_HUMAN	0.00092	0.00184	0.00276	0.0000
Cmap hydrochloro-thiazide	DOK5_HUMAN	0.0247	0.11249	0.13719	0.0000
	PBIP1_HUMAN	0.01326	0.02561	0.03887	0.0000
	SVIL_HUMAN	0.11203	0.04985	0.16187	0.0000
	GBRB2_HUMAN	0.00092	0.00229	0.00321	0.0000
	ZN230_HUMAN	0.00092	0.00184	0.00276	
	HSP76_HUMAN	0.01281	0.00138	0.01419	
	PLCD1_HUMAN	0.01326	0.00184	0.0151	
	NPHP1_HUMAN	0.00137	0.01418	0.01555	

Table S 5-2: GUILD scores of top ranking cmap genes in the case of antagonistic DDI. In the table are shown the scores obtained by the top ranking genes from the cmap database (in the first column), obtained using the targets of each drug individually as seeds (second and third columns) and for the drug combination (fourth column). Scores are calculated using the NetScore algorithm in GUILD. Section A) refers to the combination diphenhydramine-theophylline and section B) to aminophylline-theophylline.

A)

	Uniprot entry	Diphenhydramine Sc_A	Theophylline Sc_B	Combination Sc_{AB}	Δ_{AB}
Cmap Diphenhy-dramine	S10A8_HUMAN	0.13763	0.17513	0.30041	-0.0124
	NDRG2_HUMAN	0.02607	0.06402	0.07775	-0.0123
	S10A7_HUMAN	0.02653	0.0407	0.05488	-0.0123
	MIIP_HUMAN	0.0023	0.04024	0.04208	-0.0005
Cmap Theophyl-line	DNJA4_HUMAN	3e-05	0.00277	0.00279	0.0000
	RBGP1_HUMAN	0.0023	0.11568	0.11752	-0.0005
	TRPS1_HUMAN	0.05075	0.05258	0.09099	-0.0123
	AKA11_HUMAN	0.00139	0.04024	0.04118	-0.0006
	RB11B_HUMAN	0.01464	0.0279	0.04208	-0.0004
	SP100_HUMAN	0.05075	0.18701	0.22542	-0.0123

B)

	Uniprot entry	Aminophylline Sc_A	Theophylline Sc_B	Combination Sc_{AB}	Δ_{AB}
Cmap Thephylline	DNJA4_HUMAN	0.00277	0.00138	0.00277	-0,0014
	RBGP1_HUMAN	0.11568	0.00183	0.11568	-0,0018
	TRPS1_HUMAN	0.05258	0.01372	0.05258	-0,0137
	AKA11_HUMAN	0.04024	0.02561	0.04024	-0,0256
	RB11B_HUMAN	0.0279	0.00183	0.0279	-0,0018
	SP100_HUMAN	0.18701	0.02561	0.18701	-0,0256
Cmap Aminophylline	B2RAL8_HUMAN	0.02835	0.00183	0.02835	-0,0018
	EFNA3_HUMAN	0.13901	0.00138	0.13901	-0,0014
	CC85B_HUMAN	0.08825	0.03749	0.08825	-0,0375
	CO1A1_HUMAN	0.05213	0.01372	0.05213	-0,0137
	TLR2_HUMAN	0.00277	0.00093	0.00277	-0,0009

Table S 5-3: GUILD scores of top ranking cmap genes in the case of synergistic DDI. In the table are shown the scores obtained by the top ranking genes from the cmap database (in the first column), obtained using the targets of each drug individually as seeds (second and third columns) and for the drug combination (fourth column). Scores are calculated using the NetScore algorithm in GUILD. Section A) refers to the combination enalapril-hydrochlorothiazide, section B) to imatinib-vorinostat and section C) to glipizide- metformin.

A)

	Uniprot entry	Enalapril Sc_A	Hydrochloro-Thiazide Sc_B	Combination Sc_{AB}	Δ_{AB}
Cmap Hydrochloro -thiazide	DOK5_HUMAN	0.11249	0.12347	0.23595	0,0000
	PBIP1_HUMAN	0.02561	0.02516	0.05077	0,0000
	SVIL_HUMAN	0.04985	0.13627	0.17377	-0,0124
	GBRB2_HUMAN	0.00229	0.00138	0.00321	-0,0005
	ZN230_HUMAN	0.00184	0.00138	0.00321	0,0000
	HSP76_HUMAN	0.00138	0.00138	0.00275	0,0000
	PLCD1_HUMAN	0.00184	0.00093	0.00231	-0,0005
	HPHP1_HUMAN	0.01418	0.00183	0.01555	-0,0005
Cmap Enalapril	Q5ST80_HUMAN	0.02607	0.02516	0.03888	-0,0124
	DYST_HUMAN	0.23549	0.01372	0.24875	-0,0005
	LZTR1_HUMAN	0.02607	0.02561	0.05122	-0,0005
	RRBP1_HUMAN	0.03751	0.02516	0.06266	0,0000
	T22D4_HUMAN	0.02607	0.11249	0.16187	0,0233
	NDE1_HUMAN	0.00183	0.00048	0.00229	0,0000

B)

	Uniprot entry	Imatinib Sc_A	Vorinostat Sc_B	Combination Sc_{AB}	Δ_{AB}
Cmap Vorinostat	CTGF_HUMAN	0.13626	0.06585	0.20211	0,0000
	DPYL4_HUMAN	0.01372	0.01739	0.01922	-0,0119
	GLRX1_HUMAN	0.00138	0.0037	0.00507	0,0000
	CASL_HUMAN	0.14815	0.07774	0.22589	0,0000
	H10_HUMAN	0.01372	0.36215	0.37586	0,0000
	TBB2A_HUMAN	0.03749	0.18884	0.21445	-0,0119
	TBB2B_HUMAN	0.35848	0.1765	0.5231	-0,0119
Cmap Imatinib	Q5ST80_HUMAN	0.00183	0.21217	0.214	0,0000
	TGT_HUMAN	0.03749	0.07774	0.10334	-0,0119
	TRPC1_HUMAN	0.03749	0.1765	0.20211	-0,0119
	PI51A_HUMAN	0.00047	0.01331	0.01378	0,0000
	NPRL2_HUMAN	0.01372	0.0183	0.02012	-0,0119
	HSP74_HUMAN	0.01372	0.02928	0.04299	0,0000
	ENOX2_HUMAN	0.02561	0.01694	0.04255	0,0000

C)

	Uniprot entry	Glipizide Sc_A	Metformin Sc_B	Combination Sc_{AB}	Δ_{AB}
Cmap Glipizide	MILK1_HUMAN	0.00047	1e-05	0.00048	0,0000
	SFR15_HUMAN	0.00137	0.00092	0.00229	0,0000
	Q5T6X2_HUMAN	0.03704	0.00137	0.03841	0,0000
	CDK20_HUMAN	0.02515	0.00047	0.02562	0,0000
	IFI6_HUMAN	0.01326	0.00092	0.01418	0,0000
	SUGP1_HUMAN	0.00092	2e-05	0.00094	0,0000
	CUL7_HUMAN	0.01326	0.00092	0.01418	0,0000
Cmap Metformin	AP1S1_HUMAN	0.02515	0.00183	0.01463	-0,0124
	NEBL_HUMAN	0.11203	0.00137	0.1134	0,0000
	Q59H94_HUMAN	0.03704	0.02515	0.06219	0,0000
	PLK4_HUMAN	0.00137	0.00092	0.00229	0,0000
	DOK5_HUMAN	0.12391	0.00092	0.12483	0,0000
	RL10_HUMAN	0.01326	0.11157	0.12483	0,0000

6 DISCUSSION

I have presented three different studies in this thesis, all of them revolving around a common theme: discovering new components of signalling pathways leading to the activation of main regulators of diverse biological processes from different biological species. Particularly, I have applied the same rationale to tackle two completely different problems: i) the invasion of a pathogen and its subsequent interaction with the hosts, *Arabidopsis* (in chapter 3) and human (chapter 4), and ii) the study of drug-drug interactions. Despite of the inherent biologically diversity among these three pieces of work, they share a number of **common denominators** allowing for a common strategy to address them :

- **Analysis of high-throughput data**. Both the study of the mechanisms of *Salmonella spp.* infection (chapters 3 and 4), and the research for the two HPN-DREAM Challenges in which I participated (in the Appendix of this book), required the analysis of multiple types of data coming from different microarray platforms. In the study on drug-drug interactions (chapter 5) I made use of the connectivity map database that collects drug specific gene signatures derived from microarray experiments [314]. As previously introduced, this technology suffers from biased signals due to cross-hybridization and limited dynamic range from saturation of the fluorescence signal [40]. In addition, it is genome-annotation dependant. All these aspects have been solved with the advent of RNA-seq. With this technique gene expression levels of thousands of genes are measured simultaneously. The amount of additional information, with respect to microarrays, includes alternative splicing, allele-specific expression, un-annotated exons and novel transcripts (genes and non coding RNAs). The global view that can be obtained is much more detailed, with less prior knowledge. For these reasons, it is widely spreading in the

scientific community. Costs related to this technology are lowering and protocols are being unified. In part this revolutionary technology has already been used in drug discovery to identify drug-related genes [408] but clearly there is still much to learn in terms of cellular drug response, drug resistance and drugs combinations.

Apart from microarrays, I exploited several databases containing high-throughput derived transcription factor (TF) DNA-binding profiles (CIS-BP [94], PAZAR [213] and JASPAR [84]) in order to predict putative main regulators (MRs) for a set of expressed genes of interest (chapters 3, 4 and 5). I would like to remark the fact that in all the work I modelled gene expression as a function of the sole activity of TFs. Although I am aware of the limitations of this assumption, which does not consider other crucial elements in transcriptional regulation, such as epigenetic factors, I showed that, by integrating several knowledge-based databases and experimental results, it is possible to identify those TFs which could, potentially, drive major gene expression signatures. This, using a simplified, easier to understand and to process analysis model, with the sufficient amount of data it is possible to obtain an acceptable prediction level.

- **Integration of the analysis of *in-vivo* data with protein-protein interaction networks**. Despite the continuous growth of available protein-protein interaction (PPI) data, our current knowledge of the mechanisms governed by such interactions is still limited. In order to gain insight in this topic –i.e. how specific PPI contribute to the expression of gene products performing certain biological functions, I made extensive use of BIANA [199] (chapter 5) and of the BIANA-derived server

BIPS [295] (chapters 3 and 4) for the study of *Salmonella* infection. The flexibility provided by the integration protocol of BIANA is one of the most remarkable features of this framework and very unique to it. This flexibility allowed for selecting the exact type of interactions that better suited the study of each biological question posed in this thesis. Taking advantage of such feature was specially relevant in the case of bacterial infection. The use of BIANA allowed mapping proteins from both the host and the pathogen on the same network. Applying, then, the message passing algorithm contained in GUILD [282] for the identification of new function-related components allowed the discovery of the host's MRs and the prediction of more bacterial effectors (see below in this discussion).

- **Prediction of the main regulators of a group of genes**. It has been previously introduced that checking the expression of thousands of genes provides the opportunity to find similarities among them. Clustering algorithms play a central role on this task, grouping in the same cluster genes for which common transcriptional mechanisms can be hypothesized. This strategy is widely applicable in different experimental and computational settings. In these pages I consistently adopted the same strategy trying to identify one or more TFs, which I refer to as MRs, whose activity influences a larger group of downstream genes (chapters 3, 4 and 5). I used a variety of bioinformatics tools (i.e. DISPOM [82], T-reg comparator [215], FIMO [214] and Tomtom [216]) and databases (named in the first point of this list: *Analysis of high-throughput data*) to predict the MRs. Therefore

the predictions rely on a combination of genomic sequence information and database mining.

This allowed the creation of a predicted gene regulatory network (GRN) that I further integrated in the PPI network in order to understand the regulatory elements (**who**) and the molecular basis (**how**) of an observed genetic signature.

- **Application of gene prioritization algorithms**. A straight forward step after the prediction of MRs, as already anticipated in this discussion, is to understand which paths of the network are more likely to be responsible for the observed gene expression, thus for the phenotype. Gene prioritization algorithms are specifically designed to handle this task. In the study of *Salmonella* infection in *Arabidopsis* (chapter 3), I choose to apply NetCombo from the GUILD framework [282], the performances of which has been shown to be better than other state-of-the-art gene prioritization methods [359]. However, for the study on drug-drug interactions (chapter 5), the same exact algorithm could not be applied, so I opted for a modified version of NetScore from the same framework. The algorithm propagates the score assigned to the message emitters (seeds) through the network, specifically each node to its neighbours. I used a modified version of this algorithm because the original contains a normalization step at the end of each iteration. That would have limited the variability of the resulting scores for each individual drug, and for the drug combination, separately. Since the aim of this analysis was to focus on the comparison of these scores, any variation, although minimal, needed to be observable. For this reason the modified version used returns raw scores instead of normalized ones.

> Identifying key network components through gene prioritization algorithms can prove the foundation to develop or combine already approved drugs to target specific paths and, for example, disrupt the cycle of an infecting pathogen.

The application of the afore mentioned strategy lead to the production of several *in silico* predictions about the involvement of different *Arabidopsis* proteins in the regulation of the gene expression during *Salmonella* infection (see Chapter3). Notably, my predictions about the early involvement of two proteins (WRKY18 and WRKY60) in early stages of the bacterial infection have been experimentally validated (see Chapter 3.1.5.4.2). Both proteins belong to the same family (WRKY) which are known to activate stimulus-dependant, PAMP-triggered, defence response genes [371].

After deriving the putative MRs from *Salmonella* infected human data (see Chapter 4), the search result of a gene specific drug therapy addressing those TFs resulted coherent with the type of infection. Although antimicrobial therapy is not recommended for uncomplicated Salmonella gastroenteritis, the determination of antimicrobial resistance patterns is often valuable for surveillance purposes [393]. In addition to the extraordinary resistance to antimicrobials, Salmonella proteins can have different functions depending on their location [392]. This may increase its ability to evade the response of the immune system of its huge variety of hosts. Therefore, dealing with time-spatial dependant networks may add an extra level of information. The prediction of MRs involved in the cellular response to the invasion by *Salmonella* can be very useful to address this dynamic behaviour and, if applied to different species, may unveil communalities that can be, eventually, targeted by gene-specific drugs.

The network pharmacology approach for studying pharmaco-dynamic drug-drug interactions (see Chapter 5), as stated, reflects the expectations, in

terms of scoring, in the case of additive and antagonistic drug-drug interactions. Its current high potential will be definitely increased by the adoption of next generation sequencing technologies for collecting drug-specific genetic signatures and the consequent increasing completeness of the related databases. Then, by tuning the parameters of the algorithm, it will be possible to reflect the drug combination proportions, better reflecting the definition of drug synergism. On the long run, being able to calculate synergistic drug combinations might derive into treatments combining lower doses of multiple drugs and the reduction of their individual side effects.

All the studies included in this book derive from data on cellular populations. Latest development in biotechnology have seen the born of single cell RNA-seq. This will allow the characterization of the genetic profiles of groups of cells. In the case of a disease or in response to a treatment, it would be interesting to monitor the behaviour of a determined group of cells. For example it may happen that, depending on the type of cell, some MRs are activated while others are not. This technology has the potential, if applied to each individual, to be the focus around which the personalized medicine will orbit.

Computational approximation to biomedical research cannot supplant experimental analysis as of today. Regardless, with full species genomes being characterized at an increasing rate and the growing interest towards personalized medicine, it is clear that experimental approaches are not going to be able to keep up with the amount of new data they themselves are generating. Thus, the development of new methods and analytical pipelines, such as the ones I describe in this work, is meant to become a requisite towards modern research protocols that will depend on the constant feedback between experimentally validated data, computationally derived analysis and prediction of new targets of interest. If computational analysis is

precise and accurate, it will considerably speed up the scientific progress, thus the technological one; if not this development process is destined to stall.

6.1.1 Future Perspectives

A GRM within a cell is a combination of DNA regions indirectly interacting with each other via their RNAs and protein expression products, thereby determining the cell's gene expression. As mentioned in the introduction of this thesis, apart from the information on the TFs binding sites, other elements are fundamental for the creation of this type of networks. Among them, two are of special relevance: i) the causal link between the activity of the TF and the expression of the regulated gene, and ii) knowledge on the expression of the TF in space and time. The availability of small to medium size GRNs is testimony that this is not a trivial problem. The low specificity TFs have and the lack of available TFs profiles [94], together with the few causal conditions tested are among the causes for such lack of information. An additional level of complication comes from the cooperation between TFs in the formation of enhanceosomes, which may include members of the complex located far away from the gene promoter region [409]. From an experimental point of view, new technologies are being proposed, like ChIA-pet and HI-C, to study chromatin interactions. This will expand our knowledge on distal interactions between TFs and the promoters of their regulatory targets and opens new bioinformatics challenges for the integration of this information in the current methodologies.

The spatial-temporal knowledge will certainly be fundamental for a better understanding of certain cellular mechanisms. In fact the cost of high-throughput sequencing is decreasing, and more time courses experiments are being devised in order to study how a system changes through time. Eventually, this understanding will be achieved combining the newly derived GRNs with any other type of network to which the same principles can be applied: metabolic networks and signalling pathways just to name a few.

In the light of this technological rush, in a close future, out of the proposed *Salmonella* infection analysis (chapters 3 and 4), I will release an automatized version of the code implemented. Its possible future expansions may include, for example, the use of ChIP-seq. data. I made predictions on TFs and their binding motifs but being able to experimentally test a set of TFs with great accuracy, speed and at a reasonable price would increase the strength of the presented approach. Current limits to this are represented by the cost that a large-scale analysis, on many TFs, could reach, specifically for testing all the needed antibodies. Another possible improvement for the methodology implemented could be including in the analysis non-coding RNAs. This aspect has been excluded because of the microarray limitation but, when a disease occurs, non coding RNAs have been reported to be strong sources of regulation and immunity [410], specially in the case of viral infections [411].

Improvements, in a short-medium term future, of the approach for studying pharmaco-dynamic drug-drug interaction (see Chapter 5) include one more comparison, between the different types of interactions (synergistic, antagonistic and additive), in terms of common putative MRs among the positively scored nodes after the message-passing algorithm. This would give one more hint on the involvement of common paths between the two drugs studied. We are also planning to analyse directly the raw data contained in the cmap database [314]. The aim of such an effort would be to be able to determine specifically for each gene not only the differential expression, but if this happens to be an up-regulation or a down-regulation. This could represent additional information for the method that could help in discerning between synergistic and antagonistic effect.

Community experiments, called challenges, are a very good opportunity for testing one's algorithm performances as blind tests (to avoid any bias or

over-fitting) are provided. The DREAM initiative [14] is currently organizing a series of systems biology competitions. These kind of initiatives have demonstrated that the collection of outcomes from a variety of participants leads to robust and top-performing final results. In this regard, I report two submitted papers derived from my collaboration in the participation to previous DREAM Challenges in the Appendix of this thesis. By chance, one of the challenges currently opened is the AstraZeneca-Sanger drug combination prediction DREAM challenge. It has the aim to understand effective combination treatments and drug synergy through the use of baseline genomic data. The core of the challenge is the release of about ~11.5k experimentally tested drug combinations measuring cell variability over 119 drugs and 85 cancer cell lines, and monotherapy response data for each drug and cell line. In addition, gene expression data, mutations, copy number alterations and methylation data will be provided. This appears to be the perfect benchmark framework to test the predictive capabilities of my method for the study of drug-drug interactions. To this end, I've already enrolled myself in the challenge. I am currently preparing the software to deal with the provided data and I expect to have the first assessment of the double-blinded predictions on drug-drug interactions as soon as possible.

During the whole development of my thesis I dug into many topics of the system biology field. This framework relies on the very basic assumption that many different biological phenomena are interwoven and uses a network graphical representation for the elements participating in such phenomena and the connections between them. In order to find biological paths activating MRs, I used many tools that integrate data from many different sources or collect algorithms for the most diverse computations. I would like to conclude this discussion by pointing that I strongly appreciate that and I am very grateful to their creators, especially when they come with a good user manual. Not only because it shortens the necessary time to

complete the analysis, but also because it reflects the effort of the scientific community towards a universal knowledge.

7 CONCLUSIONS

The main achievements of the work presented in this thesis are:

- Main regulators (MRs) of a set of genes with similar behaviour (derived from microarray experiments) were identified by integrating DISPOM's [82] predictions and information on specific transcription factors DNA binding sites (collected from JASPAR [84] and CIS-BP [94]).
- The combination of the gene regulatory network (GRN), derived from the prediction MRs-regulated genes, with protein-protein interactions (PIN) was achieved using a bespoken unification protocol in the BIANA framework [199].
- The application of a message-passing algorithm, from the predicted MRs into the GRN+PIN network, was determined to be a successful strategy to identify the regulatory elements (**who**) and the molecular basis (**how**) in *Salmonella spp.* infection.
- Using the developed strategy it was identified a set of key proteins, or MRs, in *Arabidopsis thaliana* (a small of transcriptions factors belonging to the same family) that play a central role in the infection process by *Salmonella spp.* The predicted set of MRs were subsequently experimentally validated and proved their importance at early stages of the infection.
- In the case of the *Salmonella spp.* – *Homo sapiens* infection process, predicted MRs were confirmed by comparing them to drug-specific gene signatures collected from the cmap database [4]: the top ranking drugs for the given MRs were mainly antimicrobials.
- The same strategy was adapted and used to study pharmacodynamics drug-drug interactions. Initial encouraging results suggest that the strategy could be useful to identify synergistic and antagonistic mechanisms in drug-drug interactions. Further developments and future directions are proposed in this thesis.

8 APPENDIX

As mentioned earlier in this chapter, in the Appendix are reported two submitted manuscripts about previous participations in these challenges. In the HPN-DREAM 8 (see Appendix 8.1) breast cancer network inference challenge my contribution consisted in analysing the microarray data provided and in the computation of cross-correlations among the profiles of differentially expressed genes that were later used as features for the Random Forest classifier used to infer the network. In the DREAM 8.5 rheumatoid arthritis responder challenge (see Appendix 8.2), my contribution consisted in the normalisation and analysis of the microarray data provided by the organizers.

