Construction of gene coexpression networks and transcriptomic analysis for *Rhizobium leguminosarum*



Javier Pardo Díaz

Keble College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2022

This thesis is dedicated to
my parents, my brother, and my grandparents
for their unconditional love and support.

Acknowledgements

It seems like a lifetime since I first arrived in Oxford to start working on this piece of work. During this (very) long journey, I have been extremely fortunate to meet loads of amazing people who have made this experience an unforgettable one and to whom I am incredibly grateful.

Firstly and foremost, I would like to thank my supervisors for their guidance and advice throughout this project. Thank you to Gesine Reinert for her constant personal and professional support; I have no doubt she is the best supervisor a student can have. Thank you to Philip Poole, for sharing his enthusiasm about *Rhizobium* with me and for the opportunity to work in his lab. Thank you to Charlotte Deane for always seeing the big picture and providing the right advice. Thank you to Mariano Beguerisse for his input and ideas despite the distance.

I would also like to thank all the colleagues and friends that have selflessly contributed to this thesis. Thank you to Lyuba Bozhilova for teaching me everything about networks and procrastination mind games. Thank you to Raphael Ledermann for his supervision, knowledgeable advice and his not-that-good jokes. Thank you to Marcela Mendoza for bearing with me when I first came to the lab. Thank you to Carolin Schulte for her contributions to the mutant construction and her ability to motivate me. Thank you to Clare Cocker for her moral support during the RNA extraction. Thank you to Carmen Sánchez for always being there without being. Thank you to Carlos Outeiral for his motivational speeches and unlikely future career ideas. Thank you to Alison East, Beatriz Jorrín, Brandon Ford, Florian Klimm, Hayley Knights, Jack Parsons, James Wilsenach, Nick Crang, Oliver Crook, Tim Haskett, and Vinoy Ramachandran for their help and advice. Thank you to Helen Prescott, María Rendón, and Lida Chen for managing and taking care of the lab and the enjoyable conversations. Thank you to all the members of the OPIG and the Rhizosphere Lab for being so welcoming; it has been such a pleasure learning from them. Thank you to Carlos, Carmen, Florian, Lyuba, Raphael, and Thomas for revising an earlier version of this thesis.

Keble MCR has been a pillar during most of this journey. I will forever remember all the dinners, board games nights, croquet matches, punting trips, and BOPs; and all the incredible people I have met there. Thank you to Adam, Alba, Annie,

Bryony, Carolin, Fergus, Hanna, Klaudia, Michael, Miguel, Molly, Annie, Susy, Thomas, and many others for their friendship. Meeting them all is the best thing I get from this journey. Special thanks to James for his support whenever I needed it. I would also like to thank the Keble community; thanks to them, I have made the most out of the "Oxford experience".

Thank you to all those friends that have supported me and checked on me from outside Oxford. Thank you to my family in the UK for always being there and making me laugh and enjoy "La Isla"; certainly, these last years would not have been the same without them. Thank you to my SRUK friends, for all their work and motivation and for allowing me to explore other careers outside academia. Thank you to my colleagues at the Embassy of Spain in London, especially Javi, whose friendship has been a constant during my internship there.

Lastly, I want to thank my family. Thank you very very much to my parents, Pepe and Paloma, and my brother, Pepe. Thank you for their unconditional love, support, and encouragement. I would not be here without them, and I am extremely proud of having them in my life. Os quiero mucho.

When asked about my DPhil, I often say that "es como una montaña rusa de emociones" (it is like a roller-coaster of emotions). I am sure that I would not have enjoyed this ride without all the people that have been accompanying me. Thank you all, see you at the next adventure!

Abstract

Rhizobium leguminosarum is a soil bacteria which infects plants of the legume family such as peas, lentils, and beans. When associated with the plant, R. leguminosarum fixes atmospheric nitrogen (biological nitrogen fixation), producing ammonium that is assimilated by the plant, improving plant growth. This symbiotic relationship is very important for agriculture; increasing the knowledge of plant infection and nitrogen fixation processes might lead to a decrease in the use of chemical fertilisers.

In this thesis, I combine *R. leguminosarum* gene expression data from previous experiments to produce gene coexpression networks and increase the functional annotation of this bacterium. Gene coexpression networks are networks in which nodes represent genes and edges represent coexpression (i.e. two genes are connected if they have a similar expression pattern). Unfortunately, there is no standard method to generate gene coexpression networks from gene expression data. I introduce signed distance correlation as a measure of dependency between two variables and use it to generate self-consistent, unweighted and weighted, gene coexpression networks that are more stable and capture more biological information than those obtained from Pearson correlation or mutual information.

Subsequently, I present a pipeline that combines novel scores from gene coexpression network analysis in a principled way to identify the genes that are associated with certain growth conditions or highly coexpressed with a predefined set of genes of interest. This association can lead to putative functional annotation or to a prioritised list of genes for further study.

Lastly, I provide a detailed analysis of the transcriptome of *R. leguminosarum*. For this purpose, I generate different knock-out mutants lacking transcriptional regulators which are reported to be important in the nitrogen-fixing state of the bacteria and perform RNA-Seq experiments. I also carry out a preliminary study of the molecular signals that may affect the changes in gene expression between the free-living and the plant associated bacteria.