8.1 Empirical assessment of causal network inference through a community-based effort

Steven M. Hill1,*, Laura M. Heiser2,*, Thomas Cokelaer3, Michael Unger4, Nicole K. Nesser19, Dan Carlin5, Yang Zhang6, Artem Sokolov5, Evan Paull5, Chris K Wong5, Kiley Graim5, Adrian Bivol5, Haizhou Wang6, Fan Zhu7, Bahman Afsari10, Ludmila V. Danilova10, Alexander V. Favorov10,11,12, Wai-shing Lee10, Dane Taylor13,14, Chenyue W. Hu15, Byron L. Long15, David P. Noren15, Alexander Bisberg15, HPN-DREAM Consortium, Gordon B. Mills16, Joe W. Gray2,17,18, Michael Kellen19, Thea Norman19, Stephen Friend19, Amina A. Qutub15, Elana J. Fertig10, Yuanfang Guan7,8,9, Mingzhou Song6, Joshua Stuart5, Paul T. Spellman20, Heinz Koeppl4, Gustavo Stolovitzky21,^, Julio Saez-Rodriguez3,^, Sach Mukherjee1,22,^

1. MRC Biostatistics Unit, Cambridge Institute of Public Health, Cambridge CB2 0SR, UK
2. Department of Biomedical Engineering, Oregon Health and Science University, Portland, OR, USA
3. European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton CB10 1SD, UK
4. Automatic Control Laboratory and Institute of Biochemistry, ETH Zurich, 8092 Zurich, Switzerland
5. Biomolecular Engineering, UC Santa Cruz, Santa Cruz, CA, USA
6. Department of Computer Science, New Mexico State University, Las Cruces, NM, USA
7. Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, USA
8. Department of Internal Medicine, University of Michigan, Ann Arbor, Michigan, USA
9. Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, USA
10. Department of Oncology, Division of Biostatistics and Bioinformatics, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, Baltimore, MD, USA
11. Department of Computational Systems Biology, Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, Russia
12. Laboratory of Bioinformatics, Research Institute of Genetics and Selection of Industrial Microorganisms, Moscow, Russia
13. Statistical and Applied Mathematical Sciences Institute, Research Triangle Park, NC, USA
14. Department of Mathematics, University of North Carolina, Chapel Hill, NC, USA
15. Rice University, Department of Bioengineering, 6500 Main St. Room 613, Houston, TX, USA, 77030
16. Department of Systems Biology, MD Anderson Cancer Center, Houston, TX, USA
17. Center for Spatial Systems Biomedicine,
18. Knight Cancer Institute, Oregon Health and Science University, Portland, OR, USA
19. Sage Bionetworks, Seattle, WA, USA
20. Department of Molecular and Medical Genetics, Oregon Health and Science University, Portland, OR, USA
21. IBM Translational Systems Biology and Nanobiotechnology, Yorktown Heights, NY 10598, USA
22. Cambridge Institute, School of Clinical Medicine, University of Cambridge, Cambridge CB2 0RE, UK

Haizhou Wang Present address: SimQuest Inc, Boston, MA, USA
Yang Zhang Present address: Amyris Inc, Emeryville, CA, USA

Publicat com:

Hill SM, Heiser LM, Cokelaer T, Unger M, Nesser NK, Carlin DE, Zhang Y, Sokolov A, Paull EO, Wong CK, Graim K, Bivol A, Wang H, Zhu F, Afsari B, Danilova LV, Favorov AV, Lee WS, Taylor D, Hu CW, Long BL, Noren DP, Bisberg AJ; HPN-DREAM Consortium, Mills GB, Gray JW, Kellen M, Norman T, Friend S, Qutub AA, Fertig EJ, Guan Y, Song M, Stuart JM, Spellman PT, Koeppl H, Stolovitzky G, Saez-Rodriguez J, Mukherjee S. Inferring causal molecular networks: empirical assessment through a community-based effort. Nat Methods. 2016 Apr;13(4):310-8. doi: 10.1038/nmeth.3773.

8.2 Crowdsourced assessment of genetic contribution to predicting anti-TNF treatment response in rheumatoid arthritis

Organizers: Solveig K. Sieberts, Eli Stahl, Abhishek Pratap, Gaurav Pandey, Dimitrios Pappas, Jing Cui, Andre O. Falcao, Christine Suver, Bruce Hoff, Venkat S.K. Balagurusamy, Donna Dillenberger, Elias Chaibub Neto, Thea Norman, Stephen Friend, Robert Plenge, Gustavo Stolovitzky, Lara M. Mangravite

Solvers: Fan Zhu, Javier García-García, Daniel Aguilar, Bernat Anton, Jaume Bonet, Ridvan Eksi, Oriol Fornés, Emre Guney, Hongdong Li, Manuel Alejandro Marín, Bharat Panwar, Joan Planas-Iglesias, Daniel Poglayen, Tero Aittokallio, Muhammad Ammad-ud-din, Chloe-Agathe Azencott, Víctor Bellón, Valentina Boeva, Kerstin Bunte, Himanshu Chheda, Lu Cheng, Jukka Corander, Michel Dumontier, Anna Goldenberg, Peddinti Gopalacharyulu, Mohsen Hajiloo, Daniel Hidru, Alok Jaiswal, Samuel Kaski, Beyrem Khalfaoui, Suleiman Ali Khan, Eric R Kramer, Pekka Marttinen, Aziz M. Mezlini, Bhuvan Molparia, Matti Pirinen, Janna Saarela, Matthias Samwald, Véronique Stoven, Hao Tang, Jing Tang, Ali Torkamani, Jean-Phillipe Vert, Bo Wang, Tao Wang, Krister Wennerberg, Nathan E. Wineinger, Guanghua Xiao,

Yang Xie, Rae Yeung, Xiaowei Zhan, Cheng Zhao, The Rheumatoid Arthritis Challenge

Consortium, Baldo Oliva, Yuanfang Guan

Data Contributors: Jeff Greenberg, Joel Kremer, Kaleb Michaud, Anne Barton, Marieke Coenen, Xavier Mariette, Corinne Miceli, Nancy Shadick, Michael Weinblatt, Niek de Vries, Paul P. Tak, Danielle Gerlag, Tom W. J. Huizinga, Fina Kurreeman, Cornelia F. Allaart, S. Louis Bridges Jr., Lindsey Criswell, Larry Moreland, Lars Klareskog, Saedis Saevarsdottir, Leonid Padyukov, Peter K. Gregersen, Robert Plenge

Solveig K. Sieberts[1]*, Fan Zhu[2]*, Javier García-García[3]*, Eli Stahl[4,5], Abhishek Pratap[1], Gaurav Pandey[5], Dimitrios Pappas[6,7], Daniel Aguilar[3], Bernat Anton[3], Jaume Bonet[3], Ridvan Eksi[2], Oriol Fornés[3], Emre Guney[8], Hongdong Li[2], Manuel Alejandro Marín[3], Bharat Panwar[2], Joan Planas-Iglesias[3], Daniel Poglayen[3], Jing Cui[9], Andre O. Falcao[10], Christine Suver[1], Bruce Hoff[1], Venkat S. K. Balagurusamy[11], Donna Dillenberger[11], Elias Chaibub Neto[1], Thea Norman[1], Tero Aittokallio[12], Muhammad Ammad-ud- din[13,14], Chloe-Agathe Azencott[15,16,17], Víctor Bellón[15,16,17], Valentina Boeva[15,16,17], Kerstin

Bunte[13,14], Himanshu Chheda[12], Lu Cheng[12,13,14], Jukka Corander[14,18], Michel Dumontier[19], Anna Goldenberg[20,21], Peddinti Gopalacharyulu[12], Mohsen Hajiloo[21], Daniel Hidru[20,21], Alok Jaiswal[12], Samuel Kaski[13,14,22], Beyrem Khalfaoui[21], Suleiman Ali Khan[12,13,14], Eric R. Kramer[23], Pekka Marttinen[13,14], Aziz M. Mezlini[20,21], Bhuvan Molparia[23], Matti Pirinen[12], Janna Saarela[12], Matthias Samwald[24], Véronique Stoven[15,16,17], Hao Tang[25], Jing Tang[12], Ali Torkamani[23], Jean-Phillipe Vert[15,16,17], Bo Wang[26], Tao Wang[25], Krister Wennerberg[12], Nathan E. Wineinger[23], Guanghua Xiao[25], Yang Xie[25,27], Rae Yeung[28,29], Xiaowei Zhan[25,30], Cheng Zhao[20,21], The Rheumatoid Arthritis Challenge Consortium, Jeff Greenberg[7,31], Joel Kremer[32], Kaleb Michaud[33,34], Anne Barton[35,36], Marieke Coenen[37], Xavier Mariette[38,39], Corinne Miceli[38,39], Nancy Shadick[9], Michael Weinblatt[9], Niek de Vries[40], Paul P. Tak[40,41,42,43], Danielle Gerlag[40,44], Tom W. J. Huizinga[45], Fina Kurreeman[45], Cornelia F. Allaart[45], S. Louis Bridges Jr.[46], Lindsey Criswell[47], Larry Moreland[48], Lars Klareskog[49], Saedis Saevarsdottir[49], Leonid Padyukov[49], Peter K. Gregersen[50], Stephen Friend[1], Robert Plenge[51], Gustavo Stolovitzky[5,11], Baldo Oliva[3^], Yuanfang Guan[2^], Lara M. Mangravite[1^]

Affiliations

[1] Sage Bionetworks, Seattle, Washington, USA.

[2] Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, USA.

[3] Structural Bioinformatics Group (GRIB/IMIM), Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona, Catalonia, Spain.

[4] Center for Statistical Genetics, Division of Psychiatric Genomics, Icahn School of Medicine at Mount Sinai, New York, New York, USA.

[5] Icahn Institute for Genomics and Multiscale Biology and Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York, USA.

[6] College of Physicians and Surgeons, Columbia University, New York, New York, USA.

[7] Corrona LLC, Southborough, Massachusetts, USA.

[8] Center for Complex Network Research. Northeastern University and Center for Cancer Systems Biology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA

[9] Division of Rheumatology, Immunology, and Allergy, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA.

[10] Department of Informatics. Faculty of Sciences. University of Lisbon, Lisbon, Spain.

[11] IBM T.J.Watson Research Center, Yorktown Heights, New York, USA.

[12] Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland.

[13] Department of Computer Science, Aalto University, Espoo, Finland.

[14] Helsinki Institute for Information Technology (HIIT), Esbo, Finland.

[15] MINES ParisTech, PSL-Research University, CBIO-Centre for Computational Biology, Fontainebleau, France.

[16] Institut Curie, Paris Cedex ,France.

[17] INSERM U900, Paris Cedex, France

[18] Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland.

[19] Stanford Center for Biomedical Informatics, Stanford University, Stanford, CA, USA.

[20] Department of Computer Science, University of Toronto, Toronto, ON, Canada.

[21] Genetics & Genome Biology, SickKids Research Institute, Toronto, ON, Canada.

[22] Department of Computer Science, University of Helsinki, Helsinki, Finland.

[23] The Scripps Translational Science Institute and Department of of Integrative Structural and Computational Biology The Scripps Research Institute, La Jolla, California, USA.

[24] Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Vienna, Austria.

[25] Quantitative Biomedical Research Center, University of Texas Southwestern Medical Center, Dallas, Texas, USA.

[26] Department of Computer Science, Stanford University, Stanford, California, USA.

[27] Harold C. Simmons Comprehensive Cancer Center, University of Texas Southwestern Medical Center, Dallas, Texas, USA.

[28] Department of Paediatrics, Department of Immunology, Institute of Medical Sciences, University of Toronto, Toronto, Ontario, Canada.

[29] Cell Biology, SickKids Research Institute, Toronto, Ontario, Canada.

[30] Center for the Genetics of Host Defense, University of Texas Southwestern Medical Center, Dallas, Texas, USA.

[31] New York University School of Medicine, New York, New York, USA

[32] Albany Medical College, Albany, New York, USA.

[33] University of Nebraska Medical Center, Omaha, Nebraska, USA.

[34] National Data Bank for Rheumatic Diseases, Wichita, Kansas, USA.

[35] Arthritis Research UK Centre for Genetics and Genomics, Centre for Musculoskeletal Research, Manchester Academic Health Sciences Centre, The University of Manchester, UK.

[36] NIHR Manchester Musculoskeletal Biomedical Research Unit, Central Manchester Foundation Trust, Oxford Road, Manchester, UK.

[37] Department of Human Genetics, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands.

[38] Université Paris-Sud, Orsay, France

[39] APHP–Hôpital Bicêtre, Center of Immunology of Viral Infections and Autoimmune Diseases (IMVA) INSERM U1184, Paris, France.

[40] Department of Clinical Immunology and Rheumatology, Academic Medical Center/University of Amsterdam, Amsterdam, The Netherlands.

[41] Cambridge University, Cambridge, UK.

[42] Ghent University, Ghent, Belgium.

[43] Glaxo Smith Kline, Stevenage, UK.

[44] Clinical Unit, GlaxoSmithKline, Cambridge, UK.

[45] Department of Rheumatology, Leiden University Medical Centre, Leiden, The Netherlands.

[46] Division of Clinical Immunology and Rheumatology, Department of Medicine, University of Alabama at Birmingham, Birmingham, Alabama, USA.

[47] Rosalind Russell / Ephraim P Engleman Rheumatology Research Center, Division of Rheumatology, Department of Medicine, University of California San Francisco, San Francisco, California, USA.

[48] Division of Rheumatology and Clinical Immunology, University of Pittsburgh, Pittsburgh, Pennsylvania, USA.

[49] Rheumatology Unit, Department of Medicine, Karolinska Hospital and Karolinska Institutet, Solna, Sweden.

ADDITIONAL TITLE PAGE FOOTNOTES: (includes description of co-first authors)

* These authors contributed equally to this work

^ These authors contributed equally to this work

8.2.1 Abstract

Although one-third of RA patients fail to enter clinical remission following anti-TNF treatment, no algorithm currently exists to accurately predict likelihood of response. This study was designed as a blind open-science competition to test whether disease-lowering anti-TNF-a response could be predicted based on genetic, demographic and clinical information. 73 teams submitted predictions (top predictor, AUROC=0.62 and an AUPR=0.51). Given this relatively low predictive performance, a collaborative effort across top performing teams was conducted to formally assess the contribution to performance of genetic information. Despite a significant genetic heritability estimate of the treatment non-response trait (h^2 =

0.18, p-value = 0.02), no statistical differences were observed between models that incorporated rational SNP selection relative to models containing no genetic information, indicating that current algorithms are not able to effectively leverage polygenic signal for prediction. As such, future efforts toward predicting anti-TNF efficacy should focus on use of clinical and biomarker measures

8.2.2 Introduction

Rheumatoid arthritis (RA), a chronic inflammatory disorder affecting synovial joints, is treated with disease-modifying antirheumatic drugs, including those that block the inflammatory cytokine, tumor necrosis factor-α (anti-TNF therapy). However, response to anti-TNF therapy is variable with nearly one-third of RA patients failing to enter clinical remission[1,2]. Although clinical and lifestyle predictors of RA disease risk have been developed that are of sufficient accuracy to merit clinical consideration[3–5], no algorithm currently exists to accurately assess likelihood of response prior to treatment[6]. Technological advances in DNA genotyping and sequencing have afforded the opportunity to assess the contribution of genetic variation to heterogeneity of RA response to therapy. Meta-analysis across patient cohorts has identified many loci associated with RA disease risk independent of treatment[7] and several loci associated with therapeutic response to anti-TNF agents[8] have been identified. Here we leverage this information to assess whether common genetic variation can be used to predict anti-TNF treatment response.

This project leveraged the DREAM framework[9–12] to implement a blind, open, community-based analysis. This framework, which has been successfully implemented to solve a range of modeling problems, provides a formalized and rigorous mechanism to compare performance across independent methods and has demonstrated that, under the right conditions, the most robust predictions are developed by combining solutions across multiple complementary methods[13–15].

We further extend the competition-based DREAM framework, to include a collaborative analysis from the top performing teams in order to determine whether predictions can be improved by creating ensemble predictions of multiple methods, and to determine the degree to which genetic predictors contribute to the accuracy of treatment response predictions.

Specifically, this analysis was designed as a two-phase experiment (Fig. 1A). Algorithms best suited to address this question were identified within the Competitive Phase, which ran as a traditional DREAM challenge, by inviting research teams across the world to compete to build accurate predictions of anti-TNF response. Two tasks or "sub-challenges" were presented to the partipants: the "classification" sub-challenge, in which participants were asked to predict whether or not a patient would respond to treatment, and the "quantitative response" sub- challenge, in which participants were asked to predict the quantitative change in disease severity after treatment. Those teams with the most accurate models were brought together in a subsequent Collaborative Phase to conduct a comparative analysis that formally assessed the contribution of genetic information to accuracy in predicting anti-TNF treatment effects.

8.2.3 Results

Genetic analyses were conducted using whole genome SNP data derived from two cohorts: a primary 2,706 anti-TNF treated RA patients combined across 13 collections of European ancestry[8] and 591 patients in the CORRONA CERTAIN study[16]. Treatment efficacy was measured using the absolute change in disease activity score in 28 joints[17] (DAS28) following 3-6 months of anti-TNF treatment. Data from the two cohorts were harmonized, and the resultant data have been made publicly available as a resource for use by the research community (https://www.synapse.org/#!Synapse:syn3280809). Significant SNP-heritability was estimated for change in DAS28 (ΔDAS28), via variance component modeling (VCM)[18,19], within the 2,706 patient, primary cohort (SNP-h^2=0.18, p=0.02, Fig. 1B). Heritability estimates were strongest in the subset of patients treated with anti-TNF monoclonal antibodies relative to those treated with the circulating biologic, enteracept (Fig. 1B). These

heritability estimates are similar to those reported for other treatment response traits[20] and suggest that there may be sufficient genetic contribution to anti-TNF therapeutic response to support the use of predictive modeling methods to identify polygenic predictors of response[21,22].

The primary endpoint used throughout both phases of the challenge was the classification of nonresponse to anti-TNF therapy as defined by EULAR response criteria[23], a categorical definition based on DAS28 that is used widely in clinical practice. As a secondary endpoint, participants were also challenged to predict ΔDAS28 as a continuous measure. Throughout both phases, participants trained models using a data set containing whole genome SNP data, age, gender, anti-TNF therapy, concomitant methotrexate treatment, and baseline DAS28 in a subset of 2,031 individuals from the primary cohort (Fig. 1C, Supplementary Table 1 and Methods)[8]. Participants were provided with a leaderboard, which evaluated the performance of their predictions with real-time feedback, relative to the remaining 675 individuals from that cohort. To reduce the potential for overfitting or reverse- engineering of treatment outcomes from the leaderboard, each team was limited to 100 leaderboard submissions. Final evaluation of algorithms was conducted relative to the separate test dataset consisting of data collected from the CORRONA CERTAIN[16] study. Participants remained blinded to outcomes from both the leaderboard and test data sets throughout the experiment.

The competitive phase of the challenge attracted 242 registered participants representing 30 countries and 4 continents. 73 teams submitted a total of 4874 predictions for evaluation on the leaderboard data over the course of the 16-week training period. After that time, final models were formally compared based on their performance in the test data, and teams were allowed up to 2 final submissions per challenge endpoint. For the classification challenge, models were scored using both the area under the

receiver-operator curve (AUROC) and the area under the precision-recall curve (AUPR). Ultimately, 27 final submissions were received from 15 teams. Overall rank for each submission was determined as the average of the AUROC rank and the AUPR rank among all valid submissions. AUROC and AUPR were interpolated in the case of binary classifications or in the case of tied predictions[24]. Of 27 submissions, 11 performed significantly better than random for both AUPR and AUROC after Bonferroni correction for multiple submissions. The AUPR of all submissions ranged from 0.345 to 0.510 (null expectation 0.359), and the AUROC ranged from 0.471 to 0.624 (null expectation 0.5). Using bootstrap analysis of submission ranks (Fig. 2A), we determined that the top two submissions performed robustly better than all remaining solutions (Wilcoxon signed-rank test of bootstraps p-value = 5e-34 and 1e-66, relative to the third ranked submission, respectively) but were not distinct from one another (p-value = 0.44). These submissions had AUPR of 0.5099 and 0.5071 and AUROC of 0.6152 and 0.6237, respectively. Both of these teams used Gaussian Process Regression (GPR)[25] models but they differed in their implementation (see 'Team Guanlab' and 'Team SBI_Lab' in the Supplementary methods for more details). The code and provenance for the winning algorithms have been cataloged and made available for reuse and are available through the challenge website. Team Guanlab selected SNP predictors based on the training data and previous analyses described in the literature and applied a GPR model to predict non-response classification directly. Team SBI_Lab selected SNP predictors using only the training data, applied a GPR model to predict ΔDAS28, and refactored these predictions into classification weights. For the quantitative sub challenge, 28 final models from 17 teams were received that predicted ΔDAS28 as a continuous measure. In this case, performance was evaluated based on correlation between predicted and observed ΔDAS28 (observed range: r = 0.393 to -0.356). Of these, 18 submissions performed significantly better than random (r= 0.393 to 0.208), and the top performing submission was robustly better than all remaining

solutions (p-value = 2e-32 relative to the 2nd ranked submission, Supplementary Fig. 1). The winning model used a similar GPR model to predict ΔDAS28 as described above (see 'Team Guanlab' in the Supplementary methods for more details).

The Collaborative Phase leveraged the top performing algorithms to formally assess the contribution of genetic information to model performance. The 8 teams with the best predictive performances (7 in each subchallenge) from this competitive phase were invited to participate. This was motivated in part by the narrow range and low predictive performance observed across submitted predictions in the competitive phase, suggesting that genetic variation was not substantially contributing to predictions. First, a direct comparison of models built in the presence and absence of genetic information was performed. For this analysis, each team developed a pair of predictions based on a model built in the presence of genetic information (genetic model) and a model built using only clinical and demographic covariates (non-genetic model). Pairwise comparison across models demonstrated that there was no statistical difference between the non-genetic and genetic models (paired t-test p- value = 0.85, 0.82, for classification AUPR and AUROC, respectively, and p-value = 0.65 for continuous prediction correlation) (Fig. 2B, Supplementary Fig. 2). These results indicated that, while there may be weak underlying genetic contribution to treatment effect, such genetic effects had no detectable contribution to predictive performance. To assess the ability of modeling techniques to detect weak genetic contribution, a comparative analysis was performed between genetic models built with researcher-selected SNP sets - guided by prior biological knowledge and data-driven SNP selection - relative to 100 random SNP sets of equivalent size. For 5 of 7 classification algorithms, models using knowledge-mined SNPs significantly outperformed models using random SNPs for AUPR, AUROC or both (enrichment p- value=3.3e-05) (Fig. 2C). This suggests that there is a non-zero contribution of genetic information to treatment effect even if it is not of sufficient magnitude to

have a significant contribution to modeling performance. No relationship between modeling algorithm selection and performance was observed.

The challenge framework provided several advantages over traditional predictive modeling approaches. First, comparison with an independent, blinded test dataset reduced the contribution to estimated accuracy of overfitting to the training dataset, as indicated by comparing predictive performance between leaderboard and test data predictions for both the PR and AUROCs (Supplementary Fig. 3). Of note, for the quantitative subchallenge, the correlation between the leaderboard and final scores was negative (pearson correlation = -0.052) suggesting the presence of widespread overfitting. Second, the use of a diverse set of methodological approaches across teams provided the opportunity to assess whether performance was more robust using an ensemble approach to combine information across submissions. In previous DREAM challenges, unsupervised ensemble predictions have been demonstrated to perform as well as or better than top performing team submissions[9]. In this challenge, ensemble analysis was performed using a supervised approach to leverage the diversity across the submitted predictions[26]. Ensemble models were trained for the classification subchallenge using leave-one- out cross-validated (LOOCV) predictions generated on the original training set using the individual methods, and, as with individual submissions, analyzed in a blinded fashion using the test data. The first principal component (PC) discovered by applying supervised PC analysis[27] to these training LOOCV classifications significantly separated responders from non-responders (Wilcoxon rank-sum p- value=5.40e-62), thus indicating that learning a supervised model over these submitted classifications can help boost discriminative/predictive power. Motivated by this observation, two separate ensemble classifications were developed using the stacking method[28,29] from the class of heterogeneous ensemble learning methods[26] and both performed well: the first was based on LOOCV predictions during the

competitive phase (AUPR=0.5228, AUROC=0.622) and the second based on LOOCV predictions in the collaborative phase (AUPR=0.5209, AUROC=0.6168). Compared to the 7 collaborative phase models, these ensemble models ranked first and third, respectively (Supplementary Fig. 4). These results indicate that the individual classifications provided some complementary information that, when appropriately aggregated, improved classifications. Supervised ensembles also substantially outperformed predictions developed using an unsupervised ensemble[30] (Competitive Phase Unsupervised Ensemble Predictor: PR=0.415, ROC=0.575, Wilcoxon signed-rank test of bootstraps p-value =5.3e-167; Collaborative phases Ensemble Predictor: PR=0.415, ROC=0.576, Wilcoxon signed-rank test of bootstraps p-value =8.7e-167). The capability of ensemble methods to provide predictions with highly ranked relative accuracy supports the use of this approach to boost predictive performance in situations where analyses are performed in the absence of a gold standard with which to identify the best performing individual methodology to use. Despite this, these predictions need to be further improved to merit consideration in clinical care.

8.2.4 Discussion

Although theoretical heritability estimates for polygenic models indicate significant genetic contribution to variation in treatment response, current predictive algorithms are not able to translate this estimated signal into practical predictions likely due to the complex nature of the genetic contribution. Future studies with larger sample sizes would provide the opportunity to detect smaller SNP effects as well as include more complex relationships amongst the data including epistasis. The most effective approaches explicitly modeled drug-specific genetic signal, suggesting that

there may be heterogeneity in response mechanisms across different anti-TNF drugs. We anticipate that there is also heterogeneity across the patient population that is not genetic in nature and was not fully captured by the clinical information available, which was limited relative to the breadth of information typically available within clinical practice. Given that the most successful methodologies employed in this study used within-group predictions based on stratified subsets of patients, we anticipate that data modalities – clinical, molecular, or other - that capture the heterogeneity in RA disease progression will provide more accurate predictive information[13].

The adoption of predictive algorithms within clinical trial enrollment provides a powerful mechanism to reduce heterogeneity, to increase statistical power to observe positive outcomes and/or to decrease study size. Although genetic information did not provide a meaningful contribution to the predictions in this study, these methods were able to leverage the small set of available clinical features to develop a prediction that performed significantly better than random.

Incorporation of additional clinical information - including seropositivity, treatment compliance, and disease duration - may provide the best opportunity to leverage these methods in clinically meaningful ways.

8.2.5 Methods

8.2.5.1 Datasets

Two separate data sets were provided to participants to train and test the predictive models, respectively (Table 2). In the case of the test data, only predictor variables were released, and the teams remained blinded to the response variables. The training data consisted of a previously published collection of anti-TNF treated patients (n=2,706) of European ancestry,

compiled from 13 collections[8], of which the response variables from 675 patients were held out as a leaderboard test set. All patients met 1987 ACR criteria for RA, or were diagnosed by a board-certified rheumatologist and were required to have at least moderate DAS28[17] at baseline (DAS28>3.2). Available clinical and demographic data included DAS28 at baseline and at least one time point after treatment, gender, age, anti-TNF drug name and methotrexate use. Follow-up DAS28 was measured 3–12 months after initiating anti-TNF therapy, though precise duration of treatment was not available. Genotypes for each sample were imputed to HapMap Phase 2 (release 22) as previously described[8]. We note that although this dataset does not represent the full spectrum of patient information that may be utilized within a clinical setting to inform treatment - including synovial tissue and novel soluble biomarkers like MRP8/14 levels[2,31,32], it did present sufficient data to explicitly assess the contribution of genetics to prediction.

The final test set was derived from a subset of patients enrolled in the CORRONA CERTAIN study[16]. CERTAIN is a prospective, non-randomized cohort study of adult patients with RA fulfilling the 1987 ACR criteria, having at least moderate disease activity defined by a clinical disease activity index (CDAI) score>10 who are starting or switching biologic agents. DAS28 was provided at baseline and 3 month follow-up. At the time of challenge launch, 723 subjects had initiated anti-TNF therapy and had a 3 month follow-up visit. Of these patients, 57.4% were previously naïve to biologics. Genotypes were generated on the Illumina Infinium HumanCoreExome array and imputed to HapMap Phase 2 (release 22) using IMPUTE2[33]. While data for all 723 were released to participants, 93 patients were excluded for the purposes of scoring because their genotyping data were not consistent with European ancestry. In addition, a subset of patients in the test data set were treated with anti-TNF drugs that were not represented in the training data set: golimumab and certolizumab. The 39 patients receiving golimumab were excluded because this drug was not

represented in the training data and predictions showed that participants were unable to successfully predict response in these subjects. In contrast, prediction in certolizumab-treated patients was similar to prediction in the remaining three drugs and so these data were included in the final test set.

Two ancillary datasets were made available for participant use. The first measured TNFα protein level in HapMap cell lines [34]. The second included blood RNA-seq data and genotypes for 60 RA patients from the Arthritis Foundation-sponsored Arthritis Internet Registry (AIR), 30 who displayed high inflammatory levels and 30 who displayed low inflammatory levels. Inflammatory levels were assessed using blood concentrations of C-reactive protein (CRP), and elevated disease was defined as CRP greater than 0.8 mg/dL, while low disease activity was defined as CRP less than 0.1 mg/dL. In addition to CRP levels, rheumatoid factor (RF) antibody levels, and cyclic citrullinated peptide (CCP) levels were also assayed. Genotypes were assayed on the Illumina HumanOmniExpressExome array.

Data use within the scope of this challenge was performed with the approval of an Internal Review Board for all data sets. All four data sets can be assessed through the Synapse repository (syn3280809, doi:10.7303/syn3280809).

8.2.5.2 Scoring methods

For the classification subchallenge, teams were asked to submit an ordered list of patients ranked according to the predicted response to therapy. Special treatment was given to the computation of the curve statistics when the order was ambiguous such as in the case in the case of ties or binary predictions, in which case an average across all possible consistent solutions was used[24]. The average of the rank of the AUPR and AUROC was used to rank solutions.

For the quantitative subchallenge, teams were asked to submit predicted ΔDAS28, and the pearson correlation between the predicted and actual ΔDAS28 was used to score submissions.

8.2.5.3 Comparative phase challenge

The challenge was open to all individuals who agreed to the DREAM terms of use and obtained access to the challenge data by certifying their compliance with the Data Terms of Use. The training and ancillary data were released for use on February 10, 2014. The leaderboards opened on March 5, at which time participants were able to test their models in real-time against a held-out portion of the training dataset. The prediction variables of the test data set were released to participants on May 8 and submission queues for final submissions were open between May 21st and June 4th. Only the final two submissions per team per subchallenge were scored. Participants who didn't have enough computational resources in their home institutions were offered the option to use an IBM z-Enterprise cloud, with two virtual machines running Linux servers, one with 20 processors, 242 GB memory, 9 TB storage space and the other with 12 processors, 128 GB memory and 1 TB of storage space. Cloud users could access the Challenge data directly through the IBM system.

8.2.5.4 Evalutation of submissions

Predictions were evaluated using two data sets: 675 individuals from the training cohort (leaderboard test set) and all individuals from the CORRONA CERTAIN data (final test set). In both cases, response variables were withheld from participants. Participants were allowed 100 submissions to the classification sub-challenge leaderboard and unlimited submissions to the quantitative sub-challenge leaderboard throughout the competitive phase of the competition, and were provided near-instant results. Participants were

allowed 2 final submissions per sub-challenge and scores were revealed after the submission deadline. A permutation test was used to assess whether the classifications or ΔDAS28 quantitative predictions were better than expected at random using a one-sided p-value. In order to assess the robustness of the relative ranking of predictions, 1,000 bootstraps were performed by sampling subjects with replacement. Within each bootstrap iteration, evaluation scores were computed for each submission, along with the within-iteration rank. A prediction was deemed "robustly" better than another if the Wilcoxon signed-rank test of the 1000 bootstrap iteration estimates was significant with p-value < 0.05. While this is not the same as strict statistical significance, it was the criteria we used to differentiate models given the relatively small improvements from one to another.

8.2.5.5 Development and scoring in the collaborative phase

One of the aims of DREAM Challenges is to foster collaborative research. As such, the collaborative phase was designed to foster cooperation between the best performing teams in the competitive phase. Teams came together to develop research questions and analytical strategies to answer specific questions related to the ability to predict non-response to anti-TNF treatment. Each team submitted a number of classifications/predictions and/or sets of classifications/predictions that were designed to be able to answer questions about the degree to which genetic data were contributing to the models, and the classifications were scored and analyzed across teams by the challenge organizers. In order to compare across methods and approaches, we asked the collaborative phase participants to submit classifications/predictions using their own knowledge- and data-mined SNP lists, which they refined from the competitive phase after peer review from fellow participants. Additionally, they were asked to submit a non-genetic classification/prediction, which did not include genetic predictors. We also asked the participants to submit 100 sets of classifications/predictions in

which the SNPs used as potential predictors were randomly sampled from the genome and matched the number of SNPs in their genetic model. Eight teams participated in the collaborative phase, seven in each sub-challenge. Ranked results for the genetic models are shown in Supplementary Figure 4.

8.2.5.6 Ensemble classifications

The goal of ensemble learning was to aggregate the classifications submitted by individual teams to the classification subchallenge, including 6 from the Competitive Phase and 7 from the Collaborative Phase, by effectively leveraging the consensus as well as diversity among these predictions. We focused on learning heterogeneous ensembles[16], which are capable of aggregating classifications from a diverse set of potentially unrelated base classifiers, as is the case with the submissions to this subchallenge. Specifically, we followed the stacking methodology[28,29], which involves learning a meta-classifier (2^{nd} level predictor) on top of the base classifications. This methodology was applied to the training set classifications generated through a leave-one-out cross-validation (LOOCV) procedure applied to the training set for the initial ensemble learning. To address the potential calibration issue in this task[35], we investigated using the raw base classifications and the output of two other normalization procedures – Z-score (mean=0, std. dev.=1) and Scale0-1 (maximum=1, minimum=0) – applied to the raw base classifications. Next, sixteen different classification algorithms (Supplementary Table 3) were used to train ensemble models from each of the above normalized versions of the base classifications. The implementations of these algorithms were obtained from the Weka machine learning suite[36], and their default parameters were used.

Supplementary figure 5 shows the performance of different combinations of normalization and classification methods on the leaderboard test set in terms of (A) AUPR, (B) AUROC and (C) the overall rank. Several

observations can be made from these results. First, the ensemble learnt with normalization using Z-score and subsequent learning of a Naïve Bayes classifier that uses kernelized probability distribution functions[37] produced the best aggregate performance on the leaderboard test set (AUROC= 0.7569, AUPR=0.49), indicating the conditional independence of the base classifications and the non-normality of their underlying distributions. In general, normalization (either Z-score or Scale0-1) improved the performance for 14, 14 and 13 of the 16 classifiers examined in terms of PR, ROC and overall rank respectively, thus indicating the importance of effective calibration in such ensemble learning tasks. Of these, 10, 9 and 7 classifiers, including NaiveBayes_kdf, saw the best performance due to the use of Z-score normalization, thus giving this normalization method an edge over Scale0-1.

Based on the conclusions above, we applied the ensemble model trained using Z-score and NaiveBayes_kdf to the individual team classifications submitted for the CORRONA CERTAIN test set in the competitive and collaborative phases. The ensemble of the competitive phase (AUPR=0.5228, AUROC=0.622) performed better than each of the individual classifications and slightly better than the ensemble of the collaborative phase (AUPR=0.5209, AUROC=0.6168). These results indicate that it is indeed possible to modestly improve classifications for RA anti-TNF response by aggregating classifications submitted by individual teams to this subchallenge using supervised heterogeneous ensemble methods.

For comparison, we also generated unsupervised ensemble classifications using the Spectral Meta-Learner (SML) method[30]. Specifically, the binary input classifications needed for this method were obtained using the signs of the z-scored base classification.

8.2.6 Author contribution

The following authors contributed to organizing the challenge: LM Mangravite , SK Sieberts, G Stolovitzky, E Stahl, A Pratap, G Pandey, D Pappas, J Cui, AO Falcao, C Suver, T Norman, S Friend, R Plenge

The following authors contributed to data analysis: SK Sieberts, E Stahl, A Pratap, G Pandey, J Cui, AO Falcao, EC Neto

The following authors contributed to software and technical solutions for the challenge: A Pratap, B Hoff, VSK Balagurusamy, D Dillenberger

The following authors contributed data for the challenge: J Greenberg, J Kremer, K Michaud, A Barton, M Coenen, X Mariette, C Miceli, N Shadick, M Weinblatt, N de Vries, PP Tak, D Gerlag, TWJ Huizinga, F Kurreeman, CF Allaart, SL Bridges Jr., L Criswell, L Moreland, L Klareskog, S Saevarsdottir, L Padyukov, PK Gregersen, R Plenge

The following authors participated in the predictive modeling challenge: F Zhu, J García-García, D Aguilar, B Anton, J Bonet, R Eksi, O Fornés, E Guney, H Li, MA Marín, B Panwar, J Planas-Iglesias, D Poglayen, T Aittokallio, M Ammad-ud-din, CA Azencott, V Bellón, V Boeva, K Bunte, H Chheda, L Cheng, J Corander, M Dumontier, A Goldenberg, P Gopalacharyulu, M Hajiloo, D Hidru, A Jaiswal, S Kaski, B Khalfaoui, SA Khan, ER Kramer, P Marttinen, AM Mezlini, B Molparia, M Pirinen, J Saarela, M Samwald, V Stoven, H Tang, J Tang, A Torkamani, JP Vert, B Wang, T Wang, K Wennerberg, NE Wineinger, G Xiao, Y Xie, R Yeung, X Zhan, C Zhao, The Rheumatoid Arthritis Challenge Consortium, B Oliva, Y Guan

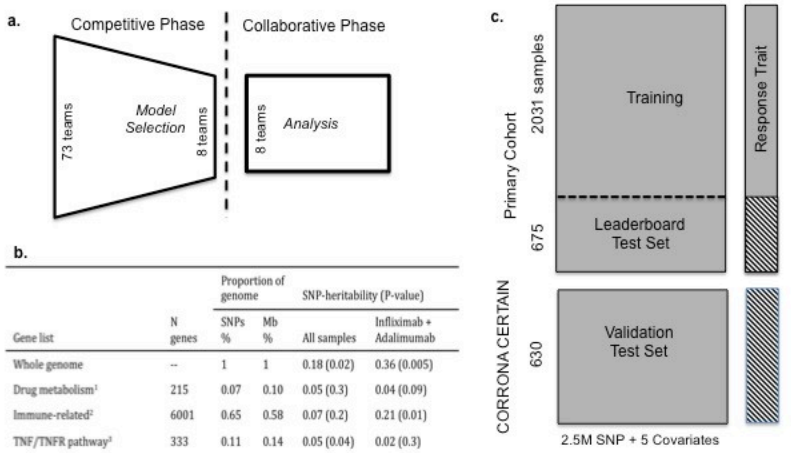
8.2.7 Bibliography

1. Vincent, F. B. *et al.* Antidrug antibodies (ADAb) to tumour necrosis factor (TNF)-specific neutralising agents in chronic inflammatory diseases: a real issue, a clinical perspective. *Ann. Rheum. Dis.* **72,** 165–78 (2013).
2. Wijbrandts, C. A. *et al.* The clinical response to infliximab in rheumatoid arthritis is in part dependent on pretreatment tumour necrosis factor alpha expression in the synovium. *Ann. Rheum. Dis.* **67,** 1139–1144 (2008).
3. Scott, I. C. *et al.* Predicting the Risk of Rheumatoid Arthritis and Its Age of Onset through Modelling Genetic Risk Variants with Smoking. *PLoS Genet.* **9,** (2013).
4. Bos, W. H. *et al.* Arthritis development in patients with arthralgia is strongly associated with anti-citrullinated protein antibody status: a prospective cohort study. *Ann. Rheum. Dis.* **69,** 490–494 (2010).
5. De Hair, M. J. H. *et al.* Smoking and overweight determine the likelihood of developing rheumatoid arthritis. *Annals of the Rheumatic Diseases* (2012). doi:10.1136/annrheumdis-2012-202254
6. Tak, P. P. A personalized medicine approach to biologic treatment of rheumatoid arthritis: A preliminary treatment algorithm. *Rheumatology* **51,** 600–609 (2012).
7. Stahl, E. A. *et al.* Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat. Genet.* **42,** 508–514 (2010).

8. Cui, J. *et al.* Genome-wide association study and gene expression analysis identifies CD84 as a predictor of response to etanercept therapy in rheumatoid arthritis. *PLoS Genet.* **9,** e1003394 (2013).
9. Costello, J. C. & Stolovitzky, G. Seeking the wisdom of crowds through challenge-based competitions in biomedical research. *Clin. Pharmacol. Ther.* **93,** 396–8 (2013).
10. Margolin, A. A. *et al.* Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer. *Sci. Transl. Med.* **5,** 181re1 (2013).
11. Dialogue on Reverse Engineering Assessment and Methods.
12. Plenge, R. M. *et al.* Crowdsourcing genetic prediction of clinical utility in the Rheumatoid Arthritis Responder Challenge. *Nat. Genet.* **45,** 468–9 (2013).
13. Costello, J. C. *et al.* A community effort to assess and improve drug sensitivity prediction algorithms. *Nat. Biotechnol.* 1–103 (2014). doi:10.1038/nbt.2877
14. Marbach, D. *et al.* Wisdom of crowds for robust gene network inference. *Nature Methods* **9,** 796–804 (2012).
15. R, K. *et al.* Crowdsourced analysis of clinical trial data to predict amyotrophic lateral sclerosis progression. (2015).
16. Pappas, D. A., Kremer, J. M., Reed, G., Greenberg, J. D. & Curtis, J. R. 'Design characteristics of the CORRONA CERTAIN study: a comparative effectiveness study of biologic agents for rheumatoid arthritis patients'. *BMC Musculoskelet. Disord.* **15,** 113 (2014).
17. Prevoo, M. L. L. *et al.* Modified disease activity scores that include twenty- eight-joint counts: Development and validation in a prospective longitudinal study of patients with rheumatoid arthritis. *Arthritis Rheum.* **38,** 44–48 (1995).
18. Yang, J. *et al.* Common {SNPs} explain a large proportion of the heritability for human height. *Nat Gen* **42,** 565–569 (2010).
19. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A tool for genome- wide complex trait analysis. *Am. J. Hum. Genet.* **88,** 76–82 (2011).
20. Chhibber, A. *et al.* Genomic architecture of pharmacological efficacy and adverse events. *Pharmacogenomics* **15,** 2025–2048 (2014).
21. Mäki-tanila, A. & Hill, W. G. Influence of gene interaction on complex trait variation with multi-locus models. *Genetics* 1–27 (2014). doi:10.1534/genetics.114.165282
22. Okser, S. *et al.* Regularized Machine Learning in the Genetic Prediction of Complex Traits. *PLoS Genet.* **10,** e1004754 (2014).
23. Van Gestel, A. M. et al. Development and validation of the European League Against Rheumatism response criteria for rheumatoid arthritis. Comparison with the preliminary American College of Rheumatology and the World Health Organization/International League Against Rheumatism Cri. Arthritis and rheumatism **39,** (1996).
24. Stolovitzky, G., Prill, R. J. & Califano, A. Lessons from the DREAM2 challenges: A community effort to assess biological network inference. *Ann. N. Y. Acad. Sci.* **1158,** 159–195 (2009).
25. Rasmussen, C. E. & Williams, C. K. I. *Gaussian Processes for Machine Learning.* (The MIT Press, 2006).

26. Whalen, S. & Pandey, G. A Comparative Analysis of Ensemble Classifiers: Case Studies in Genomics. in *13th IEEE International Conference on Data Mining (ICDM)* 807–816 (IEEE, 2013). doi:10.1109/icdm.2013.21
27. Barshan, E., Ghodsi, A., Azimifar, Z. & Zolghadri Jahromi, M. Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds. *Pattern Recognit.* **44,** 1357–1371 (2011).
28. Wolpert, D. H. Stacked Generalization. *Neural Networks* **5,** 241–259 (1992).
29. Ting, K. M. & Witten, I. H. Issues in stacked generalization. *J. Artif. Intell. Res.* **10,** 271–289 (1999).
30. Parisi, F., Strino, F., Nadler, B. & Kluger, Y. Ranking and combining multiple predictors without labeled data. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 1253–1258 (2014).
31. Klaasen, R. *et al.* The relationship between synovial lymphocyte aggregates and the clinical response to infliximab in rheumatoid arthritis: A prospective study. *Arthritis Rheum.* **60,** 3217–3224 (2009).
32. Choi, I. Y. *et al.* MRP8/14 serum levels as a strong predictor of response to biological treatments in patients with rheumatoid arthritis. *Ann. Rheum. Dis.* 1–9 (2013). doi:10.1136/annrheumdis-2013-203923
33. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5,** (2009).
34. Choy, E. *et al.* Genetic analysis of human traits in vitro: drug response and gene expression in lymphoblastoid cell lines. *PLoS Genet.* **4,** e1000287 (2008).
35. Bella, A., Ferri, C., Hernández-Orallo, J. & Ramírez-Quintana, M. On the effect of calibration in classifier combination. *Appl. Intell.* **38,** 566–585 (2013).
36. Hall, M. *et al.* The WEKA data mining software: an update. *SIGKDD Explor. Newsl.* **11,** 10–18 (2009).
37. John, G. H. & Langley, P. Estimating continuous distributions in Bayesian classifiers. *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence* 338–345 (1995).

8.2.8 Figures



b.

Gene list	N genes	Proportion of genome: SNPs %	Proportion of genome: Mb %	SNP-heritability (P-value): All samples	SNP-heritability (P-value): Infliximab + Adalimumab
Whole genome	--	1	1	0.18 (0.02)	0.36 (0.005)
Drug metabolism[1]	215	0.07	0.10	0.05 (0.3)	0.04 (0.09)
Immune-related[2]	6001	0.65	0.58	0.07 (0.2)	0.21 (0.01)
TNF/TNFR pathway[3]	333	0.11	0.14	0.05 (0.04)	0.02 (0.3)
CD84 coexpression (ImmGen)	200	0.08	0.11	0 (0.5)	0 (0.5)
r_mab chr7 top SNP PPI- direct	39	0.02	0.02	0.01 (0.2)	0.03 (0.03)

[1] Affymetrix DMET chip SNPs
[2] immport.niaid.nih.gov
[3] PPI and coexpressed genes (eQTLs)

Figure 8-1: Challenge schematic. (a) This analysis was performed in two phases. In the Competitive phase, an open competition was performed to formally evaluate and identify the best models in the world to address this research question. 73 teams representing 242 registered participants joined the challenge. Organizers evaluated model performance for test set predictions submitted by 17 teams. The 8 best performing teams were invited to join the collaborative phase. In this phase, a collectively designed experimental design was developed, in which each team independently performed analyses and challenge organizers performed a combined analysis. **(b)** Heritability estimates within the Primary Cohort. **(c)** Two datasets were used in the analysis: The discovery cohort and the CORRONA CERTAIN study.Participants were provided with 2.5 SNP genotypes + 5 covariates from two cohorts and with the response trait for 2031 individuals in the Discovery cohort ('Training Set'). At the completion of the 16 week training period, participants were required to submit a final submission containing predictions of response traits in a completely independent dataset, the CORRONA CERTAIN study ('Validation Test Set').

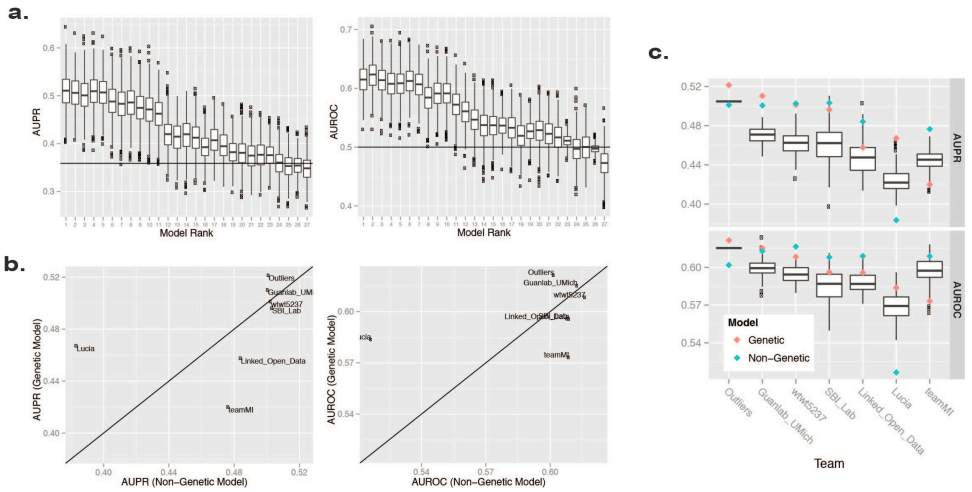


Figure 2: (a) Bootstrap distributions for each of the 27 models submitted to the classification subchallenge ordered by overall rank. The top 11 models were significantly better than random at Bonferroni corrected p-value < 0.05. **(b) AUPR and AUROC of each collaborative phase team's genetic model versus their non- genetic model.** There was no significant difference in either metric between genetic and non-genetic models (paired t-test p-value = 0.85, 0.82, for AUPR and AUROC, respectively). **(c) Distributions of the models built with SNPs selected at random, by team, along with scores for the genetic (pink) and non-genetic (blue) models.** For 5 of 7 teams, the genetic models are significantly better relative to the random SNP models for AUPR, AUROC or both.