Contents

List of Abbreviations **xiii**

1 Introduction **1**

1.1 Motivation 1

1.2 Gene expression 3

1.2.1 Molecular basis of gene expression 3

1.2.2 Techniques to quantify gene expression 5

1.2.3 Gene expression data 7

1.2.4 Differential gene expression 8

1.3 Gene coexpression networks 9

1.3.1 Construction of gene coexpression networks 9

1.3.2 Extraction of information from gene coexpression networks . 11

1.4 *Rhizobium leguminosarum* and nitrogen fixation 14

1.4.1 *Rhizobium leguminosarum* 14

1.4.2 Bacteroid metabolism 15

1.4.3 Genome of *Rhizobium leguminosarum* 16

1.4.4 INSeq library for *Rhizobium leguminosarum* 17

1.4.5 Databases and other sources of information 18

1.5 Thesis outline 18

2 Construction of gene coexpression networks using distance correlation **21**

3 Construction of weighted and thresholded gene coexpression networks **35**

4 Analysis and evaluation of gene coexpression networks for *Rhizobium leguminosarum* **53**

5 Analysis of gene expression in the bacteroid 75
5.1 Overview . . . 75
5.2 Material and methods . . . 76
5.2.1 Bacterial strains . . . 76
5.2.2 Expression data and gene coexpression networks . . . 77
5.2.3 Identification of transcriptional regulators . . . 78
5.2.4 Experimental methods . . . 80
5.2.5 RNA-Seq analysis . . . 84
5.3 Results and discussion . . . 85
5.3.1 Identification of genes which are differentially expressed in the bacteroid . . . 85
5.3.2 Acetylene Reduction Assays and phenotypic analysis of mutants 87
5.3.3 Transcriptomic analysis of the mutants . . . 88
5.3.4 Discarded results . . . 110
5.4 Conclusions . . . 111

6 Analysis of promoter regions of genes which are highly expressed in free-living conditions 115
6.1 Introduction and objectives . . . 116
6.2 Methods . . . 117
6.2.1 Identification of groups of genes which are highly expressed in free-living conditions . . . 117
6.2.2 Identification of Transcription Starting Sites . . . 118
6.2.3 Evaluation of regions surrounding the Transcription Starting Sites . . . 118
6.3 Results and discussion . . . 119
6.3.1 Identification and analysis of groups of genes which highly expressed in free-living conditions . . . 119
6.3.2 Evaluation of regions surrounding the Transcription Starting Sites . . . 120
6.3.3 Future directions . . . 123

7 Conclusions and future directions 125
7.1 Construction of gene coexpression networks . . . 125
7.2 Extracting information from gene coexpression networks . . . 127
7.3 Transcriptomic analysis of *R. leguminosarum* . . . 129
7.4 Closing remarks . . . 131

Bibliography 133

Appendices

A Online application to access *R. leguminosarum* gene coexpression networks 145

B Supplementary Information: Robust gene coexpression networks using signed distance correlation 147

C Supplementary Information: COGENT: evaluating the consistency of gene co-expression networks 163

D Supplementary Information: Analysis and evaluation of gene co-expression networks for *Rhizobium leguminosarum* 179

E Supplementary Tables 201

List of Abbreviations

ATP . . . . . Adenosine triphosphate

cDNA . . . . . Complementary DNA

CoA . . . . . Coenzyme A

CO_2 . . . . . Carbon dioxyde

DNA . . . . . Deoxyribonucleic acid

DSD . . . . . Diffusion State Distances

EPS . . . . . Exopolysaccharides

FC . . . . . Fold Change

GCN . . . . . Gene coexpression network(s)

GRN . . . . . Gene regulatory network(s)

mRNA . . . . Messenger RNA

LB . . . . . Lysogeny Both growth media

log2FC . . . . Log2 Fold Change

N_2 . . . . . Nitrogen

NADPH . . . Nicotinamide adenine dinucleotide phosphate

NH_4^+ . . . . . Ammonia

O_2 . . . . . Oxygen

PCR . . . . . Polymerase Chain Reaction

PHB . . . . . Polyhydroxybutyrate

Rlv3841 . . . . *Rhizobium leguminosarum* bv. **viciae** 3841

RNA . . . . . Ribonucleic acid

RNA-Seq . . . RNA sequencing

RPKM . . . . Reads Per Kilobase Million

rRNA . . . . . Ribosomic RNA

scRNA-Seq . . Single-cell RNA sequencing

TCA . . . . . Tricarboxylic Acid

TOM . . . . . Topologic Overlap Matrix

TPM . . . . . Transcripts Per Million

TR . . . . . Transcriptional regulator

TSS . . . . . Transcription Starting Site

TY . . . . . Tryptone Yeast growth media

WGCNA . . . Weighted Gene Coexpression Network Analysis (R package)

WT . . . . . Wild Type

1

Introduction

1.1 Motivation

Nitrogen is an essential nutrient for all organisms, including plants, and the most abundant gas in the atmosphere, where it is present as a diatomic gas (N_2). Despite this, nitrogen is, together with water, the main limiting factor in crop production since N_2 needs to be transformed into ammonium NH_4^+ or nitrate NO_3^- before it can be readily assimilated by plants (Lupwayi *et al.*, 2006). The conversion of N_2 into ammonia NH_3 (*nitrogen fixation*) to use in fertilisers can be done chemically using the Haber-Bosch process, which requires high temperatures and pressures, and contributes an estimated 1.2% of all anthropogenic carbon dioxide (CO_2) emissions (Smith *et al.*, 2020). For this reason, the development of alternatives allowing the reduction of the use of nitrogen fertilizers but maintaining crop production is desirable.

In this thesis, we concentrate on a particular bacterium which may contribute to reducing the environmental cost of nitrogen provision for crops. *Rhizobium*

leguminosarum is a soil bacteria which infects plants of the legume family such as peas, lentils, and beans. When associated with the plant, *R. leguminosarum* fixes atmospheric nitrogen (*biological nitrogen fixation*), producing NH_4^+ that is assimilated by the plant, improving its growth (Sprent and James, 2007). Thus, this symbiotic relationship has an important role in food production. Further investigating the plant infection and nitrogen fixation processes might contribute to reducing the use of chemical fertilisers.

Professor Philip Poole's lab has been studying *R. leguminosarum* for over two decades (Lodwig and Poole, 2003; Sánchez-Cañizares *et al.*, 2017; Mendoza-Suárez *et al.*, 2020; Knights *et al.*, 2021; Schulte *et al.*, 2022; Poole and Ledermann, 2022). During this time, they have obtained a rich collection of gene expression datasets that measure the expression of the genes in bacteria across different conditions (Karunakaran *et al.*, 2009; Pini *et al.*, 2017; Wheatley *et al.*, 2017). However, this data has never been analysed jointly. In this thesis, we combine gene expression data from previous experiments to produce gene coexpression networks and increase our knowledge of this bacterium. In these networks, the genes are presented as nodes which are connected if they are coexpressed under the same conditions. There exist different approaches to constructing gene coexpression networks from gene expression data, and no universally accepted *right* choice exists (Weirauch, 2011). We examine how accurately different measurements (e.g. Pearson correlation coefficient) assess the similarity of gene expression patterns, and if and how to threshold those similarity values (i.e. select only the most similar pairs of genes) to construct gene coexpression networks. We also study to extract information from such networks. In this dissertation, we tackle these points by providing a construction method of gene coexpression networks and a detailed statistical and biological analysis of weighted and unweighted networks for *R. leguminosarum*. We also generate mutants for several transcriptional regulators and analyse their transcriptomes to better understand their role in nitrogen fixation.

We continue this chapter with an overview of gene expression, as well as gene coexpression networks, and *R. leguminosarum*. We then discuss how to

construct gene coexpression networks from gene expression data, and how to analyse them. We subsequently present a detailed analysis of the gene expression in the nitrogen-fixing state of *R. leguminosarum*. Finally, we give an overview of the remainder of this thesis.

This project is related to a partnership with the company Legume Technology, who is interested in the development of "nitrogen-fixing inoculants" for legume crops. We use gene coexpression networks to increase our knowledge about *R. leguminosarum* and the nitrogen fixation process. The results derived from this work may improve the inoculants they are currently using in order to make them more efficient.

1.2 Gene expression

1.2.1 Molecular basis of gene expression

The genetic information carried by each cell is stored in deoxyribonucleic acid (DNA). DNA is generally formed by two complementary chains of nucleic acids, which coil around each other to form a double helix (Watson and Crick, 1953). Each DNA strand is a polymer of nucleotides bound together with a 5' end, which frequently contains a phosphate group attached to the 5' carbon of nucleotide, and a 3' end, which typically is unmodified. A protein-coding gene is a sequence of nucleotides that contains the information to produce a particular protein. The expression of a gene is the amount of messenger ribonucleic acid (mRNA) that there is in the cell at a given moment; mRNA is the intermediate step between genes (DNA) and proteins, which are the molecules that carry out the functions in the cell. In order to understand gene (or equivalent protein) function, gene expression can be used as a proxy for protein concentration (more difficult to measure systematically) and a reasonable assumption is that when a protein is needed in the cell, the amount of mRNA that codes for it increase (Gygi *et al.*, 1999). Different environmental conditions may require different responses, and gene expression may vary depending on environmental conditions. For example, a cell

will only produce the proteins participating in the degradation of lactose when this metabolite is available. Varying gene expression can be done by transcriptional regulators. Transcriptional regulators (TR) are proteins that control which genes are expressed at a given time. These molecules bind to DNA and increase (activators) or decrease (repressors) the expression of genes the transcriptional regulators recognise.

The synthesis of mRNA from DNA is called transcription. Transcription is initiated when RNA polymerase binds to the promoter sequence at the 5' end of the template DNA. This promoter sequence, in bacteria, is marked by two motifs, the -35 region and the Pribnow box (also known as the -10 region); these motifs appear 35 and 10 nucleotides upstream of the transcription starting site (TSS), respectively. The sigma subunit of the RNA polymerase (sigma factor) recognises the promoter region and binds to the DNA; then, a short length of DNA is unwound to create a transcription bubble, allowing ribonucleotides to bind. Afterwards, the sigma factor dissociates from the polymerase complex and transcription proceeds until the RNA polymerase reaches a termination sequence.

In bacteria, unlike in eukaryotes, genes are organised in operons (Jacob and Monod, 1961). An operon is a functioning unit of DNA containing a cluster of genes under the control of the same promoter. Thus, the genes in an operon are transcribed into the same mRNA strand. This does not necessarily mean that the genes in the same operon will be perfectly coexpressed since the RNA polymerase may detach from the template DNA before transcribing all the genes under the same promoter.

After transcription, the mRNA is translated into proteins by the ribosome. Ribosomes are macromolecular complexes made up of proteins and ribosomal RNA (rRNA). During the translation process, amino acids are bound together to form proteins; each amino acid of a protein is encoded by a triplet of mRNA bases (a codon).

1.2.2 Techniques to quantify gene expression

1.2.2.1 Microarrays

Microarrays are used to measure the expression of a set of different genes at the same time (Butte, 2002). A microarray consists of a collection of tens of thousands of microscopic DNA spots — probes of known DNA sequences — on a solid surface (generally plastic, glass or silicone) to which DNA molecules complementary in sequence can attach (Figure 1.1). The use of microarrays to measure the amount of mRNA in a sample requires the previous synthesis in silico of complementary DNA (cDNA) from the mRNA in the sample by the use of reverse transcriptase in a process called retrotranscription. The concentration of the cDNA is estimated by the luminescence of the corresponding probes on the array. A high concentration of a determined mRNA in the sample (high expression) will lead to more molecules attached to the microarray and increased luminescence. Microarray experiments provide rich information about gene expression and have been carried out for many organisms in many different conditions for differential expression studies (e.g. Walker *et al.* (2016); Salem *et al.* (2017); Karunakaran *et al.* (2009)). However, they have some well-documented problems:

- The measures they provide exhibit high randomness (and noise) (Kothapalli *et al.*, 2002; Barnes *et al.*, 2005). Even two equivalent samples — replicates — can provide different and often uncorrelated results.
- Data from different microarrays are not directly comparable because they provide a relative measurement of gene expression levels (Venkatasubbarao, 2004). In addition, the final expression values depend on different factors such as the preparation and purification of the samples, the dye employed and the way the light emission is captured.
- The analysis is restricted to probes which are present in the microarrays.
- It is impossible to distinguish genes that are not expressed from those that are expressed at a low-level (Ideker *et al.*, 2000).