9 BIBLIOGRAPHY

[1] J. D. Watson and F. H. C. Crick, "Molecular structure of nucleic acids," *Nature*, vol. 171, no. 4356, pp. 737–738, 1953.

[2] R. N. Shukla, *Analysis Of Chromosome*. AGROTECH PRESS, 2014.

[3] J. M. Berg, J. L. Tymoczko, and L. Stryer, *Biochemistry*. New York: WH Freeman and Company, 2002.

[4] G. W. Beadle and E. L. Tatum, "Genetic Control of Biochemical Reactions in Neurospora.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 27, no. 11, pp. 499–506, Nov. 1941.

[5] F. H. CRICK, "On protein synthesis.," *Symp. Soc. Exp. Biol.*, vol. 12, pp. 138–63, Jan. 1958.

[6] F. Crick, "Central dogma of molecular biology.," *Nature*, vol. 227, no. 5258, pp. 561–3, Aug. 1970.

[7] E. Birney, J. A. Stamatoyannopoulos, A. Dutta, R. Guigó, T. R. Gingeras, E. H. Margulies, Z. Weng, M. Snyder, E. T. Dermitzakis, R. E. Thurman, M. S. Kuehn, C. M. Taylor, S. Neph, C. M. Koch, S. Asthana, A. Malhotra, I. Adzhubei, J. A. Greenbaum, R. M. Andrews, P. Flicek, P. J. Boyle, H. Cao, N. P. Carter, G. K. Clelland, S. Davis, N. Day, P. Dhami, S. C. Dillon, M. O. Dorschner, H. Fiegler, P. G. Giresi, J. Goldy, M. Hawrylycz, A. Haydock, R. Humbert, K. D. James, B. E. Johnson, E. M. Johnson, T. T. Frum, E. R. Rosenzweig, N. Karnani, K. Lee, G. C. Lefebvre, P. A. Navas, F. Neri, S. C. J. Parker, P. J. Sabo, R. Sandstrom, A. Shafer, D. Vetrie, M. Weaver, S. Wilcox, M. Yu, F. S. Collins, J. Dekker, J. D. Lieb, T. D. Tullius, G. E. Crawford, S. Sunyaev, W. S. Noble, I. Dunham, F. Denoeud, A. Reymond, P. Kapranov, J. Rozowsky, D. Zheng, R. Castelo, A. Frankish, J. Harrow, S. Ghosh, A. Sandelin, I. L. Hofacker, R. Baertsch, D. Keefe, S. Dike, J. Cheng, H. A. Hirsch, E. A. Sekinger, J. Lagarde, J. F. Abril, A. Shahab, C. Flamm, C. Fried, J. Hackermüller, J. Hertel, M. Lindemeyer, K. Missal, A. Tanzer, S. Washietl, J. Korbel, O. Emanuelsson, J. S. Pedersen, N. Holroyd, R. Taylor, D. Swarbreck, N. Matthews, M. C. Dickson, D. J. Thomas, M. T. Weirauch, J. Gilbert, J. Drenkow, I. Bell, X. Zhao, K. G. Srinivasan, W.-K. Sung, H. S. Ooi, K. P. Chiu, S. Foissac, T. Alioto, M. Brent, L. Pachter, M. L. Tress, A. Valencia, S. W. Choo, C. Y. Choo, C. Ucla, C. Manzano, C. Wyss, E. Cheung, T. G. Clark, J. B. Brown, M. Ganesh, S. Patel, H. Tammana, J. Chrast, C. N. Henrichsen, C. Kai, J. Kawai, U. Nagalakshmi, J. Wu, Z. Lian, J. Lian, P. Newburger, X. Zhang, P. Bickel, J. S. Mattick, P. Carninci, Y. Hayashizaki, S. Weissman, T. Hubbard, R. M. Myers, J. Rogers, P. F. Stadler, T. M. Lowe, C.-L. Wei, Y. Ruan, K. Struhl, M. Gerstein, S. E. Antonarakis, Y. Fu, E. D. Green, U. Karaöz, A. Siepel, J. Taylor, L. A. Liefer, K. A. Wetterstrand, P. J. Good, E. A. Feingold, M. S. Guyer, G. M. Cooper, G. Asimenos, C. N. Dewey, M. Hou, S. Nikolaev, J. I. Montoya-Burgos, A. Löytynoja, S. Whelan, F. Pardi, T. Massingham, H. Huang, N. R. Zhang, I. Holmes, J. C. Mullikin, A. Ureta-Vidal, B. Paten, M. Seringhaus, D. Church, K. Rosenbloom, W. J. Kent, E. A. Stone, S. Batzoglou, N. Goldman, R. C. Hardison, D. Haussler, W. Miller, A. Sidow, N. D. Trinklein, Z. D. Zhang, L. Barrera, R. Stuart, D. C. King, A. Ameur, S. Enroth, M. C. Bieda, J. Kim, A. A. Bhinge, N. Jiang, J. Liu, F. Yao, V. B. Vega, C. W. H. Lee, P. Ng, A. Yang, Z. Moqtaderi, Z. Zhu, X. Xu, S. Squazzo, M. J. Oberley, D. Inman, M. A. Singer, T. A. Richmond, K. J. Munn, A. Rada-Iglesias, O. Wallerman, J. Komorowski, J. C. Fowler, P. Couttet, A. W. Bruce, O. M. Dovey, P. D. Ellis, C. F. Langford, D. A. Nix, G. Euskirchen, S. Hartman, A. E. Urban, P. Kraus, S. Van Calcar, N. Heintzman, T. H. Kim, K. Wang, C. Qu, G. Hon, R. Luna, C. K. Glass, M. G. Rosenfeld, S. F. Aldred, S. J. Cooper, A. Halees, J. M. Lin, H. P. Shulha, X. Zhang, M. Xu, J. N. S. Haidar, Y. Yu, V. R. Iyer, R. D. Green, C. Wadelius, P. J. Farnham, B. Ren, R. A. Harte, A. S. Hinrichs, H. Trumbower, H. Clawson, J. Hillman-Jackson, A. S. Zweig, K. Smith, A. Thakkapallayil, G. Barber, R. M. Kuhn, D. Karolchik, L. Armengol, C. P. Bird, P. I. W. de Bakker, A. D. Kern, N. Lopez-Bigas, J. D. Martin, B. E. Stranger, A. Woodroffe, E. Davydov, A. Dimas, E. Eyras, I. B. Hallgrímsdóttir, J. Huppert, M. C. Zody, G. R. Abecasis, X. Estivill, G. G. Bouffard, X. Guan, N. F. Hansen, J. R. Idol, V. V. B. Maduro, B. Maskeri, J. C. McDowell, M. Park, P. J. Thomas, A. C. Young, R. W. Blakesley, D. M. Muzny, E. Sodergren, D. A. Wheeler, K. C. Worley, H. Jiang, G. M. Weinstock, R. A. Gibbs, T. Graves, R. Fulton, E. R. Mardis, R. K. Wilson, M. Clamp, J. Cuff, S. Gnerre, D. B. Jaffe, J. L. Chang, K. Lindblad-Toh, E. S. Lander, M. Koriabine, M. Nefedov, K. Osoegawa, Y. Yoshinaga, B. Zhu, and P. J. de Jong, "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.," *Nature*, vol. 447, no. 7146, pp. 799–816, Jun. 2007.

[8] P. Carninci, J. Yasuda, and Y. Hayashizaki, "Multifaceted mammalian transcriptome.," *Curr. Opin. Cell Biol.*, vol. 20, no. 3, pp. 274–80, Jun. 2008.

[9] E. Segal and J. Widom, "From DNA sequence to transcriptional behaviour: a quantitative approach.," *Nat. Rev. Genet.*, vol. 10, no. 7, pp. 443–56, Jul. 2009.

[10] F. JACOB and J. MONOD, "Genetic regulatory mechanisms in the synthesis of proteins.," *J. Mol. Biol.*, vol. 3, pp. 318–56, Jun. 1961.

[11] H. Kawaji, M. C. Frith, S. Katayama, A. Sandelin, C. Kai, J. Kawai, P. Carninci, and Y. Hayashizaki, "Dynamic usage of transcription start sites within core promoters.," *Genome Biol.*, vol. 7, no. 12, p. R118, Jan. 2006.

[12] T. Juven-Gershon, J.-Y. Hsu, and J. T. Kadonaga, "Perspectives on the RNA polymerase II core promoter.," *Biochem. Soc. Trans.*, vol. 34, no. Pt 6, pp. 1047–50, Dec. 2006.

[13] J. T. Kadonaga, "Perspectives on the RNA polymerase II core promoter.," *Wiley Interdiscip. Rev. Dev. Biol.*, vol. 1, no. 1, pp. 40–51, Jan. .

[14] P. Carninci, A. Sandelin, B. Lenhard, S. Katayama, K. Shimokawa, J. Ponjavic, C. A. M. Semple, M. S. Taylor, P. G. Engström, M. C. Frith, A. R. R. Forrest, W. B. Alkema, S. L. Tan, C. Plessy, R. Kodzius, T. Ravasi, T. Kasukawa, S. Fukuda, M. Kanamori-Katayama, Y. Kitazume, H. Kawaji, C. Kai, M. Nakamura, H. Konno, K. Nakano, S. Mottagui-Tabar, P. Arner, A. Chesi, S. Gustincich, F. Persichetti, H. Suzuki, S. M. Grimmond, C. A. Wells, V. Orlando, C. Wahlestedt, E. T. Liu, M. Harbers, J. Kawai, V. B. Bajic, D. A. Hume, and Y. Hayashizaki, "Genome-wide analysis of mammalian promoter architecture and evolution.," *Nat. Genet.*, vol. 38, no. 6, pp. 626–35, Jun. 2006.

[15] I. K. Nordgren and A. Tavassoli, "A bidirectional fluorescent two-hybrid system for monitoring protein-protein interactions.," *Mol. Biosyst.*, vol. 10, no. 3, pp. 485–90, Mar. 2014.

[16] J. C. Alwine, D. J. Kemp, and G. R. Stark, "Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 74, no. 12, pp. 5350–4, Dec. 1977.

[17] M. Taniguchi, K. Miura, H. Iwao, and S. Yamanaka, "Quantitative assessment of DNA microarrays--comparison with Northern blot analyses.," *Genomics*, vol. 71, no. 1, pp. 34–9, Jan. 2001.

[18] S. Streit, C. W. Michalski, M. Erkan, J. Kleeff, and H. Friess, "Northern blot analysis for detection and quantification of RNA in pancreatic cancer cells and tissues.," *Nat. Protoc.*, vol. 4, no. 1, pp. 37–43, Jan. 2009.

[19] H. Towbin, T. Staehelin, and J. Gordon, "Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 76, no. 9, pp. 4350–4, Sep. 1979.

[20] P. R. Langer-Safer, M. Levine, and D. C. Ward, "Immunological method for mapping genes on Drosophila polytene chromosomes.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 79, no. 14, pp. 4381–5, Jul. 1982.

[21] C. Joyce, "Quantitative RT-PCR. A review of current methodologies.," *Methods Mol. Biol.*, vol. 193, pp. 83–92, Jan. 2002.

[22] V. E. Velculescu, L. Zhang, B. Vogelstein, and K. W. Kinzler, "Serial analysis of gene expression.," *Science*, vol. 270, no. 5235, pp. 484–7, Oct. 1995.

[23] H. Matsumura, A. Ito, H. Saitoh, P. Winter, G. Kahl, M. Reuter, D. H. Krüger, and R. Terauchi, "SuperSAGE.," *Cell. Microbiol.*, vol. 7, no. 1, pp. 11–8, Jan. 2005.

[24] S. Fodor, J. Read, M. Pirrung, L. Stryer, A. Lu, and D. Solas, "Light-directed, spatially addressable parallel chemical synthesis," *Science (80-.).*, vol. 251, no. 4995, pp. 767–773, Feb. 1991.

[25] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray," *Science (80-.).*, vol. 270, no. 5235, pp. 467–470, Oct. 1995.

[26] R. Lucito, J. Healy, J. Alexander, A. Reiner, D. Esposito, M. Chi, L. Rodgers, A. Brady, J. Sebat, J. Troge, J. A. West, S. Rostan, K. C. Q. Nguyen, S. Powers, K. Q. Ye, A. Olshen, E. Venkatraman, L. Norton, and M. Wigler, "Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation.," *Genome Res.*, vol. 13, no. 10, pp. 2291–305, Oct. 2003.

[27] T. LaFramboise, "Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances.," *Nucleic Acids Res.*, vol. 37, no. 13, pp. 4181–93, Jul. 2009.

[28] V. R. Iyer, C. E. Horak, C. S. Scafe, D. Botstein, M. Snyder, and P. O. Brown, "Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF.," *Nature*, vol. 409, no. 6819, pp. 533–8, Jan. 2001.

[29] J. D. Hoheisel, "Microarray technology: beyond transcript profiling and genotype analysis.," *Nat. Rev. Genet.*, vol. 7, no. 3, pp. 200–10, Mar. 2006.

[30] D. Gresham, M. J. Dunham, and D. Botstein, "Comparing whole genomes using DNA microarrays.," *Nat. Rev. Genet.*, vol. 9, no. 4, pp. 291–302, Apr. 2008.

[31] C. M. Perou, T. Sørlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, O. Fluge, A. Pergamenschikov, C. Williams, S. X. Zhu, P. E. Lønning, A. L. Børresen-Dale, P. O. Brown, and D. Botstein, "Molecular portraits of human breast tumours.," *Nature*, vol. 406, no. 6797, pp. 747–52, Aug. 2000.

[32] M. J. Buck and J. D. Lieb, "ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments.," *Genomics*, vol. 83, no. 3, pp. 349–60, Mar. 2004.

[33] P. J. Park, "ChIP-seq: advantages and challenges of a maturing technology.," *Nat. Rev. Genet.*, vol. 10, no. 10, pp. 669–80, Oct. 2009.

[34] M. Sultan, M. H. Schulz, H. Richard, A. Magen, A. Klingenhoff, M. Scherf, M. Seifert, T. Borodina, A. Soldatov, D. Parkhomchuk, D. Schmidt, S. O'Keeffe, S. Haas, M. Vingron, H. Lehrach, and M.-L. Yaspo, "A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome," *Science (80-.).*, vol. 321, no. 5891, pp. 956–960, Aug. 2008.

[35] A. Mandlik, J. Livny, W. P. Robins, J. M. Ritchie, J. J. Mekalanos, and M. K. Waldor, "RNA-Seq-Based Monitoring of Infection-Linked Changes in Vibrio cholerae Gene Expression," *Cell Host Microbe*, vol. 10, no. 2, pp. 165–174, Aug. 2011.

[36] F. Mitelman, B. Johansson, and F. Mertens, "The impact of translocations and gene fusions on cancer causation.," *Nat. Rev. Cancer*, vol. 7, no. 4, pp. 233–45, Apr. 2007.

[37] T. E. Consortium, "The ENCODE (ENCyclopedia Of DNA Elements) Project.," *Science*, vol. 306, no. 5696, pp. 636–40, Oct. 2004.

[38] J. H. Malone and B. Oliver, "Microarrays, deep sequencing and the true measure of the transcriptome.," *BMC Biol.*, vol. 9, p. 34, Jan. 2011.

[39] A. Agarwal, D. Koppstein, J. Rozowsky, A. Sboner, L. Habegger, L. W. Hillier, R. Sasidharan, V. Reinke, R. H. Waterston, and M. Gerstein, "Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays.," *BMC Genomics*, vol. 11, p. 383, Jan. 2010.

[40] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics.," *Nat. Rev. Genet.*, vol. 10, no. 1, pp. 57–63, Jan. 2009.

[41] E. E. Schadt, M. D. Linderman, J. Sorenson, L. Lee, and G. P. Nolan, "Computational solutions to large-scale data management and analysis.," *Nat. Rev. Genet.*, vol. 11, no. 9, pp. 647–57, Sep. 2010.

[42] W. W. Wasserman and A. Sandelin, "Applied bioinformatics for the identification of regulatory elements.," *Nat. Rev. Genet.*, vol. 5, no. 4, pp. 276–87, Apr. 2004.

[43] S. P. Parker and M.-H. Companies, *McGraw-Hill Encyclopedia of Science & Technology*. 1997.

[44] J. J. Fischer, J. Toedling, T. Krueger, M. Schueler, W. Huber, and S. Sperling, "Combinatorial effects of four histone modifications in transcription and differentiation.," *Genomics*, vol. 91, no. 1, pp. 41–51, Jan. 2008.

[45] S. Hahn, "Transcriptional regulation. Meeting on regulatory mechanisms in eukaryotic transcription.," *EMBO Rep.*, vol. 9, no. 7, pp. 612–6, Jul. 2008.

[46] P. A. Wade, "Methyl CpG-binding proteins and transcriptional repression.," *Bioessays*, vol. 23, no. 12, pp. 1131–7, Dec. 2001.

[47] M. I. Arnone and E. H. Davidson, "The hardwiring of development: organization and function of genomic regulatory systems.," *Development*, vol. 124, no. 10, pp. 1851–64, May 1997.

[48] S. J. Furney, D. G. Higgins, C. A. Ouzounis, and N. López-Bigas, "Structural and functional properties of genes involved in human cancer.," *BMC Genomics*, vol. 7, p. 3, Jan. 2006.

[49] S. A. Boyadjiev and E. W. Jabs, "Online Mendelian Inheritance in Man (OMIM) as a knowledgebase for human developmental disorders.," *Clin. Genet.*, vol. 57, no. 4, pp. 253–66, Apr. 2000.

[50] E. van Nimwegen, "Scaling laws in the functional content of genomes.," *Trends Genet.*, vol. 19, no. 9, pp. 479–84, Sep. 2003.

[51] P. A. Gray, H. Fu, P. Luo, Q. Zhao, J. Yu, A. Ferrari, T. Tenzen, D.-I. Yuk, E. F. Tsung, Z. Cai, J. A. Alberta, L.-P. Cheng, Y. Liu, J. M. Stenman, M. T. Valerius, N. Billings, H. A. Kim, M. E. Greenberg, A. P. McMahon, D. H. Rowitch, C. D. Stiles, and Q. Ma, "Mouse brain organization revealed through direct genome-scale TF expression analysis.," *Science*, vol. 306, no. 5705, pp. 2255–7, Dec. 2004.

[52] D. N. Messina, J. Glasscock, W. Gish, and M. Lovett, "An ORFeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression.," *Genome Res.*, vol. 14, no. 10B, pp. 2041–7, Oct. 2004.

[53] J. S. Reece-Hoyes, B. Deplancke, J. Shingles, C. A. Grove, I. A. Hope, and A. J. M. Walhout, "A compendium of Caenorhabditis elegans regulatory transcription factors: a resource for mapping transcription regulatory networks.," *Genome Biol.*, vol. 6, no. 13, p. R110, Jan. 2005.

[54] B. Adryan and S. A. Teichmann, "FlyTF: a systematic review of site-specific transcription factors in the fruit fly Drosophila melanogaster.," *Bioinformatics*, vol. 22, no. 12, pp. 1532–3, Jun. 2006.

[55] S.-W. Ho, G. Jona, C. T. L. Chen, M. Johnston, and M. Snyder, "Linking DNA-binding proteins to their recognition sequences by using protein microarrays.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 103, no. 26, pp. 9940–5, Jun. 2006.

[56] J. M. Vaquerizas, S. K. Kummerfeld, S. A. Teichmann, and N. M. Luscombe, "A census of human transcription factors: function, expression and evolution.," *Nat. Rev. Genet.*, vol. 10, no. 4, pp. 252–63, Apr. 2009.

[57] W. F. Shen, K. Detmer, T. A. Simonitch-Eason, H. J. Lawrence, and C. Largman, "Alternative splicing of the HOX 2.2 homeobox gene in human hematopoietic cells and murine embryonic and adult tissues.," *Nucleic Acids Res.*, vol. 19, no. 3, pp. 539–45, Feb. 1991.

[58] D. R. Morris and A. P. Geballe, "Upstream open reading frames as regulators of mRNA translation.," *Mol. Cell. Biol.*, vol. 20, no. 23, pp. 8635–42, Dec. 2000.

[59] M. C. King and A. C. Wilson, "Evolution at two levels in humans and chimpanzees.," *Science*, vol. 188, no. 4184, pp. 107–16, Apr. 1975.

[60] J. L. Riechmann, J. Heard, G. Martin, L. Reuber, C. Jiang, J. Keddie, L. Adam, O. Pineda, O. J. Ratcliffe, R. R. Samaha, R. Creelman, M. Pilgrim, P. Broun, J. Z. Zhang, D. Ghandehari, B. K. Sherman, and G. Yu, "Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes.," *Science*, vol. 290, no. 5499, pp. 2105–10, Dec. 2000.

[61] J. W. Fondon and H. R. Garner, "Molecular origins of rapid and continuous morphological evolution.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101, no. 52, pp. 18058–63, Dec. 2004.

[62] G. A. Wray, M. W. Hahn, E. Abouheif, J. P. Balhoff, M. Pizer, M. V Rockman, and L. A. Romano, "The evolution of transcriptional regulation in eukaryotes.," *Mol. Biol. Evol.*, vol. 20, no. 9, pp. 1377–419, Sep. 2003.

[63] L. Kuras and K. Struhl, "Binding of TBP to promoters in vivo is stimulated by activators and requires Pol II holoenzyme.," *Nature*, vol. 399, no. 6736, pp. 609–13, Jun. 1999.

[64] Y. Cang, D. T. Auble, and G. Prelich, "A new regulatory domain on the TATA-binding protein.," *EMBO J.*, vol. 18, no. 23, pp. 6662–71, Dec. 1999.

[65] S. M. Kraemer, R. T. Ranallo, R. C. Ogg, and L. A. Stargell, "TFIIA interacts with TFIID via association with TATA-binding protein and TAF40.," *Mol. Cell. Biol.*, vol. 21, no. 5, pp. 1737–46, Mar. 2001.

[66] W. Deng and S. G. E. Roberts, "TFIIB and the regulation of transcription by RNA polymerase II.," *Chromosoma*, vol. 116, no. 5, pp. 417–29, Oct. 2007.

[67] D. A. Bushnell, K. D. Westover, R. E. Davis, and R. D. Kornberg, "Structural basis of transcription: an RNA polymerase II-TFIIB cocrystal at 4.5 Angstroms.," *Science*, vol. 303, no. 5660, pp. 983–8, Feb. 2004.

[68] D. Kostrewa, M. E. Zeller, K.-J. Armache, M. Seizl, K. Leike, M. Thomm, and P. Cramer, "RNA polymerase II-TFIIB structure and mechanism of transcription initiation.," *Nature*, vol. 462, no. 7271, pp. 323–30, Nov. 2009.

[69] X. Liu, D. A. Bushnell, D. Wang, G. Calero, and R. D. Kornberg, "Structure of an RNA polymerase II-TFIIB complex and the transcription initiation mechanism.," *Science*, vol. 327, no. 5962, pp. 206–9, Jan. 2010.