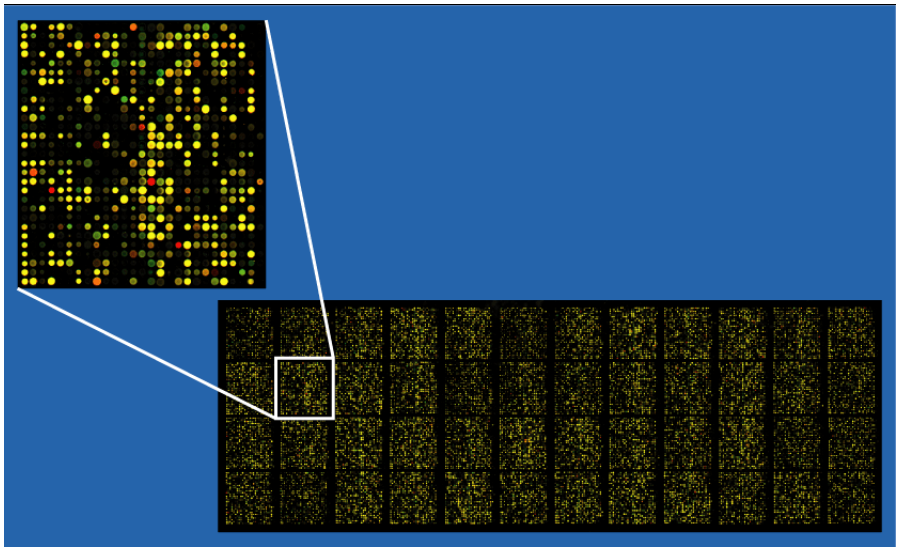


Figure 1.1: Example of an approximately 37,500 probe spotted oligo microarray with enlarged inset to show detail. Labelled molecules from the sample get attached to the probes in the microarray. Their concentration is estimated by measuring the luminescence of the array. In this example, the use of two different dyes (Cy5 — reddish — and Cy3 — greenish-yellow) allows differentiation between two different samples. Image originally produced by Paphrag (Wikipedia, 2022).

1.2.2.2 RNA Sequencing and single-cell RNA Sequencing

RNA Sequencing (RNA-Seq) is a sequencing technique that measures the presence and quantity of mRNA in a biological sample at a given moment (Wang *et al.*, 2009). This method requires, firstly, the isolation of the total RNA from the sample. Then, non-coding RNA such as for example rRNA, which highly abundant in the cell, are depleted. As most of the current DNA sequencing methods can only directly sequence relatively short DNA fragments (300–1000 nucleotides long) in a single reaction, it is necessary to digest the RNA molecules into smaller fragments using an RNase since average bacterial genes are around 1,500 nucleotides long. The next step consists of the synthesis of a double-stranded complementary DNA (cDNA) from the RNA (reverse transcription). Once the cDNA library is ready, it is sequenced using DNA sequence technologies. The sequencing results file contains the sequence of the DNA fragments in the prepared library (reads). These reads are mapped into the genome to obtain an estimate of the expression of each gene; the more reads mapping to a gene, the higher the expression of the gene. RNA-Seq results can be

expressed as raw counts (number of reads mapped to the gene), Reads Per Kilobase Million (RPKM, one million times the number of gene reads divided by the total number of reads and the length of the gene), or Transcripts Per Million (TPM, the quotient between one million times the number of reads divided by the length of the gene and the sum of reads divided by the length of the genes for all genes).

A variant of this technique is single-cell RNA-Seq, which measures the presence and quantity of mRNA in each cell of a biological sample separately. This approach allows the researcher to investigate which genes are expressed in each cell at a given time. These techniques overcome some of the problems derived from the use of microarrays. The most important differences are that RNA-Seq does not require a reference genome for the sequencing and does not limit the analysis to a set number of genes (as microarrays do). For well-characterized microorganisms, there is a good agreement between the RNA-Seq and microarray platforms (Nookaew *et al.*, 2012).

1.2.3 Gene expression data

Gene expression data is rich but, as already stated, also usually very noisy. For this reason, preprocessing of the data is required before its analysis. This preprocessing has three main objectives: minimising the expression value of the genes that were not expressed (noise), making the different arrays comparable, and conserving as much information as possible (Herrero *et al.*, 2003). To achieve them, the first step is selecting which genes to include in the analysis. One possibility is filtering out those genes that seem to have no significant change across the microarrays, known as *flat genes* (Prieto *et al.*, 2006). In Prieto *et al.* (2008) for example, genes with a mean expression value lower than the median expression value and variance across the different samples lower than the median variance are deleted from the analysis. Another possibility is employing only those genes in the analysis which are highly expressed in a particular condition of interest, as in Stertz *et al.* (2018).

The second step consists of normalising the data from the different microarrays to make them comparable (Venkatasubbarao, 2004). It is possible to normalise the values within each microarray between zero (lowest expression) and one (highest

expression). However, it is often more convenient to use quantile normalisation (Amaratunga and Cabrera, 2001) since it offers more robust and accurate estimations of the expression values and the expression correlations (Bolstad *et al.*, 2003; Lim *et al.*, 2007). The quantile normalisation technique assigns to the n-th most expressed gene in each microarray the mean expression value of all the n-th most expressed genes across all the microarrays in the set. This method makes the distribution of values in each microarray identical in statistical properties such as quantile distribution.

Gene expression data is commonly used to perform differential expression experiments in which the levels of expression of the genes are compared across two conditions. Other uses of gene expression data are the construction of gene regulatory networks and gene coexpression networks. Gene regulatory networks are directed networks which aim to identify regulatory pathways (i.e. which genes are responsible for the activation/repression of the expression of other genes). In contrast, gene coexpression networks are undirected networks in which genes are connected if they are expressed at the same time.

1.2.4 Differential gene expression

Differential gene expression studies compare the expression of the organism's genes across two different conditions. This setup allows the identification of those genes that change their expression as a response to a change in the environment. They allow the identification of genes that are *overexpressed* or *upregulated* (their expression is higher in the sample than in the control used to compare) and those that are *downregulated* (their expression is lower in the sample than in the control used to compare). The unit used to assess differential expression is Fold Change (FC), which is the quotient of the expression of the gene in both conditions (i.e. the FC value of a gene is the expression of the gene in the sample divided by the expression of the gene in the control). FC values can be expressed as logarithms; for instance, in this thesis, we use *log2FC* to denote the log base 2 of the FC values.

In microarrays, the FC values correspond to the ratio of the intensity of the two dyes employed (Butte, 2002). In RNA-Seq experiments, the R package DESeq2 (Love *et al.*, 2014) estimates variance-mean dependence in count data (raw counts) and tests for differential expression based on a model using the negative binomial distribution. DESeq2 provides a log2FC value and an adjusted p-value for each gene in the dataset. DESeq2 is reported to work well with a small number of replicates. We can also obtain FC values by dividing the TPM values for the two conditions of interest as TPM values are generally comparable across different samples.

1.3 Gene coexpression networks

Gene coexpression networks are networks where nodes are genes and edges indicate covariability in their expression across multiple samples (Lee *et al.*, 2004). Representing gene coexpression as networks eases the study and visualisation of the expression data (Weirauch, 2011; Magwene and Kim, 2004) and helps exploit the structure of interactions between genes at a whole-system level. One motivation behind creating these networks is that genes which are highly coexpressed across multiple samples are likely to have related functions (Hughes *et al.*, 2000; Stuart *et al.*, 2003; van Noort *et al.*, 2003; Makrodimitris *et al.*, 2020), allowing inference of gene function using *guilt by association* approaches (Wolfe *et al.*, 2005). This procedure is especially useful if the studied organism is poorly annotated. However, the lack of reliable genomic information may hinder the validation of the accuracy of the network models generated from gene expression data. For this reason, it is important to develop ways to construct networks that can extract valuable information from the data, yet are robust to noise in the data and changes in the gene expression dataset.

1.3.1 Construction of gene coexpression networks

In the literature, there are many different methods proposed to construct gene coexpression networks. However, all of them require preprocessing of the data to overcome some of the problems outlined in Section 1.2, as well as as a measure

that evaluates how similar the expression patterns of two genes across the different samples are. This measure is typically based on a correlation coefficient, but other measurements such as those based on mutual information have also been used (Butte and Kohane, 1999). The two most commonly used correlation measurements are the absolute value of the Pearson correlation coefficient, which assesses linear dependencies between the variables, and of the Spearman rank correlation, which measures how well the relationship between two variables can be described using a monotonic function (Weirauch, 2011). After obtaining the expression similarity values for each pair of genes, there are two alternatives: considering the coexpression as a continuous value and constructing weighted networks; or thresholding the coexpression to obtain a Boolean value and constructing an unweighted network. The latter approach (e.g. George *et al.* (2019)) keeps the strongest relationships; however, it is not obvious which threshold value strikes the right balance between sparsity and richness of information. A threshold that is too low results in overly dense networks that are noisy and difficult to analyse; a threshold that is too high risks discarding valuable information.

The Weighted Gene Coexpression Network Analysis (WGCNA) (Zhang and Horvath, 2005; Langfelder and Horvath, 2008, 2014; Ghazalpour *et al.*, 2006; Fuller *et al.*, 2011) is a method widely used to generate weighted gene coexpression networks (e.g. Guo *et al.* (2017); Presson *et al.* (2008); Min and Kramer (2020)). It assigns values (weights) from 0 to 1 to each pair of genes. These values are the absolute Pearson correlation of their expression, raised to a power. The power is chosen so that the resulting degree distribution is scale-free, implying that most genes will have a small number of edges, with only a few hub genes having many edges (Carter *et al.*, 2004). This operation emphasizes high correlations at the expense of low correlations, potentially leading to a loss of information.

1.3.2 Extraction of information from gene coexpression networks

Gene coexpression networks are a powerful tool for studying transcriptomic data since genes that are expressed at the same time are likely to be related functionally. Different approaches have been described to extract information from gene networks in general and gene coexpression networks in particular (Thalamuthu *et al.*, 2006). Below we describe three commonly used methods to identify highly-connected groups of nodes, with a specific application to gene coexpression networks, which may later be associated with a particular biological function. This can be achieved after assigning descriptors to the groups of nodes using additional node information; in the case of gene coexpression networks, this additional information may come from experimental results or the literature. For example, one might identify a community with a high proportion of genes (*enriched*) highly expressed in particular experimental conditions or known to participate in a defined metabolic pathway.

1.3.2.1 Community detection

Communities are usually defined as groups of nodes that are connected "more densely" to each other than to the rest of the network. Detecting communities allows non-trivial internal network organisation to be unveiled (Yang *et al.*, 2016). Furthermore, once the communities have been obtained, it is possible to coarse-grain the network by merging all the nodes in each community (and their edges) to create meta-nodes and then analyse the relationships between the communities through the resulting meta-node network.

Some community detection methods are based on the maximization of *modularity*, which is the difference between the number of edges in communities and the expected number under a random model whih preserves only the degrees of each node (Newman and Girvan, 2004). More formally, for a given partition P of the sets of nodes, the modularity $Q(P)$ is given by $Q(P) = \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(\sigma_i, \sigma_j)$, where A is the adjacency matrix whose elements indicate whether pairs of nodes are

adjacent or not in the graph, k_i is the degree of node i, m is the number of edges, and $\delta(\sigma_i, \sigma_j) = 1$ if i and j are in the same set in the partition P and 0 otherwise.