[70] E. Compe and J.-M. Egly, "TFIIH: when transcription met DNA repair.," *Nat. Rev. Mol. Cell Biol.*, vol. 13, no. 6, pp. 343–54, Jun. 2012.

[71] A. D. Basehoar, S. J. Zanton, and B. F. Pugh, "Identification and distinct regulation of yeast TATA box-containing genes.," *Cell*, vol. 116, no. 5, pp. 699–709, Mar. 2004.

[72] M. Baumann, J. Pontiller, and W. Ernst, "Structure and basal transcription complex of RNA polymerase II core promoters in the mammalian genome: an overview.," *Mol. Biotechnol.*, vol. 45, no. 3, pp. 241–7, Jul. 2010.

[73] G. D. Stormo and D. S. Fields, "Specificity, free energy and information content in protein-DNA interactions.," *Trends Biochem. Sci.*, vol. 23, no. 3, pp. 109–13, Mar. 1998.

[74] G. D. Stormo, "DNA binding sites: representation and discovery.," *Bioinformatics*, vol. 16, no. 1, pp. 16–23, Jan. 2000.

[75] T. G. do Rego, H. G. Roider, F. A. T. de Carvalho, and I. G. Costa, "Inferring epigenetic and transcriptional regulation during blood cell development with a mixture of sparse linear models.," *Bioinformatics*, vol. 28, no. 18, pp. 2297–303, Sep. 2012.

[76] P. V Benos, M. L. Bulyk, and G. D. Stormo, "Additivity in protein-DNA interactions: how good an approximation is it?," *Nucleic Acids Res.*, vol. 30, no. 20, pp. 4442–51, Oct. 2002.

[77] M. C. Frith, Y. Fu, L. Yu, J.-F. Chen, U. Hansen, and Z. Weng, "Detection of functional DNA motifs via statistical over-representation.," *Nucleic Acids Res.*, vol. 32, no. 4, pp. 1372–81, Jan. 2004.

[78] D. J. Galas, M. Eggert, and M. S. Waterman, "Rigorous pattern-recognition methods for DNA sequences. Analysis of promoter sequences from Escherichia coli.," *J. Mol. Biol.*, vol. 186, no. 1, pp. 117–28, Nov. 1985.

[79] G. Casella and E. I. George, "Explaining the Gibbs Sampler," *Am. Stat.*, vol. 46, no. 3, pp. 167–174, 1992.

[80] T. L. Bailey and C. Elkan, "The value of prior knowledge in discovering motifs with MEME.," *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, vol. 3, pp. 21–9, Jan. 1995.

[81] C. T. Workman and G. D. Stormo, "ANN-Spec: a method for discovering transcription factor binding sites with improved specificity.," *Pac. Symp. Biocomput.*, pp. 467–78, Jan. 2000.

[82] J. Keilwagen, J. Grau, I. A. Paponov, S. Posch, M. Strickert, and I. Grosse, "De-novo discovery of differentially abundant transcription factor binding sites including their positional preference.," *PLoS Comput. Biol.*, vol. 7, no. 2, p. e1001070, Jan. 2011.

[83] E. Wingender, "The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation.," *Brief. Bioinform.*, vol. 9, no. 4, pp. 326–32, Jul. 2008.

[84] A. Mathelier, X. Zhao, A. W. Zhang, F. Parcy, R. Worsley-Hunt, D. J. Arenillas, S. Buchman, C. Chen, A. Chou, H. Ienasescu, J. Lim, C. Shyr, G. Tan, M. Zhou, B. Lenhard, A. Sandelin, and W. W. Wasserman, "JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles.," *Nucleic Acids Res.*, vol. 42, no. Database issue, pp. D142–7, Jan. 2014.

[85] D. E. Schones, A. D. Smith, and M. Q. Zhang, "Statistical significance of cis-regulatory modules.," *BMC Bioinformatics*, vol. 8, p. 19, Jan. 2007.

[86] S. Aerts, G. Thijs, B. Coessens, M. Staes, Y. Moreau, and B. De Moor, "Toucan: deciphering the cis-regulatory logic of coregulated genes.," *Nucleic Acids Res.*, vol. 31, no. 6, pp. 1753–64, Mar. 2003.

[87] Y. Fu, M. C. Frith, P. M. Haverty, and Z. Weng, "MotifViz: an analysis and visualization tool for motif discovery.," *Nucleic Acids Res.*, vol. 32, no. Web Server issue, pp. W420–3, Jul. 2004.

[88] N. Bellora, D. Farré, and M. Mar Albà, "PEAKS: identification of regulatory motifs by their position in DNA sequences.," *Bioinformatics*, vol. 23, no. 2, pp. 243–4, Jan. 2007.

[89] D. T. Odom, R. D. Dowell, E. S. Jacobsen, W. Gordon, T. W. Danford, K. D. MacIsaac, P. A. Rolfe, C. M. Conboy, D. K. Gifford, and E. Fraenkel, "Tissue-specific transcriptional regulation has diverged significantly between human and mouse.," *Nat. Genet.*, vol. 39, no. 6, pp. 730–2, Jun. 2007.

[90] S. Veerla, M. Ringnér, and M. Höglund, "Genome-wide transcription factor binding site/promoter databases for the analysis of gene sets and co-occurrence of transcription factor binding motifs.," *BMC Genomics*, vol. 11, p. 145, Jan. 2010.

[91] A. Sandelin, "JASPAR: an open-access database for eukaryotic transcription factor binding profiles," *Nucleic Acids Res.*, vol. 32, no. 90001, p. 91D–94, Jan. 2004.

[92] G. Robertson, M. Bilenky, K. Lin, A. He, W. Yuen, M. Dagpinar, R. Varhol, K. Teague, O. L. Griffith, X. Zhang, Y. Pan, M. Hassel, M. C. Sleumer, W. Pan, E. D. Pleasance, M. Chuang, H. Hao, Y. Y. Li, N. Robertson, C. Fjell, B. Li, S. B. Montgomery, T. Astakhova, J. Zhou, J. Sander, A. S. Siddiqui, and S. J. M. Jones, "cisRED: a database system for genome-scale computational discovery of regulatory elements.," *Nucleic Acids Res.*, vol. 34, no. Database issue, pp. D68–73, Jan. 2006.

[93] S. Gama-Castro, H. Salgado, M. Peralta-Gil, A. Santos-Zavaleta, L. Muñiz-Rascado, H. Solano-Lira, V. Jimenez-Jacinto, V. Weiss, J. S. García-Sotelo, A. López-Fuentes, L. Porrón-Sotelo, S. Alquicira-Hernández, A. Medina-Rivera, I. Martínez-Flores, K. Alquicira-Hernández, R. Martínez-Adame, C. Bonavides-Martínez, J. Miranda-Ríos, A. M. Huerta, A. Mendoza-Vargas, L. Collado-Torres, B. Taboada, L. Vega-Alvarado, M. Olvera, L. Olvera, R. Grande, E. Morett, and J. Collado-Vides, "RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units).," *Nucleic Acids Res.*, vol. 39, no. Database issue, pp. D98–105, Jan. 2011.

[94] M. T. Weirauch, A. Yang, M. Albu, A. G. Cote, A. Montenegro-Montero, P. Drewe, H. S. Najafabadi, S. A. Lambert, I. Mann, K. Cook, H. Zheng, A. Goity, H. van Bakel, J.-C. Lozano, M. Galli, M. G. Lewsey, E. Huang, T. Mukherjee, X. Chen, J. S. Reece-Hoyes, S. Govindarajan, G. Shaulsky, A. J. M. Walhout, F.-Y. Bouget, G. Ratsch, L. F. Larrondo, J. R. Ecker, and T. R. Hughes, "Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity," *Cell*, vol. 158, no. 6, pp. 1431–1443, Sep. 2014.

[95] M. L. Crowe, C. Serizet, V. Thareau, S. Aubourg, P. Rouzé, P. Hilson, J. Beynon, P. Weisbeek, P. van Hummelen, P. Reymond, J. Paz-Ares, W. Nietfeld, and M. Trick, "CATMA: a complete Arabidopsis GST database.," *Nucleic Acids Res.*, vol. 31, no. 1, pp. 156–158, Jan. 2003.

[96] T. Schiex, A. Moisan, and P. Rouzé, "Eugène: An Eukaryotic Gene Finder That Combines Several Sources of EvidenceComputational Biology," in *2066*, O. Gascuel and M.-F. Sagot, Eds. Springer Berlin / Heidelberg, 2001, pp. 111–125.

[97] J. Allemeersch, S. Durinck, R. Vanderhaeghen, P. Alard, R. Maes, K. Seeuws, T. Bogaert, K. Coddens, K. Deschouwer, P. Van Hummelen, M. Vuylsteke, Y. Moreau, J. Kwekkeboom, A. H. M. Wijfjes, S. May, J. Beynon, P. Hilson, and M. T. R. Kuiper, "Benchmarking the CATMA Microarray. A Novel Tool forArabidopsis Transcriptome Analysis1[w]," *Plant Physiol.*, vol. 137, no. 2, pp. 588–601, 2005.

[98] M. F. Elshal and J. P. McCoy, "Multiplex bead array assays: performance evaluation and comparison of sensitivity to ELISA.," *Methods*, vol. 38, no. 4, pp. 317–23, Apr. 2006.

[99] B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.," *Bioinformatics*, vol. 19, no. 2, pp. 185–93, Jan. 2003.

[100] B. Bolstad, "Low Level Analysis of High-density Oligonucleotide Array Data: Background, Normalization and Summarization," University of California, Berkely, 2004.

[101] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed, "Exploration, normalization, and summaries of high density oligonucleotide array probe level data.," *Biostatistics*, vol. 4, no. 2, pp. 249–64, Apr. 2003.

[102] M. J. Zilliox and R. A. Irizarry, "A gene expression bar code for microarray data.," *Nat. Methods*, vol. 4, no. 11, pp. 911–3, Nov. 2007.

[103] M. Lin, L.-J. Wei, W. R. Sellers, M. Lieberfarb, W. H. Wong, and C. Li, "dChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data.," *Bioinformatics*, vol. 20, no. 8, pp. 1233–40, May 2004.

[104] K. A. Baggerly, K. R. Coombes, and E. S. Neeley, "Run batch effects potentially compromise the usefulness of genomic signatures for ovarian cancer.," *J. Clin. Oncol.*, vol. 26, no. 7, pp. 1186–7; author reply 1187–8, Mar. 2008.

[105] J. T. Leek, R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly, and R. A. Irizarry, "Tackling the widespread and critical impact of batch effects in high-throughput data.," *Nat. Rev. Genet.*, vol. 11, no. 10, pp. 733–9, Oct. 2010.

[106] E. S. Lander, "Array of hope.," *Nat. Genet.*, vol. 21, no. 1 Suppl, pp. 3–4, Jan. 1999.

[107] O. Alter, P. O. Brown, and D. Botstein, "Singular value decomposition for genome-wide expression data processing and modeling.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 97, no. 18, pp. 10101–6, Aug. 2000.

[108] M. Benito, J. Parker, Q. Du, J. Wu, D. Xiang, C. M. Perou, and J. S. Marron, "Adjustment of systematic microarray data biases.," *Bioinformatics*, vol. 20, no. 1, pp. 105–14, Jan. 2004.

[109] W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical Bayes methods.," *Biostatistics*, vol. 8, no. 1, pp. 118–27, Jan. 2007.

[110] C. Chen, K. Grennan, J. Badner, D. Zhang, E. Gershon, L. Jin, and C. Liu, "Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods.," *PLoS One*, vol. 6, no. 2, p. e17238, Jan. 2011.

[111] J. T. Leek and J. D. Storey, "Capturing heterogeneity in gene expression studies by surrogate variable analysis.," *PLoS Genet.*, vol. 3, no. 9, pp. 1724–35, Sep. 2007.

[112] R. A. Irizarry, C. Wang, Y. Zhou, and T. P. Speed, "Gene set enrichment analysis made simple.," *Stat. Methods Med. Res.*, vol. 18, no. 6, pp. 565–75, Dec. 2009.

[113] R. S. Ron Shamir, "Algorithmic Approaches to Clustering Gene Expression Data," in *Current Topics in Computational Biology*, T. Jiang, Y. Xu, and M. Q. Zhang, Eds. MIT Press, 2002, pp. 259–299.

[114] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 95, no. 25, pp. 14863–8, Dec. 1998.

[115] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, 1967.

[116] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub, "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 96, no. 6, pp. 2907–2912, 1999.

[117] A. Schliep, a. Schonhuth, C. Steinhoff, A. Schönhuth, and C. Steinhoff, "Using hidden Markov models to analyze gene expression time course data.," *Bioinformatics*, vol. 19 Suppl 1, no. Suppl 1, pp. i255–63, Jan. 2003.

[118] Z. Bar-Joseph, G. K. Gerber, D. K. Gifford, T. S. Jaakkola, and I. Simon, "Continuous representations of time-series gene expression data.," *J. Comput. Biol.*, vol. 10, no. 3–4, pp. 341–356, 2003.

[119] M. F. Ramoni, P. Sebastiani, and I. S. Kohane, "Cluster analysis of gene expression dynamics.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 99, no. 14, pp. 9121–9126, 2002.

[120] L. P. Zhao, R. Prentice, and L. Breeden, "Statistical modeling of large microarray data sets to identify stimulus-response profiles.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 98, no. 10, pp. 5631–5636, 2001.

[121] X. Lu, W. Zhang, Z. S. Qin, K. E. Kwast, and J. S. Liu, "Statistical resynchronization and Bayesian detection of periodically expressed genes," *Nucleic Acids Res.*, vol. 32, no. 2, pp. 447–455, 2004.

[122] C. S. Möller-Levet, K.-H. Cho, and O. Wolkenhauer, "Microarray data clustering based on temporal variation: FCV with TSD preclustering.," *Appl. Bioinformatics*, vol. 2, no. 1, pp. 35–45, 2003.

[123] S. D. Peddada, E. K. Lobenhofer, L. Li, C. A. Afshari, C. R. Weinberg, and D. M. Umbach, "Gene selection and clustering for time-course and dose-response microarray experiments using order-restricted inference," *Bioinformatics*, vol. 19, no. 7, pp. 834–841, 2003.

[124] M. J. De Hoon, S. Imoto, and S. Miyano, "Statistical analysis of a small set of time-ordered gene expression data using linear splines," *Bioinformatics*, vol. 18, no. 11, pp. 1477–1485, 2002.

[125] J. Ernst, G. J. Nau, and Z. Bar-Joseph, "Clustering short time series gene expression data.," *Bioinformatics*, vol. 21 Suppl 1, no. suppl_1, pp. i159–68, Jun. 2005.

[126] A.-L. Barabási and Z. N. Oltvai, "Network biology: understanding the cell's functional organization.," *Nat. Rev. Genet.*, vol. 5, no. 2, pp. 101–13, Feb. 2004.

[127] H. Kitano, *Foundations of Systems Biology*. MIT Press, 2001.

[128] L. von Bertalanffy, *General system theory: foundations, development, applications*. New York: George Braziller, 1968.

[129] L. Euler, "Solutio problematis ad geometriam situs pertinentis," *Comment. Acad. Sci. Petropolitanae*, vol. 8, no. 1741, pp. 128–140, 1735.

[130] R. Diestel, *Graph Theory*, Electronic. New York: Springer-Verlag, 2000.

[131] P. Erdős and A. Rényi, "On the Evolution of Random Graphs," *Publ. Math. Inst. Hungarian Acad. Sci.*, pp. 17–61, 1960.

[132] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabási, "The large-scale organization of metabolic networks.," *Nature*, vol. 407, no. 6804, pp. 651–4, Oct. 2000.

[133] J. Loscalzo, I. Kohane, and A.-L. Barabasi, "Human disease classification in the postgenomic era: a complex systems approach to human pathobiology.," *Mol. Syst. Biol.*, vol. 3, p. 124, Jan. 2007.

[134] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease.," *Nat. Rev. Genet.*, vol. 12, no. 1, pp. 56–68, Jan. 2011.

[135] R. Albert, "Scale-free networks in cell biology.," *J. Cell Sci.*, vol. 118, no. Pt 21, pp. 4947–57, Nov. 2005.

[136] L. da F. Costa, F. A. Rodrigues, and A. S. Cristino, "Complex networks: the key to systems biology," *Genet. Mol. Biol.*, vol. 31, no. 3, pp. 591–601, 2008.

[137] L. Freeman, "A Set of Measures of Centrality Based on Betweenness," *Sociometry*, vol. 40, no. 1, pp. 35 – 41, 1977.

[138] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numer. Math.*, vol. 1, no. 1, pp. 269–271, Dec. 1959.

[139] G. Sabidussi, "The centrality of a graph.," *Psychometrika*, vol. 31, no. 4, pp. 581–603, Dec. 1966.

[140] T. Opsahl, F. Agneessens, and J. Skvoretz, "Node centrality in weighted networks: Generalizing degree and shortest paths," *Soc. Networks*, vol. 32, no. 3, pp. 245–251, Jul. 2010.

[141] H. Yu, P. M. Kim, E. Sprecher, V. Trifonov, and M. Gerstein, "The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics.," *PLoS Comput. Biol.*, vol. 3, no. 4, p. e59, Apr. 2007.

[142] J. Loscalzo and A.-L. Barabasi, "Systems biology and the future of medicine.," *Wiley Interdiscip. Rev. Syst. Biol. Med.*, vol. 3, no. 6, pp. 619–27, Jan. 2011.

[143] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: simple building blocks of complex networks.," *Science*, vol. 298, no. 5594, pp. 824–7, Oct. 2002.

[144] M. M. Babu, N. M. Luscombe, L. Aravind, M. Gerstein, and S. A. Teichmann, "Structure and evolution of transcriptional regulatory networks.," *Curr. Opin. Struct. Biol.*, vol. 14, no. 3, pp. 283–91, Jun. 2004.

[145] A. Ma'ayan, "Insights into the organization of biochemical regulatory networks using graph theory analyses.," *J. Biol. Chem.*, vol. 284, no. 9, pp. 5451–5, Feb. 2009.

[146] D. K. Arrell and A. Terzic, "Network systems biology for drug discovery.," *Clin. Pharmacol. Ther.*, vol. 88, no. 1, pp. 120–5, Jul. 2010.

[147] A. Ma'ayan, S. L. Jenkins, S. Neves, A. Hasseldine, E. Grace, B. Dubin-Thaler, N. J. Eungdamrong, G. Weng, P. T. Ram, J. J. Rice, A. Kershenbaum, G. A. Stolovitzky, R. D. Blitzer, and R. Iyengar, "Formation of regulatory patterns during signal propagation in a Mammalian cellular network.," *Science*, vol. 309, no. 5737, pp. 1078–83, Aug. 2005.

[148] A. Bossi and B. Lehner, "Tissue specificity and the human protein interaction network.," *Mol. Syst. Biol.*, vol. 5, p. 260, Jan. 2009.

[149] L. Giot, J. S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y. L. Hao, C. E. Ooi, B. Godwin, E. Vitols, G. Vijayadamodar, P. Pochart, H. Machineni, M. Welsh, Y. Kong, B. Zerhusen, R. Malcolm, Z. Varrone, A. Collis, M. Minto, S. Burgess, L. McDaniel, E. Stimpson, F. Spriggs, J. Williams, K. Neurath, N. Ioime, M. Agee, E. Voss, K. Furtak, R. Renzulli, N. Aanensen, S. Carrolla, E. Bickelhaupt, Y. Lazovatsky, A. DaSilva, J. Zhong, C. A. Stanyon, R. L. Finley, K. P. White, M. Braverman, T. Jarvie, S. Gold, M. Leach, J. Knight, R. A. Shimkets, M. P. McKenna, J. Chant, and J. M. Rothberg, "A protein interaction map of Drosophila melanogaster.," *Science*, vol. 302, no. 5651, pp. 1727–36, Dec. 2003.

[150] Y. Li, P. Agarwal, and D. Rajagopalan, "A global pathway crosstalk network.," *Bioinformatics*, vol. 24, no. 12, pp. 1442–7, Jun. 2008.

[151] K. Basso, A. A. Margolin, G. Stolovitzky, U. Klein, R. Dalla-Favera, and A. Califano, "Reverse engineering of regulatory networks in human B cells.," *Nat. Genet.*, vol. 37, no. 4, pp. 382–90, Apr. 2005.

[152] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási, "The human disease network.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 104, no. 21, pp. 8685–90, May 2007.

[153] S. Orchard and H. Hermjakob, "The HUPO proteomics standards initiative--easing communication and minimizing data loss in a changing world.," *Brief. Bioinform.*, vol. 9, no. 2, pp. 166–73, Mar. 2008.

[154] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S.-A. Sansone, R. H. Scheuermann, N. Shah, P. L. Whetzel, and S. Lewis, "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration.," *Nat. Biotechnol.*, vol. 25, no. 11, pp. 1251–5, Nov. 2007.

[155] S. W. Michnick, "Exploring protein interactions by interaction-induced folding of proteins from complementary peptide fragments.," *Curr. Opin. Struct. Biol.*, vol. 11, no. 4, pp. 472–7, Aug. 2001.

[156] A. Brückner, C. Polge, N. Lentze, D. Auerbach, and U. Schlattner, "Yeast two-hybrid, a powerful tool for systems biology.," *Int. J. Mol. Sci.*, vol. 10, no. 6, pp. 2763–88, Jun. 2009.

[157] X. Xu, Y. Song, Y. Li, J. Chang, H. Zhang, and L. An, "The tandem affinity purification method: an efficient system for protein complex purification and protein interaction identification.," *Protein Expr. Purif.*, vol. 72, no. 2, pp. 149–56, Aug. 2010.

[158] C.-D. Hu, Y. Chinenov, and T. K. Kerppola, "Visualization of interactions among bZIP and Rel family proteins in living cells using bimolecular fluorescence complementation.," *Mol. Cell*, vol. 9, no. 4, pp. 789–98, Apr. 2002.

[159] T. J. Magliery, C. G. M. Wilson, W. Pan, D. Mishler, I. Ghosh, A. D. Hamilton, and L. Regan, "Detecting protein-protein interactions with a green fluorescent protein fragment reassembly trap: scope and mechanism.," *J. Am. Chem. Soc.*, vol. 127, no. 1, pp. 146–57, Jan. 2005.

[160] R. N. Day, A. Periasamy, and F. Schaufele, "Fluorescence resonance energy transfer microscopy of localized protein interactions in the living cell nucleus.," *Methods*, vol. 25, no. 1, pp. 4–18, Sep. 2001.

[161] Y. Xu, D. W. Piston, and C. H. Johnson, "A bioluminescence resonance energy transfer (BRET) system: application to interacting circadian clock proteins.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 96, no. 1, pp. 151–6, Jan. 1999.

[162] G. MacBeath and S. L. Schreiber, "Printing proteins as microarrays for high-throughput function determination.," *Science*, vol. 289, no. 5485, pp. 1760–3, Sep. 2000.

[163] C. Boozer, G. Kim, S. Cong, H. Guan, and T. Londergan, "Looking towards label-free biomolecular interaction analysis in a high-throughput format: a review of new surface plasmon resonance technologies.," *Curr. Opin. Biotechnol.*, vol. 17, no. 4, pp. 400–5, Aug. 2006.

[164] R. P. Bahadur, P. Chakrabarti, F. Rodier, and J. Janin, "A dissection of specific and non-specific protein-protein interfaces.," *J. Mol. Biol.*, vol. 336, no. 4, pp. 943–55, Feb. 2004.

[165] H. Yu, "Extending the size limit of protein nuclear magnetic resonance.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 96, no. 2, pp. 332–4, Jan. 1999.

[166] E. M. Marcotte, "Detecting Protein Function and Protein-Protein Interactions from Genome Sequences," *Science (80-.).*, vol. 285, no. 5428, pp. 751–753, Jul. 1999.