The model based in Reichardt and Bornholdt (2006) adds a resolution parameter (γ) to the quality function: $Q(P) = \sum_{ij} \left(A_{ij} - \gamma \frac{k_i k_j}{2m} \right) \delta(\sigma_i, \sigma_j)$. High values of the resolution parameter γ result in small communities; conversely, lower γ values lead to larger (and fewer) communities. Therefore, different network partitions arise by selecting different resolution values. As an example, the Louvain algorithm (Blondel *et al.*, 2008) calculates the quality of the network partition by considering its modularity and optimising it. It starts by optimising the modularity locally on all the nodes and then combines small communities if this combination increases the overall modularity. However, the Louvain algorithm may yield arbitrarily badly connected communities, and even internally disconnected communities (Traag *et al.*, 2019). The Leiden algorithm (Traag *et al.*, 2019) solves these problems by introducing an improved methodology to combine the individual nodes to generate the communities. It is worth mentioning that optimising modularity is computationally prohibitive for large networks and that these algorithms are approximate.

1.3.2.2 Module detection using WGCNA

The main application of the WGCNA R package (Langfelder and Horvath, 2008) is the detection of subsets of genes (modules) that are tightly connected to each other. To do this, WGCNA calculates a *topological overlap matrix* (TOM), which provides a similarity measure for each pair of genes. This matrix indicates how similar the patterns of connection strengths of each pair of genes are. A pair of genes is said to have a high topological overlap if they are both strongly connected to the same group of genes and to each other. WGCNA uses hierarchical clustering to group the genes in different modules according to the TOM.

1.3.2.3 PageRank and personalised PageRank

A typical objective when studying a network is to identify which nodes are the most *important* ones (those with the highest centrality). The PageRank algorithm,

originally developed to rank web pages (Brin and Page, 1998), tackles this problem. This algorithm produces a probability distribution over the N nodes of a directed network G that represents the fraction of time that a random walker spends in each node in an infinitely long random walk. At each step, the random walker can, with probability α, choose to follow an outgoing edge from the current node (chosen with probability proportional to its weight), or, with probability $1-\alpha$, "teleport" to any node chosen from a distribution v (with the notation that v_i is the probability that node i receives a teleported walker). This algorithm can also be applied to undirected networks. Let $x(t) \geq 0$ be the probability that each node receives a random walker at time t (i.e., $\sum_{i=1}^{N} x_i(t) = 1$), for all $t > 0$, which evolves according to

$$x(t) = \alpha \left(D^{-1}\right)^{\intercal} x(t-1) + (1-\alpha)v. \tag{1.1}$$

The matrix D is the diagonal matrix of (possibly weighted) degrees of the nodes, so that D_{ii} is the degree of node i and $D_{ij} = 0$ if $i \neq j$. Note that all the nodes need to have at least one edge. The PageRank vector π is the solution of equation (1.1) when $t \to \infty$.

Typically the teleportation distribution v is chosen to be uniform (i.e., $v_i = \frac{1}{N}$ for all i); however, we can manipulate v to bias teleported walkers to a specific node or group of nodes. For example if $v_i = 0.9$ and $v_j = \frac{0.1}{N-1}$ for $j \neq i$, teleported walkers will go to node i with probability 0.9. This has the effect that nodes near i will have a higher Pagerank than with uniform teleportation, and is known as Personalised Pagerank (Gleich, 2015). In the context of gene coexpression networks, Personalised Pagerank may be useful to identify which genes are highly coexpressed with a pre-defined set of genes of interest.

1.4 *Rhizobium leguminosarum* and nitrogen fixation

In this dissertation, we generate and analyse gene coexpression networks using gene expression data from *R. leguminosarum*. In this section, we present some key information about the bacteria.

1.4.1 *Rhizobium leguminosarum*

R. leguminosarum is an α-proteobacterium that fixes atmospheric nitrogen when associated with legumes (eg. peas, beans, lentils). *R. leguminosarum* reduces molecular nitrogen into ammonia which can be assimilated by plants. Nitrogen fixation improves the growth of plants as nitrogen is one of the limiting factors during the growth process (Gutiérrez, 2012). Bacteria such as *R. leguminosarum* grow in the soil in the absence of their legume host (Downie, 2010) (*free-living conditions*). Legumes secrete compounds such as phenolic flavonoids and isoflavonoids (exudates) through their roots (Recourt *et al.*, 1989) which the bacteria living close by (in the *rhizosphere*) can detect. In response to this stimulus, bacteria approach the root hairs of the plant and attach — a process called chemotaxis — and secrete lipochitooligosaccharides known as Nod factors (Denarie *et al.*, 1996). The plants sense the Nod factors, which induce root hairs to curl and trap the bacteria (Oldroyd, 2013). The bacteria continue to grow and in due course, the plant forms an infection thread in which the bacteria grow. The bacteria from the infection thread are eventually released into cortical cells in the developing nodule. The bacteria in the nodule are surrounded by a plant membrane in a structure called the symbiosome, within which the bacteria differentiate to a state called *bacteroid*. In this stage, bacteroids fix nitrogen (Downie, 2010). Fig. 1.2 illustrates this root colonization process.

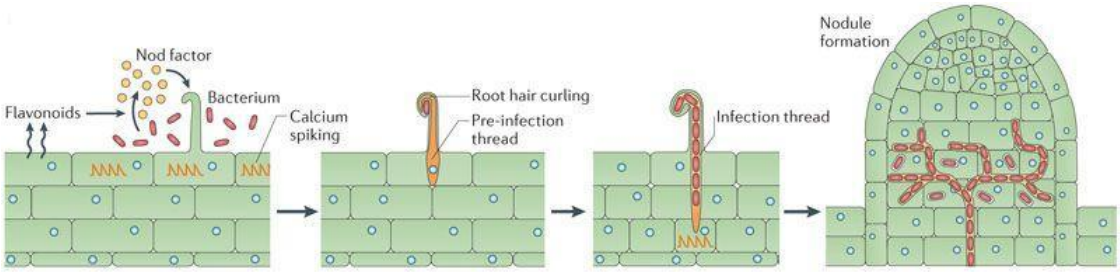


Figure 1.2: Legume root colonization by *R. leguminosarum*. (Far left) Bacteria in the rhizosphere detect flavonoids released by the plant root. Bacteria, in turn, produce nodulation factors (Nod factors) that are recognized by the plant. (Left) The root hair — where the bacteria are already attached — starts to bend and the bacteria gain entry into the plant. (Right) The formation of an infection thread allows the invasion of the bacteria into the root tissue. (Far right) The initiation of the formation of the nodule takes place, where the bacteria can differentiate into a nitrogen-fixing state. Adapted from Oldroyd (2013).

1.4.2 Bacteroid metabolism

R. leguminosarum experiences large changes in its metabolism during the infection process, from the *free-living* bacteria to the *plant-associated* bacteroid (Wheatley *et al.*, 2017), all of which respond directly or indirectly to biological fixation of nitrogen. These changes are also reflected in gene expression (Karunakaran *et al.*, 2009) and metabolite (Terpolilli *et al.*, 2016) levels.

Biological fixation of nitrogen is carried out by the nitrogenase enzyme in a reaction requiring a considerable amount of energy and reductive power (16 moles of ATP, eight H^+, and eight electrons are required to reduce one mole of N_2 to NH_3). *R. leguminosarum* oxidises dicarboxylates provided by the host plant to yield ATP and electrons required by the nitrogenase. The main dicarboxylates provided by the plant are malate and succinate. These metabolites need to be transformed into oxaloacetate and acetyl-CoA, which can be condensed to form citrate in the first step of the tricarboxylic acid (TCA) cycle.

In bacteroids, the TCA cycle is the main pathway for dicarboxylate metabolism, resulting in large amounts of reduced electron carriers (NADH and FADH2) required for ATP synthesis and nitrogen fixation. Some of the enzymes in the TCA cycle are upregulated in the transcriptome datasets across different rhizobial species. According to Karunakaran *et al.* (2009), there is an upregulation of citrate synthase,

aconitase, and 2-ketoglutarate dehydrogenase in *R. leguminosarum* bacteroids. However, additional mutant studies with careful evaluation of both enzyme activities and redirection of metabolic fluxes are required to elucidate the essentiality of flux through different parts of the TCA cycle.

Inside the nodules, the concentration of free oxygen is very low, which favours nitrogenase activity since this enzyme is highly susceptible to molecular oxygen (Imlay, 2006). *R. leguminosarum* has two copies of a cytochrome c oxidase, which is a high-affinity Cbb3-type oxidase that allows respiration under microaerobic conditions. The two copies are encoded by the gene clusters $fixNOQP1$ and $fixNOQP2$, and are highly upregulated in the bacteroid compared to the free-living bacteria (Karunakaran *et al.*, 2009).

Due to the reduced amount of available oxygen in the nodule, high TCA cycle activity may become disadvantageous because oxidation of electron carriers in the respiratory chain is restricted, and accumulation of reducing equivalents (e.g. NADH and NADPH) inhibits pyruvate dehydrogenase and a-ketoglutarate dehydrogenase. Bacteroids therefore must balance the production of NAD(P)H from the oxidation of acetyl-CoA in the TCA cycle with its storage into polymers, such as poly-hydroxybutyrate (PHB), glycogen, and lipids, consuming NADPH (Terpolilli *et al.*, 2016). PHB is synthesised by the condensation of two acetyl-CoA molecules to form acetoacetyl-CoA which is subsequently reduced to form β-hydroxybutyryl-CoA. These residues are the monomers which are polymerized by a PHB synthetase to form the polymer PHB (Mandon *et al.*, 1998). *R. leguminosarum* has a type III PHB synthase (encoded by *phaC2* and *phaE*) which is upregulated in the bacteroid (Karunakaran *et al.*, 2009), unlike the genes *phaA* and *phaB*, coding for the enzymes involved in the previous steps of the pathway.

1.4.3 Genome of *Rhizobium leguminosarum*

The strain we use in this thesis is *R.leguminosarum* bv.*viciae* 3841 (Rlv3841). Its genome is composed of the chromosome (RL) and six plasmids (pRL7, pRL8, pRL9, pRL10, pRL11 and pRL12) (Young *et al.*, 2006), having the plasmids a

smaller size than the chromosome. It has 7,288 predicted genes, which are named using an alphanumerical code that indicates the replicon where they are located and their relative position (e.g. pRL100196 is in pRL10 and RL1254 is in the chromosome; RL1254 is adjacent to RL1255). In 2021, the NCBI modified the notation of the genes of this organism to a different alphanumeric code that does not include information on the replicon (e.g. RL_RS25610). In this dissertation, we use the terms *old* and *new* notations to indicate the different identification names that a gene can have. Around 25% of the genes (1,824) have an assigned name (a gene ID; e.g. $nifA$ or $rhiB$). Gene IDs have been assigned to those genes that have been studied or that have a high similarity to annotated genes in other organisms. We have little or no functional information for most of the genes in Rlv3841 genome. A total of 1,795 genes are annotated as *hypothetical conserved protein*, *hypothetical protein*,*conserved hypothetical exported protein* or *hypothetical exported protein*). In addition, 138 genes are described as pseudogenes, since they are not thought not to code for proteins. This lack of reliable genomic information makes it difficult to evaluate the results of gene function prediction approaches designed to understand the bacteria.