[167] T. Dandekar, B. Snel, M. Huynen, and P. Bork, "Conservation of gene order: a fingerprint of proteins that physically interact.," *Trends Biochem. Sci.*, vol. 23, no. 9, pp. 324–8, Sep. 1998.

[168] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates, "Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles," *Proc. Natl. Acad. Sci.*, vol. 96, no. 8, pp. 4285–4288, Apr. 1999.

[169] H. Yu, N. M. Luscombe, H. X. Lu, X. Zhu, Y. Xia, J.-D. J. Han, N. Bertin, S. Chung, M. Vidal, and M. Gerstein, "Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs.," *Genome Res.*, vol. 14, no. 6, pp. 1107–18, Jun. 2004.

[170] J. Wojcik and V. Schächter, "Protein-protein interaction map inference using interacting domain profile pairs.," *Bioinformatics*, vol. 17 Suppl 1, pp. S296–305, Jan. 2001.

[171] E. Sprinzak and H. Margalit, "Correlated sequence-signatures as markers of protein-protein interaction.," *J. Mol. Biol.*, vol. 311, no. 4, pp. 681–92, Aug. 2001.

[172] F. Pazos, M. Helmer-Citterich, G. Ausiello, and A. Valencia, "Correlated mutations contain information about protein-protein interaction.," *J. Mol. Biol.*, vol. 271, no. 4, pp. 511–23, Aug. 1997.

[173] D. Juan, F. Pazos, and A. Valencia, "High-confidence prediction of global interactomes based on genome-wide coevolutionary networks.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, no. 3, pp. 934–9, Jan. 2008.

[174] I. H. Moal and P. A. Bates, "Kinetic rate constant prediction supports the conformational selection mechanism of protein binding.," *PLoS Comput. Biol.*, vol. 8, no. 1, p. e1002351, Jan. 2012.

[175] M. N. Wass, G. Fuentes, C. Pons, F. Pazos, and A. Valencia, "Towards the prediction of protein interaction partners using physical docking.," *Mol. Syst. Biol.*, vol. 7, p. 469, Mar. 2011.

[176] W. S. Valdar and J. M. Thornton, "Protein-protein interfaces: analysis of amino acid conservation in homodimers.," *Proteins*, vol. 42, no. 1, pp. 108–24, Jan. 2001.

[177] J. Hoskins, S. Lovell, and T. L. Blundell, "An algorithm for predicting protein-protein interaction sites: Abnormally exposed amino acid residues and secondary structure elements.," *Protein Sci.*, vol. 15, no. 5, pp. 1017–29, May 2006.

[178] S. Jones and J. M. Thornton, "Prediction of protein-protein interaction sites using patch analysis.," *J. Mol. Biol.*, vol. 272, no. 1, pp. 133–43, Sep. 1997.

[179] Y. Ofran and B. Rost, "Predicted protein-protein interaction sites from local sequence information.," *FEBS Lett.*, vol. 544, no. 1–3, pp. 236–9, Jun. 2003.

[180] J. Fernandez-Recio, M. Totrov, C. Skorodumov, and R. Abagyan, "Optimal docking area: a new method for predicting protein-protein interaction sites.," *Proteins*, vol. 58, no. 1, pp. 134–43, Jan. 2005.

[181] Y. Ofran and B. Rost, "ISIS: interaction sites identified from sequence.," *Bioinformatics*, vol. 23, no. 2, pp. e13–6, Jan. 2007.

[182] J. Segura, P. F. Jones, and N. Fernandez-Fuentes, "Improving the prediction of protein binding sites by combining heterogeneous data and Voronoi diagrams.," *BMC Bioinformatics*, vol. 12, p. 352, Jan. 2011.

[183] U. Göbel, C. Sander, R. Schneider, and A. Valencia, "Correlated mutations and residue contacts in proteins.," *Proteins*, vol. 18, no. 4, pp. 309–17, May 1994.

[184] I. Halperin, H. Wolfson, and R. Nussinov, "Correlated mutations: advances and limitations. A study on fusion proteins and on the Cohesin-Dockerin families.," *Proteins*, vol. 63, no. 4, pp. 832–45, Jun. 2006.

[185] A. Henschel, C. Winter, W. Kim, and M. Schroeder, "Using structural motif descriptors for sequence-based binding site prediction," *BMC Bioinformatics*, vol. 8, no. Suppl 4, p. S5, Jan. 2007.

[186] A. A. Bogan and K. S. Thorn, "Anatomy of hot spots in protein interfaces.," *J. Mol. Biol.*, vol. 280, no. 1, pp. 1–9, Jul. 1998.

[187] Z. Hu, B. Ma, H. Wolfson, and R. Nussinov, "Conservation of polar residues as hot spots at protein interfaces.," *Proteins*, vol. 39, no. 4, pp. 331–42, Jun. 2000.

[188] R. Aragues, A. Sali, J. Bonet, M. A. Marti-Renom, and B. Oliva, "Characterization of protein hubs by inferring interacting motifs from protein interactions.," *PLoS Comput. Biol.*, vol. 3, no. 9, pp. 1761–71, Sep. 2007.

[189] J. Garcia-Garcia, J. Bonet, E. Guney, O. Fornes, J. Planas, and B. Oliva, "Networks of protein-protein interactions: from uncertainty to molecular details," *Mol. Inform.*, vol. 31, no. 5, pp. 342–362, May 2012.

[190] C. Alfarano, C. E. Andrade, K. Anthony, N. Bahroos, M. Bajec, K. Bantoft, D. Betel, B. Bobechko, K. Boutilier, E. Burgess, K. Buzadzija, R. Cavero, C. D'Abreo, I. Donaldson, D. Dorairajoo, M. J. Dumontier, M. R. Dumontier, V. Earles, R. Farrall, H. Feldman, E. Garderman, Y. Gong, R. Gonzaga, V. Grytsan, E. Gryz, V. Gu, E. Haldorsen, A. Halupa, R. Haw, A. Hrvojic, L. Hurrell, R. Isserlin, F. Jack, F. Juma, A. Khan, T. Kon, S. Konopinsky, V. Le, E. Lee, S. Ling, M. Magidin, J. Moniakis, J. Montojo, S. Moore, B. Muskat, I. Ng, J. P. Paraiso, B. Parker, G. Pintilie, R. Pirone, J. J. Salama, S. Sgro, T. Shan, Y. Shu, J. Siew, D. Skinner, K. Snyder, R. Stasiuk, D. Strumpf, B. Tuekam, S. Tao, Z. Wang, M. White, R. Willis, C. Wolting, S. Wong, A. Wrong, C. Xin, R. Yao, B. Yates, S. Zhang, K. Zheng, T. Pawson, B. F. F. Ouellette, and C. W. V Hogue, "The Biomolecular Interaction Network Database and

related tools 2005 update.," *Nucleic Acids Res.*, vol. 33, no. Database issue, pp. D418–24, Jan. 2005.

[191] A. Chatr-Aryamontri, B.-J. Breitkreutz, R. Oughtred, L. Boucher, S. Heinicke, D. Chen, C. Stark, A. Breitkreutz, N. Kolas, L. O'Donnell, T. Reguly, J. Nixon, L. Ramage, A. Winter, A. Sellam, C. Chang, J. Hirschman, C. Theesfeld, J. Rust, M. S. Livstone, K. Dolinski, and M. Tyers, "The BioGRID interaction database: 2015 update.," *Nucleic Acids Res.*, vol. 43, no. Database issue, pp. D470–8, Jan. 2015.

[192] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg, "The Database of Interacting Proteins: 2004 update.," *Nucleic Acids Res.*, vol. 32, no. Database issue, pp. D449–51, Jan. 2004.

[193] T. S. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. Harrys Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady, and A. Pandey, "Human Protein Reference Database--2009 update," *Nucleic Acids Res.*, vol. 37, no. Database, pp. D767–D772, Jan. 2009.

[194] S. Orchard, M. Ammari, B. Aranda, L. Breuza, L. Briganti, F. Broackes-Carter, N. H. Campbell, G. Chavali, C. Chen, N. del-Toro, M. Duesbury, M. Dumousseau, E. Galeota, U. Hinz, M. Iannuccelli, S. Jagannathan, R. Jimenez, J. Khadake, A. Lagreid, L. Licata, R. C. Lovering, B. Meldal, A. N. Melidoni, M. Milagros, D. Peluso, L. Perfetto, P. Porras, A. Raghunath, S. Ricard-Blum, B. Roechert, A. Stutz, M. Tognolli, K. van Roey, G. Cesareni, and H. Hermjakob, "The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases.," *Nucleic Acids Res.*, vol. 42, no. Database issue, pp. D358–63, Jan. 2014.

[195] H. W. Mewes, A. Ruepp, F. Theis, T. Rattei, M. Walter, D. Frishman, K. Suhre, M. Spannagl, K. F. X. Mayer, V. Stümpflen, and A. Antonov, "MIPS: curated databases and comprehensive secondary data resources in 2010.," *Nucleic Acids Res.*, vol. 39, no. Database issue, pp. D220–4, Jan. 2011.

[196] U. Güldener, M. Münsterkötter, M. Oesterheld, P. Pagel, A. Ruepp, H.-W. Mewes, and V. Stümpflen, "MPact: the MIPS protein interaction resource on yeast.," *Nucleic Acids Res.*, vol. 34, no. Database issue, pp. D436–41, Jan. 2006.

[197] J. Köhler, J. Baumbach, J. Taubert, M. Specht, A. Skusa, A. Rüegg, C. Rawlings, P. Verrier, and S. Philippi, "Graph-based analysis and visualization of experimental results with ONDEX.," *Bioinformatics*, vol. 22, no. 11, pp. 1383–90, Jun. 2006.

[198] R. Aragues, D. Jaeggi, and B. Oliva, "PIANA: protein interactions and network analysis.," *Bioinformatics*, vol. 22, no. 8, pp. 1015–7, Apr. 2006.

[199] J. Garcia-Garcia, E. Guney, R. Aragues, J. Planas-Iglesias, and B. Oliva, "Biana: a software framework for compiling biological interactions and analyzing networks.," *BMC Bioinformatics*, vol. 11, p. 56, Jan. 2010.

[200] M. Dreze, D. Monachello, C. Lurin, M. E. Cusick, D. E. Hill, M. Vidal, and P. Braun, "High-quality binary interactome mapping.," *Methods Enzymol.*, vol. 470, pp. 281–315, Jan. 2010.

[201] R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. J. Krogan, S. Chung, A. Emili, M. Snyder, J. F. Greenblatt, and M. Gerstein, "A Bayesian networks approach for predicting protein-protein interactions from genomic data.," *Science*, vol. 302, no. 5644, pp. 449–53, Oct. 2003.

[202] K. Venkatesan, J.-F. Rual, A. Vazquez, U. Stelzl, I. Lemmens, T. Hirozane-Kishikawa, T. Hao, M. Zenkner, X. Xin, K.-I. Goh, M. A. Yildirim, N. Simonis, K. Heinzmann, F. Gebreab, J. M. Sahalie, S. Cevik, C. Simon, A.-S. de Smet, E. Dann, A. Smolyar, A. Vinayagam, H. Yu, D. Szeto, H. Borick, A. Dricot, N. Klitgord, R. R. Murray, C. Lin, M. Lalowski, J. Timm, K. Rau, C. Boone, P. Braun, M. E. Cusick, F. P. Roth, D. E. Hill, J. Tavernier, E. E. Wanker, A.-L. Barabási, and M. Vidal, "An empirical framework for binary interactome mapping.," *Nat. Methods*, vol. 6, no. 1, pp. 83–90, Jan. 2009.

[203] G. T. Hart, A. K. Ramani, and E. M. Marcotte, "How complete are current yeast and human protein-interaction networks?," *Genome Biol.*, vol. 7, no. 11, p. 120, Jan. 2006.

[204] M. E. Cusick, H. Yu, A. Smolyar, K. Venkatesan, A.-R. Carvunis, N. Simonis, J.-F. Rual, H. Borick, P. Braun, M. Dreze, J. Vandenhaute, M. Galli, J. Yazaki, D. E. Hill, J. R. Ecker, F. P. Roth, and M. Vidal, "Literature-curated protein interaction datasets.," *Nat. Methods*, vol. 6, no. 1, pp. 39–46, Jan. 2009.

[205] W.-H. Jang, S.-H. Jung, and D.-S. Han, "A computational model for predicting protein interactions based on multidomain collaboration.," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 9, no. 4, pp. 1081–90, Jan. 2012.

[206] J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li, and H. Jiang, "Predicting protein-protein interactions based only on sequences information.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 104, no. 11, pp. 4337–41, Mar. 2007.

[207] C.-Y. Yu, L.-C. Chou, and D. T.-H. Chang, "Predicting protein-protein interactions in unbalanced data using the primary structure of proteins.," *BMC Bioinformatics*, vol. 11, p. 167, Jan. 2010.

[208] L. G. Trabuco, M. J. Betts, and R. B. Russell, "Negative protein-protein interaction datasets derived from large-scale two-hybrid experiments.," *Methods*, vol. 58, no. 4, pp. 343–8, Dec. 2012.

[209] P. Smialowski, P. Pagel, P. Wong, B. Brauner, I. Dunger, G. Fobo, G. Frishman, C. Montrone, T. Rattei, D. Frishman, and A. Ruepp, "The Negatome database: a reference set of non-interacting protein pairs.," *Nucleic Acids Res.*, vol. 38, no. Database issue, pp. D540–4, Jan. 2010.

[210] M. R. Chikka, D. D. McCabe, H. M. Tyra, and D. T. Rutkowski, "C/EBP homologous protein (CHOP) contributes to suppression of metabolic genes during endoplasmic reticulum stress in the liver.," *J. Biol. Chem.*, vol. 288, no. 6, pp. 4405–15, Feb. 2013.

[211] V. Matys, O. V Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A. E. Kel, and E. Wingender, "TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes.," *Nucleic Acids Res.*, vol. 34, no. Database issue, pp. D108–110, Jan. 2006.

[212] O. L. Griffith, S. B. Montgomery, B. Bernier, B. Chu, K. Kasaian, S. Aerts, S. Mahony, M. C. Sleumer, M. Bilenky, M. Haeussler, M. Griffith, S. M. Gallo, B. Giardine, B. Hooghe, P. Van Loo, E. Blanco, A. Ticoll, S. Lithwick, E. Portales-Casamar, I. J. Donaldson, G. Robertson, C. Wadelius, P. De Bleser, D. Vlieghe, M. S. Halfon, W. Wasserman, R. Hardison, C. M. Bergman, and S. J. M. Jones, "ORegAnno: an open-access community-driven resource for regulatory annotation.," *Nucleic Acids Res.*, vol. 36, no. Database issue, pp. D107–13, Jan. 2008.

[213] E. Portales-Casamar, D. Arenillas, J. Lim, M. I. Swanson, S. Jiang, A. McCallum, S. Kirov, and W. W. Wasserman, "The PAZAR database of gene regulatory information coupled to the ORCA toolkit for the study of regulatory sequences.," *Nucleic Acids Res.*, vol. 37, no. Database issue, pp. D54–60, Jan. 2009.

[214] C. E. Grant, T. L. Bailey, and W. S. Noble, "FIMO: scanning for occurrences of a given motif.," *Bioinformatics*, vol. 27, no. 7, pp. 1017–8, Apr. 2011.

[215] S. Roepcke, S. Grossmann, S. Rahmann, and M. Vingron, "T-Reg Comparator: an analysis tool for the comparison of position weight matrices.," *Nucleic Acids Res.*, vol. 33, no. Web Server issue, pp. W438–41, Jul. 2005.

[216] S. Gupta, J. A. Stamatoyannopoulos, T. L. Bailey, and W. S. Noble, "Quantifying similarity between motifs.," *Genome Biol.*, vol. 8, no. 2, p. R24, Jan. 2007.

[217] A. J. Butte, P. Tamayo, D. Slonim, T. R. Golub, and I. S. Kohane, "Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 97, no. 22, pp. 12182–6, Oct. 2000.

[218] C. E. Shannon, "Communication Theory of Secrecy Systems*," *Bell Syst. Tech. J.*, vol. 28, no. 4, pp. 656–715, Oct. 1949.

[219] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera, and A. Califano, "ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context.," *BMC Bioinformatics*, vol. 7 Suppl 1, p. S7, Jan. 2006.

[220] B. Zhang and S. Horvath, "A general framework for weighted gene co-expression network analysis.," *Stat. Appl. Genet. Mol. Biol.*, vol. 4, p. Article17, Jan. 2005.

[221] S. L. Lauritzen, *Graphical Models*. 1996.

[222] A. Dobra, C. Hans, B. Jones, J. R. Nevins, G. Yao, and M. West, "Sparse graphical models for exploring gene expression data," *J. Multivar. Anal.*, vol. 90, no. 1, pp. 196–212, Jul. 2004.

[223] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso.," *Biostatistics*, vol. 9, no. 3, pp. 432–41, Jul. 2008.

[224] A. de la Fuente, N. Bing, I. Hoeschele, and P. Mendes, "Discovery of meaningful associations in genomic data using partial correlation coefficients.," *Bioinformatics*, vol. 20, no. 18, pp. 3565–74, Dec. 2004.

[225] M. C. Robert Castelo, Alberto Roverato, "A robust procedure for gaussian graphical model search from microarray data with p larger than n," *J. Mach. Learn. Res.*, vol. 7, pp. 2621–2650, 2006.

[226] R. Castelo and A. Roverato, "Reverse engineering molecular regulatory networks from microarray data with qp-graphs.," *J. Comput. Biol.*, vol. 16, no. 2, pp. 213–27, Mar. 2009.

[227] C.-H. Yeang, T. Ideker, and T. Jaakkola, "Physical network models.," *J. Comput. Biol.*, vol. 11, no. 2–3, pp. 243–62, Jan. 2004.

[228] O. Ourfali, T. Shlomi, T. Ideker, E. Ruppin, and R. Sharan, "SPINE: a framework for signaling-regulatory pathway inference from cause-effect experiments.," *Bioinformatics*, vol. 23, no. 13, pp. i359–66, Jul. 2007.

[229] T. Peleg, N. Yosef, E. Ruppin, and R. Sharan, "Network-free inference of knockout effects in yeast.," *PLoS Comput. Biol.*, vol. 6, no. 1, p. e1000635, Jan. 2010.

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

[231] S.-S. C. Huang and E. Fraenkel, "Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks.," *Sci. Signal.*, vol. 2, no. 81, p. ra40, Jan. 2009.

[232] N. Yosef, L. Ungar, E. Zalckvar, A. Kimchi, M. Kupiec, E. Ruppin, and R. Sharan, "Toward accurate reconstruction of functional protein networks.," *Mol. Syst. Biol.*, vol. 5, p. 248, Jan. 2009.

[233] M. Bailly-Bechet, C. Borgs, A. Braunstein, J. Chayes, A. Dagkessamanskaia, J.-M. François, and R. Zecchina, "Finding undetected protein associations in cell signaling by belief propagation.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 108, no. 2, pp. 882–7, Jan. 2011.

[234] G. Giaever, A. M. Chu, L. Ni, C. Connelly, L. Riles, S. Véronneau, S. Dow, A. Lucau-Danila, K. Anderson, B. André, A. P. Arkin, A. Astromoff, M. El-Bakkoury, R. Bangham, R. Benito, S. Brachat, S. Campanaro, M. Curtiss, K. Davis, A. Deutschbauer, K.-D. Entian, P. Flaherty, F. Foury, D. J. Garfinkel, M. Gerstein, D. Gotte, U. Güldener, J. H. Hegemann, S. Hempel, Z. Herman, D. F. Jaramillo, D. E. Kelly, S. L. Kelly, P. Kötter, D. LaBonte, D. C. Lamb, N. Lan, H. Liang, H. Liao, L. Liu, C. Luo, M. Lussier, R. Mao, P. Menard, S. L. Ooi, J. L. Revuelta, C. J. Roberts, M. Rose, P. Ross-Macdonald, B. Scherens, G. Schimmack, B. Shafer, D. D. Shoemaker, S. Sookhai-Mahadeo, R. K. Storms, J. N. Strathern, G. Valle, M. Voet, G. Volckaert, C. Wang, T. R. Ward, J. Wilhelmy, E. A. Winzeler, Y. Yang, G. Yen, E. Youngman, K. Yu, H. Bussey, J. D. Boeke, M. Snyder, P. Philippsen, R. W. Davis, and M. Johnston, "Functional profiling of the Saccharomyces cerevisiae genome.," *Nature*, vol. 418, no. 6896, pp. 387–91, Jul. 2002.

[235] R. Kafri, A. Bar-Even, and Y. Pilpel, "Transcription control reprogramming in genetic backup circuits.," *Nat. Genet.*, vol. 37, no. 3, pp. 295–9, Mar. 2005.

[236] R. Kafri, O. Dahan, J. Levy, and Y. Pilpel, "Preferential protection of protein interaction network hubs in yeast: evolved functionality of genetic redundancy.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, no. 4, pp. 1243–8, Jan. 2008.

[237] A. Gitter, Z. Siegfried, M. Klutstein, O. Fornes, B. Oliva, I. Simon, and Z. Bar-Joseph, "Backup in gene regulatory networks explains differences between binding and knockout results.," *Mol. Syst. Biol.*, vol. 5, p. 276, Jan. 2009.

[238] A. Gitter, Y. Lu, and Z. Bar-Joseph, "Computational methods for analyzing dynamic regulatory networks.," *Methods Mol. Biol.*, vol. 674, pp. 419–41, Jan. 2010.

[239] A. Gitter, M. Carmi, N. Barkai, and Z. Bar-Joseph, "Linking the signaling cascades and dynamic regulatory networks controlling stress responses.," *Genome Res.*, vol. 23, no. 2, pp. 365–76, Mar. 2013.

[240] J. Ernst, O. Vainas, C. T. Harbison, I. Simon, and Z. Bar-Joseph, "Reconstructing dynamic regulatory maps.," *Mol. Syst. Biol.*, vol. 3, no. 1, p. 74, Jan. 2007.

[241] E. E. Schadt, S. H. Friend, and D. A. Shaywitz, "A network view of disease and compound screening.," *Nat. Rev. Drug Discov.*, vol. 8, no. 4, pp. 286–95, Apr. 2009.

[242] M. Isalan, C. Lemerle, K. Michalodimitrakis, C. Horn, P. Beltrao, E. Raineri, M. Garriga-Canut, and L. Serrano, "Evolvability and hierarchy in rewired bacterial gene networks.," *Nature*, vol. 452, no. 7189, pp. 840–5, Apr. 2008.

[243] R. Pastor-Satorras and A. Vespignani, "Immunization of complex networks.," *Phys. Rev. E. Stat. Nonlin. Soft Matter Phys.*, vol. 65, no. 3 Pt 2A, p. 036104, Mar. 2002.

[244] H. Jeong, S. P. Mason, A. L. Barabási, and Z. N. Oltvai, "Lethality and centrality in protein networks.," *Nature*, vol. 411, no. 6833, pp. 41–2, May 2001.

[245] R. Albert, H. Jeong, and A. Barabasi, "Error and attack tolerance of complex networks," *Nature*, vol. 406, no. 6794, pp. 378–82, Jul. 2000.

[246] J.-D. J. Han, N. Bertin, T. Hao, D. S. Goldberg, G. F. Berriz, L. V Zhang, D. Dupuy, A. J. M. Walhout, M. E. Cusick, F. P. Roth, and M. Vidal, "Evidence for dynamically organized modularity in the yeast protein-protein interaction network.," *Nature*, vol. 430, no. 6995, pp. 88–93, Jul. 2004.

[247] A. L. Hopkins, "Network pharmacology: the next paradigm in drug discovery.," *Nat. Chem. Biol.*, vol. 4, no. 11, pp. 682–90, Nov. 2008.

[248] J. Park, D.-S. Lee, N. A. Christakis, and A.-L. Barabási, "The impact of cellular networks on disease comorbidity.," *Mol. Syst. Biol.*, vol. 5, p. 262, Jan. 2009.

[249] S. L. Ooi, X. Pan, B. D. Peyser, P. Ye, P. B. Meluh, D. S. Yuan, R. A. Irizarry, J. S. Bader, F. A. Spencer, and J. D. Boeke, "Global synthetic-lethality analysis and yeast functional profiling.," *Trends Genet.*, vol. 22, no. 1, pp. 56–63, Jan. 2006.

[250] M. C. Alles, M. Gardiner-Garden, D. J. Nott, Y. Wang, J. A. Foekens, R. L. Sutherland, E. A. Musgrove, and C. J. Ormandy, "Meta-analysis and gene set enrichment relative to er status reveal elevated activity of MYC and E2F in the 'basal' breast cancer subgroup.," *PLoS One*, vol. 4, no. 3, p. e4710, Jan. 2009.

[251] L. I. Furlong, "Human diseases through the lens of network biology.," *Trends Genet.*, vol. 29, no. 3, pp. 150–9, Mar. 2013.

[252] M. G. Kann, "Advances in translational bioinformatics: computational approaches for the hunting of disease genes.," *Brief. Bioinform.*, vol. 11, no. 1, pp. 96–110, Jan. 2010.

[253] L.-C. Tranchevent, F. B. Capdevila, D. Nitsch, B. De Moor, P. De Causmaecker, and Y. Moreau, "A guide to web tools to prioritize candidate genes.," *Brief. Bioinform.*, vol. 12, no. 1, pp. 22–32, Jan. 2011.

[254] E. Capriotti, N. L. Nehrt, M. G. Kann, and Y. Bromberg, "Bioinformatics for personal genome interpretation.," *Brief. Bioinform.*, vol. 13, no. 4, pp. 495–512, Jul. 2012.

[255] R. Karchin, "Next generation tools for the annotation of human SNPs.," *Brief. Bioinform.*, vol. 10, no. 1, pp. 35–52, Jan. 2009.

[256] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.," *Nat. Genet.*, vol. 25, no. 1, pp. 25–9, May 2000.

[257] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, and J. P. Mesirov, "Molecular signatures database (MSigDB) 3.0.," *Bioinformatics*, vol. 27, no. 12, pp. 1739–40, Jun. 2011.

[258] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe, "KEGG for integration and interpretation of large-scale molecular data sets.," *Nucleic Acids Res.*, vol. 40, no. Database issue, pp. D109–14, Jan. 2012.

[259] N. Salomonis, K. Hanspers, A. C. Zambon, K. Vranizan, S. C. Lawlor, K. D. Dahlquist, S. W. Doniger, J. Stuart, B. R. Conklin, and A. R. Pico, "GenMAPP 2: new features and resources for pathway analysis.," *BMC Bioinformatics*, vol. 8, p. 217, Jan. 2007.

[260] D. Croft, A. F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu, M. Caudy, P. Garapati, M. Gillespie, M. R. Kamdar, B. Jassal, S. Jupe, L. Matthews, B. May, S. Palatnik, K. Rothfels, V. Shamovsky, H. Song, M. Williams, E. Birney, H. Hermjakob, L. Stein, and P. D'Eustachio, "The Reactome pathway knowledgebase.," *Nucleic Acids Res.*, vol. 42, no. Database issue, pp. D472–7, Jan. 2014.

[261] R. Caspi, T. Altman, K. Dreher, C. A. Fulcher, P. Subhraveti, I. M. Keseler, A. Kothari, M. Krummenacker, M. Latendresse, L. A. Mueller, Q. Ong, S. Paley, A. Pujar, A. G. Shearer, M. Travers, D. Weerasinghe, P. Zhang, and P. D. Karp, "The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases.," *Nucleic Acids Res.*, vol. 40, no. Database issue, pp. D742–53, Jan. 2012.

[262] J. Xu and Y. Li, "Discovering disease-genes by topological features in human protein-protein interaction network.," *Bioinformatics*, vol. 22, no. 22, pp. 2800–5, Dec. 2006.

[263] M. Oti, B. Snel, M. A. Huynen, and H. G. Brunner, "Predicting disease genes using protein-protein interactions.," *J. Med. Genet.*, vol. 43, no. 8, pp. 691–8, Aug. 2006.

[264] K. Lage, E. O. Karlberg, Z. M. Størling, P. I. Olason, A. G. Pedersen, O. Rigina, A. M. Hinsby, Z. Tümer, F. Pociot, N. Tommerup, Y. Moreau, and S. Brunak, "A human phenome-interactome network of protein complexes implicated in genetic disorders.," *Nat. Biotechnol.*, vol. 25, no. 3, pp. 309–16, Mar. 2007.

[265] M. A. Pujana, J.-D. J. Han, L. M. Starita, K. N. Stevens, M. Tewari, J. S. Ahn, G. Rennert, V. Moreno, T. Kirchhoff, B. Gold, V. Assmann, W. M. Elshamy, J.-F. Rual, D. Levine, L. S. Rozek, R. S. Gelman, K. C. Gunsalus, R. A. Greenberg, B. Sobhian, N. Bertin, K. Venkatesan, N. Ayivi-Guedehoussou, X. Solé, P. Hernández, C. Lázaro, K. L. Nathanson, B. L. Weber, M. E. Cusick, D. E. Hill, K. Offit, D. M. Livingston, S. B. Gruber, J. D. Parvin, and M. Vidal, "Network modeling links breast cancer susceptibility and centrosome dysfunction.," *Nat. Genet.*, vol. 39, no. 11, pp. 1338–49, Dec. 2007.

[266] X. Wu, R. Jiang, M. Q. Zhang, and S. Li, "Network-based global inference of human disease genes.," *Mol. Syst. Biol.*, vol. 4, p. 189, Jan. 2008.

[267] R. Aragues, C. Sander, and B. Oliva, "Predicting cancer involvement of genes from heterogeneous data.," *BMC Bioinformatics*, vol. 9, p. 172, Jan. 2008.

[268] T. Milenkovic, V. Memisevic, A. K. Ganesan, and N. Przulj, "Systems-level cancer gene identification from protein interaction network topology applied to melanogenesis-related functional genomics data.," *J. R. Soc. Interface*, vol. 7, no. 44, pp. 423–37, Mar. 2010.

[269] S. Navlakha and C. Kingsford, "The power of protein interaction networks for associating genes with diseases.," *Bioinformatics*, vol. 26, no. 8, pp. 1057–63, Apr. 2010.

[270] L. Franke, H. van Bakel, L. Fokkens, E. D. de Jong, M. Egmont-Petersen, and C. Wijmenga, "Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes.," *Am. J. Hum. Genet.*, vol. 78, no. 6, pp. 1011–25, Jul. 2006.

[271] S. Köhler, S. Bauer, D. Horn, and P. N. Robinson, "Walking the interactome for prioritization of candidate disease genes.," *Am. J. Hum. Genet.*, vol. 82, no. 4, pp. 949–58, Apr. 2008.

[272] D. Smedley, S. Köhler, J. C. Czeschik, J. Amberger, C. Bocchini, A. Hamosh, J. Veldboer, T. Zemojtel, and P. N. Robinson, "Walking the interactome for candidate prioritization in exome sequencing studies of Mendelian diseases.," *Bioinformatics*, vol. 30, no. 22, pp. 3215–22, Dec. 2014.

[273] Z. Dezso, Y. Nikolsky, T. Nikolskaya, J. Miller, D. Cherba, C. Webb, and A. Bugrim, "Identifying disease-specific genes based on their topological significance in protein networks.," *BMC Syst. Biol.*, vol. 3, p. 36, Jan. 2009.

[274] X. Ma, H. Lee, L. Wang, and F. Sun, "CGI: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data.," *Bioinformatics*, vol. 23, no. 2, pp. 215–21, Jan. 2007.

[275] D. Nitsch, J. P. Gonçalves, F. Ojeda, B. de Moor, and Y. Moreau, "Candidate gene prioritization by network analysis of differential expression using machine learning approaches.," *BMC Bioinformatics*, vol. 11, p. 460, Jan. 2010.

[276] Y.-Q. Qiu, S. Zhang, X.-S. Zhang, and L. Chen, "Detecting disease associated modules and prioritizing active genes based on high throughput data.," *BMC Bioinformatics*, vol. 11, p. 26, Jan. 2010.

[277] J. Chen, B. J. Aronow, and A. G. Jegga, "Disease candidate gene identification and prioritization using protein interaction networks.," *BMC Bioinformatics*, vol. 10, p. 73, Jan. 2009.

[278] O. Vanunu, O. Magger, E. Ruppin, T. Shlomi, and R. Sharan, "Associating genes and protein complexes with disease via network propagation.," *PLoS Comput. Biol.*, vol. 6, no. 1, p. e1000641, Jan. 2010.

[279] S. Erten, G. Bebek, R. M. Ewing, and M. Koyutürk, "DADA: Degree-Aware Algorithms for Network-Based Disease Gene Prioritization.," *BioData Min.*, vol. 4, p. 19, Jan. 2011.

[280] B. Linghu, E. S. Snitkin, Z. Hu, Y. Xia, and C. Delisi, "Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network.," *Genome Biol.*, vol. 10, no. 9, p. R91, Jan. 2009.

[281] U. Ala, R. M. Piro, E. Grassi, C. Damasco, L. Silengo, M. Oti, P. Provero, and F. Di Cunto, "Prediction of human disease genes by human-mouse conserved coexpression analysis.," *PLoS Comput. Biol.*, vol. 4, no. 3, p. e1000043, Mar. 2008.

[282] E. Guney and B. Oliva, "Exploiting protein-protein interaction networks for genome-wide disease-gene prioritization.," *PLoS One*, vol. 7, no. 9, p. e43557, Jan. 2012.

[283] P. Uetz, Y.-A. Dong, C. Zeretzke, C. Atzler, A. Baiker, B. Berger, S. V Rajagopala, M. Roupelieva, D. Rose, E. Fossum, and J. Haas, "Herpesviral protein networks and their interaction with the human proteome.," *Science*, vol. 311, no. 5758, pp. 239–42, Jan. 2006.

[284] M. Magrane and U. Consortium, "UniProt Knowledgebase: a hub of integrated protein data.," *Database (Oxford).*, vol. 2011, p. bar009, 2011.

[285] R. D. Finn, A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. L. Sonnhammer, J. Tate, and M. Punta, "Pfam: the protein families database.," *Nucleic Acids Res.*, vol. 42, no. Database issue, pp. D222–230, Jan. 2014.

[286] H. M. Berman, G. J. Kleywegt, H. Nakamura, and J. L. Markley, "The Protein Data Bank archive as an open data resource.," *J. Comput. Aided. Mol. Des.*, vol. 28, no. 10, pp. 1009–14, Oct. 2014.

[287] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, and A. Soboleva, "NCBI GEO: archive for functional genomics data sets--update.," *Nucleic Acids Res.*, vol. 41, no. Database issue, pp. D991–5, Jan. 2013.

[288] R. D. Finn, M. Marshall, and A. Bateman, "iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions.," *Bioinformatics*, vol. 21, no. 3, pp. 410–2, Mar. 2005.

[289] R. Mosca, A. Céol, A. Stein, R. Olivella, and P. Aloy, "3did: a catalog of domain-based interactions of known three-dimensional structure.," *Nucleic Acids Res.*, vol. 42, no. Database issue, pp. D374–9, Jan. 2014.

[290] T. Driscoll, M. D. Dyer, T. M. Murali, and B. W. Sobral, "PIG--the pathogen interaction gateway.," *Nucleic Acids Res.*, vol. 37, no. Database issue, pp. D647–50, Jan. 2009.

[291] L. R. Matthews, P. Vaglio, J. Reboul, H. Ge, B. P. Davis, J. Garrels, S. Vincent, and M. Vidal, "Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or 'interologs'.," *Genome Res.*, vol. 11, no. 12, pp. 2120–6, Dec. 2001.

[292] A. M. Wiles, M. Doderer, J. Ruan, T.-T. Gu, D. Ravi, B. Blackman, and A. J. R. Bishop, "Building and analyzing protein interactome networks by cross-species comparisons.," *BMC Syst. Biol.*, vol. 4, p. 36, Jan. 2010.

[293] C.-C. Chen, C.-Y. Lin, Y.-S. Lo, and J.-M. Yang, "PPISearch: a web server for searching homologous protein-protein interactions across multiple species.," *Nucleic Acids Res.*, vol. 37, no. Web Server issue, pp. W369–75, Jul. 2009.

[294] G. Gallone, T. I. Simpson, J. D. Armstrong, and A. P. Jarman, "Bio::Homology::InterologWalk--a Perl module to build putative protein-protein interaction networks through interolog mapping.," *BMC Bioinformatics*, vol. 12, p. 289, Jan. 2011.

[295] J. Garcia-Garcia, S. Schleker, J. Klein-Seetharaman, and B. Oliva, "BIPS: BIANA Interolog Prediction Server. A tool for protein-protein interaction inference.," *Nucleic Acids Res.*, vol. 40, no. Web Server issue, pp. W147–51, Jul. 2012.

[296] S. Kerrien, Y. Alam-Faruque, B. Aranda, I. Bancarz, A. Bridge, C. Derow, E. Dimmer, M. Feuermann, A. Friedrichsen, R. Huntley, C. Kohler, J. Khadake, C. Leroy, A. Liban, C. Lieftink, L. Montecchi-Palazzi, S. Orchard, J. Risse, K. Robbe, B. Roechert, D. Thorneycroft, Y. Zhang, R. Apweiler, and H. Hermjakob, "IntAct--open source resource for molecular interaction data.," *Nucleic Acids Res.*, vol. 35, no. Database issue, pp. D561–5, Jan. 2007.

[297] L. Licata, L. Briganti, D. Peluso, L. Perfetto, M. Iannuccelli, E. Galeota, F. Sacco, A. Palma, A. P. Nardozza, E. Santonico, L. Castagnoli, and G. Cesareni, "MINT, the molecular interaction database: 2012 update.," *Nucleic Acids Res.*, vol. 40, no. Database issue, pp. D857–61, Jan. 2012.

[298] R. Winnenburg, M. Urban, A. Beacham, T. K. Baldwin, S. Holland, M. Lindeberg, H. Hansen, C. Rawlings, K. E. Hammond-Kosack, and J. Köhler, "PHI-base update: additions to the pathogen host interaction database.," *Nucleic Acids Res.*, vol. 36, no. Database issue, pp. D572–6, Jan. 2008.

[299] G. D. Bader, D. Betel, and C. W. V Hogue, "BIND: the Biomolecular Interaction Network Database.," *Nucleic Acids Res.*, vol. 31, no. 1, pp. 248–50, Jan. 2003.

[300] A. Chatr-aryamontri, A. Ceol, D. Peluso, A. Nardozza, S. Panni, F. Sacco, M. Tinti, A. Smolyar, L. Castagnoli, M. Vidal, M. E. Cusick, and G. Cesareni, "VirusMINT: a viral protein interaction database.," *Nucleic Acids Res.*, vol. 37, no. Database issue, pp. D669–73, Jan. 2009.

[301] R. Kumar and B. Nanduri, "HPIDB--a unified resource for host-pathogen interactions.," *BMC Bioinformatics*, vol. 11 Suppl 6, p. S16, Jan. 2010.

[302] S. Durmuş Tekir, T. Çakır, E. Ardiç, A. S. Sayılırbaş, G. Konuk, M. Konuk, H. Sarıyer, A. Uğurlu, İ. Karadeniz, A. Özgür, F. E. Sevilgen, and K. Ö. Ülgen, "PHISTO: pathogen-host interaction search tool.," *Bioinformatics*, vol. 29, no. 10, pp. 1357–8, May 2013.

[303] G. V Paolini, R. H. B. Shapland, W. P. van Hoorn, J. S. Mason, and A. L. Hopkins, "Global mapping of pharmacological space.," *Nat. Biotechnol.*, vol. 24, no. 7, pp. 805–15, Jul. 2006.

[304] M. J. Keiser, V. Setola, J. J. Irwin, C. Laggner, A. I. Abbas, S. J. Hufeisen, N. H. Jensen, M. B. Kuijer, R. C. Matos, T. B. Tran, R. Whaley, R. A. Glennon, J. Hert, K. L. H. Thomas, D. D. Edwards, B. K. Shoichet, and B. L. Roth, "Predicting new molecular targets for known drugs.," *Nature*, vol. 462, no. 7270, pp. 175–81, Nov. 2009.

[305] M. Campillos, M. Kuhn, A.-C. Gavin, L. J. Jensen, and P. Bork, "Drug target identification using side-effect similarity.," *Science*, vol. 321, no. 5886, pp. 263–6, Jul. 2008.

[306] M. Kuhn, M. Campillos, I. Letunic, L. J. Jensen, and P. Bork, "A side effect resource to capture phenotypic effects of drugs.," *Mol. Syst. Biol.*, vol. 6, p. 343, Jan. 2010.

[307] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool.," *J. Mol. Biol.*, vol. 215, no. 3, pp. 403–10, Oct. 1990.

[308] J. Hert, M. J. Keiser, J. J. Irwin, T. I. Oprea, and B. K. Shoichet, "Quantifying the relationships among drug classes.," *J. Chem. Inf. Model.*, vol. 48, no. 4, pp. 755–65, Apr. 2008.

[309] M. Lukk, M. Kapushesky, J. Nikkilä, H. Parkinson, A. Goncalves, W. Huber, E. Ukkonen, and A. Brazma, "A global map of human gene expression.," *Nat. Biotechnol.*, vol. 28, no. 4, pp. 322–4, Apr. 2010.

[310] P. Y. Lum, J. M. J. Derry, and E. E. Schadt, "Integrative genomics and drug development.," *Pharmacogenomics*, vol. 10, no. 2, pp. 203–12, Feb. 2009.

[311] J. T. Dudley, R. Tibshirani, T. Deshpande, and A. J. Butte, "Disease signatures are robust across tissues and experiments.," *Mol. Syst. Biol.*, vol. 5, p. 307, Jan. 2009.

[312] C. Harrison, "Signatures for drug repositioning.," *Nat. Rev. Genet.*, vol. 12, no. 10, p. 668, Oct. 2011.

[313] Y. A. Lussier and J. L. Chen, "The emergence of genome-based drug repositioning.," *Sci. Transl. Med.*, vol. 3, no. 96, p. 96ps35, Aug. 2011.

[314] J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J.-P. Brunet, A. Subramanian, K. N. Ross, M. Reich, H. Hieronymus, G. Wei, S. A. Armstrong, S. J. Haggarty, P. A. Clemons, R. Wei, S. A. Carr, E. S. Lander, and T. R. Golub, "The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease.," *Science*, vol. 313, no. 5795, pp. 1929–35, Sep. 2006.

[315] J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J. Wilson, J. Lehár, G. V Kryukov, D. Sonkin, A. Reddy, M. Liu, L. Murray, M. F. Berger, J. E. Monahan, P. Morais, J. Meltzer, A. Korejwa, J. Jané-Valbuena, F. A. Mapa, J. Thibault, E. Bric-Furlong, P. Raman, A. Shipway, I. H. Engels, J. Cheng, G. K. Yu, J. Yu, P. Aspesi, M. de Silva, K. Jagtap, M. D. Jones, L. Wang, C. Hatton, E. Palescandolo, S. Gupta, S. Mahan, C. Sougnez, R. C. Onofrio, T. Liefeld, L. MacConaill, W. Winckler, M. Reich, N. Li, J. P. Mesirov, S. B. Gabriel, G. Getz, K. Ardlie, V. Chan, V. E. Myer, B. L. Weber, J. Porter, M. Warmuth, P. Finan, J. L. Harris, M. Meyerson, T. R. Golub, M. P. Morrissey, W. R. Sellers, R. Schlegel, and L. A. Garraway, "The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.," *Nature*, vol. 483, no. 7391, pp. 603–7, Mar. 2012.

[316] M. Sirota, J. T. Dudley, J. Kim, A. P. Chiang, A. A. Morgan, A. Sweet-Cordero, J. Sage, and A. J. Butte, "Discovery and preclinical validation of drug indications using compendia of public gene expression data.," *Sci. Transl. Med.*, vol. 3, no. 96, p. 96ra77, Aug. 2011.

[317] J. T. Dudley, M. Sirota, M. Shenoy, R. K. Pai, S. Roedder, A. P. Chiang, A. A. Morgan, M. M. Sarwal, P. J. Pasricha, and A. J. Butte, "Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease.," *Sci. Transl. Med.*, vol. 3, no. 96, p. 96ra76, Aug. 2011.

[318] A. Gottlieb, G. Y. Stein, Y. Oron, E. Ruppin, and R. Sharan, "INDI: a computational framework for inferring drug interactions and their associated recommendations.," *Mol. Syst. Biol.*, vol. 8, p. 592, Jan. 2012.

[319] V. Law, C. Knox, Y. Djoumbou, T. Jewison, A. C. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson, V. Neveu, A. Tang, G. Gabriel, C. Ly, S. Adamjee, Z. T. Dame, B. Han, Y. Zhou, and D. S. Wishart, "DrugBank 4.0: shedding new light on drug metabolism.," *Nucleic Acids Res.*, vol. 42, no. Database issue, pp. D1091–7, Jan. 2014.

[320] World Organisation for Animal Health, "Manual of Diagnostic Tests and Vaccines for Terrestrial Animals," 2008.

[321] E. Scallan, R. M. Hoekstra, F. J. Angulo, R. V Tauxe, M.-A. Widdowson, S. L. Roy, J. L. Jones, and P. M. Griffin, "Foodborne illness acquired in the United States--major pathogens.," *Emerg. Infect. Dis.*, vol. 17, no. 1, pp. 7–15, Jan. 2011.

[322] M. B. Cooley, W. G. Miller, and R. E. Mandrell, "Colonization of Arabidopsis thaliana with Salmonella enterica and enterohemorrhagic Escherichia coli O157:H7 and competition by Enterobacter asburiae.," *Appl. Environ. Microbiol.*, vol. 69, no. 8, pp. 4915–26, Aug. 2003.

[323] A. L. Iniguez, Y. Dong, H. D. Carter, B. M. M. Ahmer, J. M. Stone, and E. W. Triplett, "Regulation of enteric endophytic bacterial colonization by plant defenses.," *Mol. Plant. Microbe. Interact.*, vol. 18, no. 2, pp. 169–78, Feb. 2005.

[324] A. Schikora, A. Carreri, E. Charpentier, and H. Hirt, "The dark side of the salad: Salmonella typhimurium overcomes the innate immune response of Arabidopsis thaliana and shows an endopathogenic lifestyle.," *PLoS One*, vol. 3, no. 5, p. e2279, Jan. 2008.

[325] C. N. Berger, R. K. Shaw, D. J. Brown, H. Mather, S. Clare, G. Dougan, M. J. Pallen, and G. Frankel, "Interaction of Salmonella enterica with basil and other salad leaves.," *ISME J.*, vol. 3, no. 2, pp. 261–5, Feb. 2009.

[326] Y. Kroupitski, D. Golberg, E. Belausov, R. Pinto, D. Swartzberg, D. Granot, and S. Sela, "Internalization of Salmonella enterica in leaves is induced by light and involves chemotaxis and penetration through open stomata.," *Appl. Environ. Microbiol.*, vol. 75, no. 19, pp. 6076–86, Oct. 2009.

[327] J. D. Barak, L. C. Kramer, and L. Hao, "Colonization of tomato plants by Salmonella enterica is cultivar dependent, and type 1 trichomes are preferred colonization sites.," *Appl. Environ. Microbiol.*, vol. 77, no. 2, pp. 498–504, Jan. 2011.

[328] C. N. Berger, D. J. Brown, R. K. Shaw, F. Minuzzi, B. Feys, and G. Frankel, "Salmonella enterica strains belonging to O serogroup 1,3,19 induce chlorosis and wilting of Arabidopsis thaliana leaves.," *Environ. Microbiol.*, vol. 13, no. 5, pp. 1299–308, May 2011.

[329] D. Golberg, Y. Kroupitski, E. Belausov, R. Pinto, and S. Sela, "Salmonella Typhimurium internalization is variable in leafy vegetables and fresh herbs.," *Int. J. Food Microbiol.*, vol. 145, no. 1, pp. 250–7, Jan. 2011.

[330] A. Schikora, I. Virlogeux-Payant, E. Bueso, A. V Garcia, T. Nilau, A. Charrier, S. Pelletier, P. Menanteau, M. Baccarini, P. Velge, and H. Hirt, "Conservation of Salmonella infection mechanisms in plants and animals.," *PLoS One*, vol. 6, no. 9, p. e24112, Jan. 2011.

[331] J. D. Barak and B. K. Schroeder, "Interrelationships of food safety and plant pathology: the life cycle of human pathogens on plants.," *Annu. Rev. Phytopathol.*, vol. 50, pp. 241–66, Jan. 2012.

[332] N. Shirron and S. Yaron, "Active suppression of early immune response in tobacco by the human pathogen Salmonella Typhimurium.," *PLoS One*, vol. 6, no. 4, p. e18855, Jan. 2011.

[333] D. Roy, S. Panchal, B. A. Rosa, and M. Melotto, "Escherichia coli O157:H7 induces stronger plant immunity than Salmonella enterica Typhimurium SL1344.," *Phytopathology*, vol. 103, no. 4, pp. 326–32, Apr. 2013.

[334] C. Zipfel, "Early molecular events in PAMP-triggered immunity.," *Curr. Opin. Plant Biol.*, vol. 12, no. 4, pp. 414–20, Aug. 2009.

[335] J. E. Galán, "Common themes in the design and function of bacterial effectors.," *Cell Host Microbe*, vol. 5, no. 6, pp. 571–9, Jun. 2009.

[336] I. Behlau and S. I. Miller, "A PhoP-repressed gene promotes Salmonella typhimurium invasion of epithelial cells.," *J. Bacteriol.*, vol. 175, no. 14, pp. 4475–84, Jul. 1993.

[337] J. M. Pawelek, S. Sodi, A. K. Chakraborty, J. T. Platt, S. Miller, D. W. Holden, M. Hensel, and K. B. Low, "Salmonella pathogenicity island-2 and anticancer activity in mice.," *Cancer Gene Ther.*, vol. 9, no. 10, pp. 813–8, Oct. 2002.

[338] A. V. Garcia, A. Charrier, A. Schikora, J. Bigeard, S. Pateyron, M.-L. L. de Tauzia-Moreau, A. Evrard, A. Mithöfer, M. L. Martin-Magniette, I. Virlogeux-Payant, and H. Hirt, "Salmonella enterica flagellin is recognized via FLS2 and activates PAMP-triggered immunity in Arabidopsis thaliana," *Mol. Plant*, vol. 7, no. 4, pp. 657–74, Nov. 2013.

[339] M. N. Arbeitman, E. E. M. Furlong, F. Imam, E. Johnson, B. H. Null, B. S. Baker, M. A. Krasnow, M. P. Scott, R. W. Davis, and K. P. White, "Gene expression during the life cycle of Drosophila melanogaster.," *Science (80-.).*, vol. 297, no. 5590, pp. 2270–2275, Sep. 2002.

[340] G. Laenen, L. Thorrez, D. Börnigen, and Y. Moreau, "Finding the targets of a drug by integration of gene expression data with a protein interaction network.," *Mol. Biosyst.*, vol. 9, no. 7, pp. 1676–85, Jul. 2013.

[341] H. Shatkay, S. Edwards, W. J. Wilbur, and M. Boguski, "Genes, themes and microarrays: using information retrieval for large-scale gene analysis.," *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, vol. 8, pp. 317–328, Jan. 2000.

[342] B. Futcher, G. I. Latter, P. Monardo, C. S. McLaughlin, and J. I. Garrels, "A sampling of the yeast proteome.," *Mol. Cell. Biol.*, vol. 19, no. 11, pp. 7357–68, Nov. 1999.

[343] D. Greenbaum, R. Jansen, and M. Gerstein, "Analysis of mRNA expression and protein abundance data: an approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts.," *Bioinformatics*, vol. 18, no. 4, pp. 585–96, Apr. 2002.

[344] I. A. Maraziotis, K. Dimitrakopoulou, and A. Bezerianos, "Growing functional modules from a seed protein via integration of protein interaction and gene expression data.," *BMC Bioinformatics*, vol. 8, no. 1, p. 408, Jan. 2007.

[345] R. Jansen, D. Greenbaum, and M. Gerstein, "Relating whole-genome expression data with protein-protein interactions.," *Genome Res.*, vol. 12, no. 1, pp. 37–46, Jan. 2002.

[346] S. Tornow and H. W. Mewes, "Functional modules by relating protein interaction networks and gene expression.," *Nucleic Acids Res.*, vol. 31, no. 21, pp. 6283–9, Nov. 2003.

[347] Y. Moreau and L.-C. Tranchevent, "Computational tools for prioritizing candidate genes: boosting disease gene discovery.," *Nat. Rev. Genet.*, vol. 13, no. 8, pp. 523–536, Aug. 2012.

[348] R. M. Piro and F. Di Cunto, "Computational approaches to disease-gene prediction: rationale, classification and successes.," *FEBS J.*, vol. 279, no. 5, pp. 678–96, Mar. 2012.

[349] D. Börnigen, L.-C. Tranchevent, F. Bonachela-Capdevila, K. Devriendt, B. De Moor, P. De Causmaecker, and Y. Moreau, "An unbiased evaluation of gene prioritization tools.," *Bioinformatics*, vol. 28, no. 23, pp. 3081–3088, Dec. 2012.

[350] A. Gitter and Z. Bar-Joseph, "Identifying proteins controlling key disease signaling pathways.," *Bioinformatics*, vol. 29, no. 13, pp. i227–36, Jul. 2013.

[351] P. Hilson, J. Allemeersch, T. Altmann, S. Aubourg, A. Avon, J. Beynon, R. P. Bhalerao, F. Bitton, M. Caboche, B. Cannoot, V. Chardakov, C. Cognet-Holliger, V. Colot, M. Crowe, C. Darimont, S. Durinck, H. Eickhoff, A. F. de Longevialle, E. E. Farmer, M. Grant, M. T. R. Kuiper, H. Lehrach, C. Léon, A. Leyva, J. Lundeberg, C. Lurin, Y. Moreau, W. Nietfeld, J. Paz-Ares, P. Reymond, P. Rouzé, G. Sandberg, M. D. Segura, C. Serizet, A. Tabrett, L. Taconnat, V. Thareau, P. Van Hummelen, S. Vercruysse, M. Vuylsteke, M. Weingartner, P. J. Weisbeek, V.

Wirta, F. R. A. Wittink, M. Zabeau, and I. Small, "Versatile gene-specific sequence tags for Arabidopsis functional genomics: transcript profiling and reverse genetics applications.," *Genome Res.*, vol. 14, no. 10B, pp. 2176–2189, Oct. 2004.

[352] G. Sclep, J. Allemeersch, R. Liechti, B. De Meyer, J. Beynon, R. Bhalerao, Y. Moreau, W. Nietfeld, J.-P. Renou, P. Reymond, M. T. Kuiper, and P. Hilson, "CATMA, a comprehensive genome-scale resource for silencing and transcript profiling of Arabidopsis genes.," *BMC Bioinformatics*, vol. 8, p. 400, Jan. 2007.

[353] C. Li and W. Hung Wong, "Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application.," *Genome Biol.*, vol. 2, no. 8, p. RESEARCH0032, Jan. 2001.

[354] C. Li, W. H. Wong, and W. Hung Wong, "Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection," *Proc. Natl. Acad. Sci.*, vol. 98, no. 1, pp. 31–36, Jan. 2001.

[355] R. V Davuluri, H. Sun, S. K. Palaniswamy, N. Matthews, C. Molina, M. Kurtz, and E. Grotewold, "AGRIS: Arabidopsis gene regulatory information server, an information resource of Arabidopsis cis-regulatory elements and transcription factors.," *BMC Bioinformatics*, vol. 4, p. 25, Jun. 2003.

[356] S. K. Palaniswamy, S. James, H. Sun, R. S. Lamb, R. V Davuluri, and E. Grotewold, "AGRIS and AtRegNet. a platform to link cis-regulatory elements and transcription factors into regulatory networks.," *Plant Physiol.*, vol. 140, no. 3, pp. 818–829, Mar. 2006.

[357] A. Yilmaz, M. K. Mejia-Guerra, K. Kurz, X. Liang, L. Welch, and E. Grotewold, "AGRIS: the Arabidopsis Gene Regulatory Information Server, an update.," *Nucleic Acids Res.*, vol. 39, no. Database issue, pp. D1118–22, Jan. 2011.

[358] M. H. Schulz, W. E. Devanny, A. Gitter, S. Zhong, J. Ernst, and Z. Bar-Joseph, "DREM 2.0: Improved reconstruction of dynamic regulatory networks from time-series expression data.," *BMC Syst. Biol.*, vol. 6, p. 104, Jan. 2012.

[359] E. Guney and B. Oliva, "Analysis of the robustness of network-based disease-gene prioritization methods reveals redundancy in the human interactome and functional diversity of disease-genes.," *PLoS One*, vol. 9, no. 4, p. e94686, Jan. 2014.

[360] E. Guney, J. Garcia-Garcia, and B. Oliva, "GUILDify: a web server for phenotypic characterization of genes through biological data integration and network-based prioritization algorithms.," *Bioinformatics*, vol. 30, no. 12, pp. 1789–90, Mar. 2014.

[361] A. A. Hagberg, D. A. Schult, and P. J. Swart, "Exploring Network Structure, Dynamics, and Function using NetworkX," in *Proceedings of the 7th Python in Science Conference (SciPy 2008)*, 2008, pp. 11–15.

[362] B. Rost, "Twilight zone of protein sequence alignments.," *Protein Eng.*, vol. 12, no. 2, pp. 85–94, Feb. 1999.

[363] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.," *Nucleic Acids Res.*, vol. 25, no. 17, pp. 3389–3402, 1997.

[364] R. D. Finn, J. Clements, and S. R. Eddy, "HMMER web server: interactive sequence similarity searching.," *Nucleic Acids Res.*, vol. 39, no. Web Server issue, pp. W29–37, Jul. 2011.

[365] J. Ernst and Z. Bar-Joseph, "STEM: a tool for the analysis of short time series gene expression data.," *BMC Bioinformatics*, vol. 7, p. 191, Jan. 2006.

[366] S. Maere, K. Heymans, and M. Kuiper, "BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks.," *Bioinformatics*, vol. 21, no. 16, pp. 3448–9, Aug. 2005.

[367] O. Garcia, C. Saveanu, M. Cline, M. Fromont-Racine, A. Jacquier, B. Schwikowski, and T. Aittokallio, "GOlorize: a Cytoscape plug-in for network visualization with Gene Ontology-based layout and coloring.," *Bioinformatics*, vol. 23, no. 3, pp. 394–6, Feb. 2007.

[368] Z.-Q. Liu, L. Yan, Z. Wu, C. Mei, K. Lu, Y.-T. Yu, S. Liang, X.-F. Zhang, X.-F. Wang, and D.-P. Zhang, "Cooperation of three WRKY-domain transcription factors WRKY18, WRKY40, and WRKY60 in repressing two ABA-responsive genes ABI4 and ABI5 in Arabidopsis.," *J. Exp. Bot.*, vol. 63, no. 18, pp. 6371–6392, 2012.

[369] H. Chen, Z. Lai, J. Shi, Y. Xiao, Z. Chen, and X. Xu, "Roles of arabidopsis WRKY18, WRKY40 and WRKY60 transcription factors in plant responses to abscisic acid and abiotic stress.," *BMC Plant Biol.*, vol. 10, no. 1, p. 281, Jan. 2010.

[370] F. Schweizer, N. Bodenhausen, S. Lassueur, F. G. Masclaux, and P. Reymond, "Differential Contribution of Transcription Factors to Arabidopsis thaliana Defense Against Spodoptera littoralis.," *Front. Plant Sci.*, vol. 4, p. 13, Jan. 2013.

[371] Q.-H. Shen, Y. Saijo, S. Mauch, C. Biskup, S. Bieri, B. Keller, H. Seki, B. Ulker, I. E. Somssich, and P. Schulze-Lefert, "Nuclear activity of MLA immune receptors links isolate-specific and basal disease-resistance responses.," *Science*, vol. 315, no. 5815, pp. 1098–103, Feb. 2007.

[372] Y. Brotman, U. Landau, Ã . Cuadros-Inostroza, T. Takayuki, A. R. Fernie, I. Chet, A. Viterbo, L. Willmitzer, Á. Cuadros-Inostroza, and T. Tohge, "Trichoderma-plant root colonization: escaping early plant defense responses and activation of the antioxidant machinery for saline stress tolerance.," *PLoS Pathog.*, vol. 9, no. 3, p. e1003221, Apr. 2013.

[373] J. Dong, C. Chen, and Z. Chen, "Expression profiles of the Arabidopsis WRKY gene superfamily during plant defense response.," *Plant Mol. Biol.*, vol. 51, no. 1, pp. 21–37, Jan. 2003.

[374] D. Yu, C. Chen, and Z. Chen, "Evidence for an important role of WRKY DNA binding proteins in the regulation of NPR1 gene expression.," *Plant Cell*, vol. 13, no. 7, pp. 1527–1540, Jul. 2001.

[375] X. Xu, C. Chen, B. Fan, and Z. Chen, "Physical and functional interactions between pathogen-induced Arabidopsis WRKY18, WRKY40, and WRKY60 transcription factors.," *Plant Cell*, vol. 18, no. 5, pp. 1310–1326, 2006.

[376] A. Takaya, T. Tomoyasu, H. Matsui, and T. Yamamoto, "The DnaK/DnaJ chaperone machinery of Salmonella enterica serovar Typhimurium is essential for invasion of epithelial cells and survival within macrophages, leading to systemic infection.," *Infect. Immun.*, vol. 72, no. 3, pp. 1364–1373, 2004.

[377] P. J. Pomposiello and B. Demple, "Identification of SoxS-regulated genes in Salmonella enterica serovar typhimurium.," *J. Bacteriol.*, vol. 182, no. 1, pp. 23–29, Jan. 2000.

[378] S. P. Salcedo and D. W. Holden, "Bacterial interactions with the eukaryotic secretory pathway.," *Curr. Opin. Microbiol.*, vol. 8, no. 1, pp. 92–8, Feb. 2005.

[379] K. T. Ly and J. E. Casanova, "Mechanisms of Salmonella entry into host cells.," *Cell. Microbiol.*, vol. 9, no. 9, pp. 2103–11, Sep. 2007.

[380] J. E. Galán, "Salmonella interactions with host cells: type III secretion at work.," *Annu. Rev. Cell Dev. Biol.*, vol. 17, pp. 53–86, Jan. 2001.

[381] A. E. Ramsden, D. W. Holden, and L. J. Mota, "Membrane dynamics and spatial distribution of Salmonella-containing vacuoles.," *Trends Microbiol.*, vol. 15, no. 11, pp. 516–24, Nov. 2007.

[382] J. R. Parrish, K. D. Gulyas, and R. L. Finley, "Yeast two-hybrid contributions to interactome mapping.," *Curr. Opin. Biotechnol.*, vol. 17, no. 4, pp. 387–93, Aug. 2006.

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

[384] P. Beltrao, C. Kiel, and L. Serrano, "Structures in systems biology.," *Curr. Opin. Struct. Biol.*, vol. 17, no. 3, pp. 378–84, Jun. 2007.

[385] H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker, "Network-based classification of breast cancer metastasis.," *Mol. Syst. Biol.*, vol. 3, p. 140, Jan. 2007.

[386] R. Sharan, I. Ulitsky, and R. Shamir, "Network-based prediction of protein function.," *Mol. Syst. Biol.*, vol. 3, p. 88, Jan. 2007.

[387] P. F. Jonsson and P. A. Bates, "Global topological features of cancer proteins in the human interactome.," *Bioinformatics*, vol. 22, no. 18, pp. 2291–7, Sep. 2006.

[388] M. D. Dyer, T. M. Murali, and B. W. Sobral, "Computational prediction of host-pathogen protein-protein interactions.," *Bioinformatics*, vol. 23, no. 13, pp. i159–66, Jul. 2007.

[389] F. P. Davis, D. T. Barkan, N. Eswar, J. H. McKerrow, and A. Sali, "Host pathogen protein interactions predicted by comparative modeling.," *Protein Sci.*, vol. 16, no. 12, pp. 2585–96, Dec. 2007.

[390] L. M. Stuart, J. Boulais, G. M. Charriere, E. J. Hennessy, S. Brunet, I. Jutras, G. Goyette, C. Rondeau, S. Letarte, H. Huang, P. Ye, F. Morales, C. Kocks, J. S. Bader, M. Desjardins, and R. A. B. Ezekowitz, "A systems biology analysis of the Drosophila phagosome.," *Nature*, vol. 445, no. 7123, pp. 95–101, Jan. 2007.

[391] R. Dreos, G. Ambrosini, R. C. Périer, and P. Bucher, "The Eukaryotic Promoter Database: expansion of EPDnew and new promoter analysis tools.," *Nucleic Acids Res.*, vol. 43, no. Database issue, pp. D92–6, Jan. 2015.

[392] J. C. Patel, K. Hueffer, T. T. Lam, and J. E. Galán, "Diversification of a Salmonella virulence protein function by ubiquitin-dependent differential localization.," *Cell*, vol. 137, no. 2, pp. 283–94, Apr. 2009.

[393] J. Versalovic, J. H. Jorgensen, G. Funke, D. W. Warnock, M. L. Landry, and K. C. Carroll, Eds., *Manual of Clinical Microbiology, 10th Edition*. American Society of Microbiology, 2011.

[394] C. Wu, C. M. Bell, and W. P. Wodchis, "Incidence and economic burden of adverse drug reactions among elderly patients in Ontario emergency departments: a retrospective study.," *Drug Saf.*, vol. 35, no. 9, pp. 769–81, Sep. 2012.

[395] S. Vilar, N. P. Tatonetti, and G. Hripcsak, "3D pharmacophoric similarity improves multi adverse drug event identification in pharmacovigilance.," *Sci. Rep.*, vol. 5, p. 8809, Jan. 2015.

[396] M. Kuhn, D. Szklarczyk, S. Pletscher-Frankild, T. H. Blicher, C. von Mering, L. J. Jensen, and P. Bork, "STITCH 4: integration of protein-chemical interactions with user data.," *Nucleic Acids Res.*, vol. 42, no. Database issue, pp. D401–7, Jan. 2014.

[397] M. Boolell, M. J. Allen, S. A. Ballard, S. Gepi-Attee, G. J. Muirhead, A. M. Naylor, I. H. Osterloh, and C. Gingell, "Sildenafil: an orally active type 5 cyclic GMP-specific phosphodiesterase inhibitor for the treatment of penile erectile dysfunction.," *Int. J. Impot. Res.*, vol. 8, no. 2, pp. 47–52, Jun. 1996.

[398] E. O. Major, "Progressive multifocal leukoencephalopathy in patients on immunomodulatory therapies.," *Annu. Rev. Med.*, vol. 61, pp. 35–47, Jan. 2010.

[399] R. A. Pache, A. Zanzoni, J. Naval, J. M. Mas, and P. Aloy, "Towards a molecular characterisation of pathological pathways.," *FEBS Lett.*, vol. 582, no. 8, pp. 1259–65, Apr. 2008.

[400] M. A. Yildirim, K.-I. Goh, M. E. Cusick, A.-L. Barabási, and M. Vidal, "Drug-target network.," *Nat. Biotechnol.*, vol. 25, no. 10, pp. 1119–26, Oct. 2007.

[401] S. I. Berger and R. Iyengar, "Network analyses in systems pharmacology.," *Bioinformatics*, vol. 25, no. 19, pp. 2466–72, Oct. 2009.

[402] J. Huang, C. Niu, C. D. Green, L. Yang, H. Mei, and J.-D. J. Han, "Systematic prediction of pharmacodynamic drug-drug interactions through protein-protein-interaction network.," *PLoS Comput. Biol.*, vol. 9, no. 3, p. e1002998, Jan. 2013.

[403] M. Kuhn, C. von Mering, M. Campillos, L. J. Jensen, and P. Bork, "STITCH: interaction networks of chemicals and proteins.," *Nucleic Acids Res.*, vol. 36, no. Database issue, pp. D684–8, Jan. 2008.

[404] J. D. Duke, X. Han, Z. Wang, A. Subhadarshini, S. D. Karnik, X. Li, S. D. Hall, Y. Jin, J. T. Callaghan, M. J. Overhage, D. A. Flockhart, R. M. Strother, S. K. Quinney, and L. Li, "Literature based drug interaction prediction with clinical assessment using electronic medical records: novel myopathy associated drug interactions.," *PLoS Comput. Biol.*, vol. 8, no. 8, p. e1002614, Jan. 2012.

[405] N. P. Tatonetti, P. P. Ye, R. Daneshjou, and R. B. Altman, "Data-driven prediction of drug effects and interactions.," *Sci. Transl. Med.*, vol. 4, no. 125, p. 125ra31, Mar. 2012.

[406] The UniProt Consortium, "UniProt: a hub for protein information," *Nucleic Acids Res.*, vol. 43, no. Database issue, pp. D204–12, Oct. 2014.

[407] Y. Liu, Q. Wei, G. Yu, W. Gai, Y. Li, and X. Chen, "DCDB 2.0: a major update of the drug combination database.," *Database (Oxford).*, vol. 2014, p. bau124, Jan. 2014.

[408] Z. Khatoon, B. Figler, H. Zhang, and F. Cheng, "Introduction to RNA-Seq and its applications to drug discovery and development.," *Drug Dev. Res.*, vol. 75, no. 5, pp. 324–30, Aug. 2014.

[409] D. Panne, "The enhanceosome.," *Curr. Opin. Struct. Biol.*, vol. 18, no. 2, pp. 236–42, Apr. 2008.

[410] A. D. Yu, Z. Wang, and K. V Morris, "Long noncoding RNAs: a potent source of regulation in immunity and disease.," *Immunol. Cell Biol.*, vol. 93, no. 3, pp. 277–83, Mar. 2015.

[411] Q. Zhang and K.-T. Jeang, "Long non-coding RNAs (lncRNAs) and viral infections.," *Biomed. Pharmacother.*, vol. 3, no. 1, pp. 34–42, Mar. 2013.

[412] K. A. Whitehead, J. E. Dahlman, R. S. Langer, and D. G. Anderson, "Silencing or Stimulation? siRNA Delivery and the Immune System," Jun. 2011.

[413] S. Jana, C. Chakraborty, S. Nandi, and J. K. Deb, "RNA interference: potential therapeutic targets.," *Appl. Microbiol. Biotechnol.*, vol. 65, no. 6, pp. 649–57, Nov. 2004.

[414] U. Schultz, B. Kaspers, and P. Staeheli, "The interferon system of non-mammalian vertebrates.," *Dev. Comp. Immunol.*, vol. 28, no. 5, pp. 499–508, May 2004.

[415] T. Blevins, R. Rajeswaran, P. V Shivaprasad, D. Beknazariants, A. Si-Ammour, H.-S. Park, F. Vazquez, D. Robertson, F. Meins, T. Hohn, and M. M. Pooggin, "Four plant Dicers mediate viral small RNA biogenesis and DNA virus induced silencing.," *Nucleic Acids Res.*, vol. 34, no. 21, pp. 6233–46, Jan. 2006.