1.4.4 INSeq library for *Rhizobium leguminosarum*

In Wheatley *et al.* (2020), a study of the essentiality of the genes in Rlv3841 is performed. According to this study, a gene is said to be essential if its removal results in an important decrease in fitness. For this purpose, they use mariner-based transposon insertion sequencing (INSeq) across different conditions. The conditions tested were: free-living conditions (input), rhizosphere, roots, nodule bacteria, and bacteroid. Throughout this thesis, we refer several times to this data.

The approach followed to construct this database involves the generation of a collection of mutants. Each of these mutants has a single copy of the transposon inserted in a random region of the genome. The growth of the mutants results in a competition experiment: those bacteria in which the insertion has knocked down an essential gene will grow less, or not grow at all. If the observed number of

insertions in a gene is higher than expected, the loss of that gene is interpreted as resulting in an increase in fitness (Chao *et al.*, 2016; Cain *et al.*, 2020).

1.4.5 Databases and other sources of information

In this thesis, we use several databases to study Rlv3841. We use a collection of microarrays from Professor Poole's lab that has been used in previous studies (Karunakaran *et al.* (2009); Pini *et al.* (2017); Wheatley *et al.* (2017)). This microarray database is the starting point of this project, as well as a tool to analyse the results we obtain.

We also use the information available on online databases such as STRING (Szklarczyk *et al.*, 2019), which information about the physical interactions and functional associations between proteins, and Biocyc (Karp *et al.*, 2019), which includes metabolic information.

1.5 Thesis outline

This dissertation has two main objectives: The first objective is the development of a methodology to construct gene coexpression networks using *R. leguminosarum* gene expression microarray data and extracting insightful biological information from them. The second objective is the use of these networks, alongside other biological information, to increase our knowledge of the bacteria and nitrogen fixation.

In Chapter 2 we explore different approaches to construct unweighted gene coexpression networks, including our novel correlation measurement, namely *signed distance correlation*. We find that this method, out of those tested, results in the most self-consistent unweighted networks, which also capture the highest amount of biological information. The methodology described requires the selection of a threshold value, which we select using the R package COGENT. This chapter is presented as two published papers:

- *Pardo-Diaz, J., Bozhilova, L. V., Beguerisse-Díaz, M., Deane, C. M., & Reinert, G. (2021) Robust gene coexpression networks using signed distance correlation. Bioinformatics, 37(14), 1982-1989.*

- *Bozhilova, L. V., Pardo-Diaz, J., Reinert, G., & Deane, C. M (2020) COGENT: evaluating the consistency of gene co-expression networks. Bioinformatics, 37(13), 1928-1929.*

In Chapter 3, we make use of signed distance correlation to generate weighted and thresholded gene coexpression networks. We observe that again, the use of signed distance correlation results in networks which are more robust and capture more biological information than those obtained using other correlations. We compare the weighted and thresholded networks with those unweighted networks from Chapter 2 to conclude that the former contains more biological information than the latter ones. This chapter is presented as a published paper:

- *Pardo-Diaz, J., Poole, P., Beguerisse-Díaz, M., Deane, C., & Reinert, G. (2022). Generating weighted and thresholded gene coexpression networks using signed distance correlation. Network Science, 10(2), 131-145.*

In Chapter 4, we analyse *R. leguminosarum* gene coexpression obtained following the pipeline described in Chapter 3. For this purpose, we use some of the network analysis approaches described in this introduction, such as community detection and PageRank, as well as new ones. We first evaluate the methods to extract information from the networks and we subsequently use them in two case examples to assign a putative molecular function to part of those genes in *R. leguminosarum* genome without functional annotations. This chapter is presented as a paper published in a Journal of Computational Biology special issue for the 80th birthday of Michael Waterman:

- *Pardo-Diaz J, Beguerisse-Díaz M, Poole PS, Deane CM, Reinert G. Extracting Information from Gene Coexpression Networks of Rhizobium leguminosarum. Journal of Computational Biology, 29(7):752-768.*

In Chapter 5, we generate a set of *R. leguminosarum* mutants to explore the role of transcriptional regulators in the processes taking place in the bacteroid. We inoculate pea plants with the mutants, collect the nodules from the roots of the plants and extract and sequence the mRNA in the bacteroids. The data derived from these experiments, in combination with the information obtained from analysing the gene coexpression using the methods described in Chapter 4, help us to increase our knowledge from *R. leguminosarum*. Chapter 6 is a preliminary study of the role of the promoter regions in the differential expression of the genes in the bacteroid, compared to the free-living conditions.

Lastly, in Chapter 7, we conclude by discussing possible directions for future work. We also include in this thesis two appendices that complement the work presented in the main chapters. The possible avenues for future work are also detailed in this chapter. Appendix A presents a tool developed for the study of *R. leguminosarum* gene coexpression networks, that will be used by Professor Poole's lab members in future research.

All results chapters in this thesis are written in publication style, using present tense and "we" although in the context of this dissertation it should be read as "I". For example, "to our knowledge", it should be read as "to my knowledge". When Chapter 5 is integrated into submission to a biological sciences journal, the chosen tense will likely be past tense.

The published code and data used in this thesis are openly available on GitHub: https://github.com/javier-pardodiaz/.

2

Construction of gene coexpression networks using distance correlation

The Supplementary Information for the publications integrated into this chapter is available in Appendices B and C.

Bioinformatics, 37(14), 2021, 1982–1989
doi: 10.1093/bioinformatics/btab041
Advance Access Publication Date: 1 February 2021
Original Paper

OXFORD

Gene expression

Robust gene coexpression networks using signed distance correlation

Javier Pardo-Diaz [1,2,*], Lyuba V. Bozhilova [1], Mariano Beguerisse-Díaz[3], Philip S. Poole[2], Charlotte M. Deane [1] and Gesine Reinert[1,*]

[1]Department of Statistics, University of Oxford, Oxford OX1 3LB, UK, [2]Department of Plant Sciences, University of Oxford, Oxford OX1 3RB, UK and [3]Mathematical Institute, University of Oxford, Oxford OX2 6GG, UK

*To whom correspondence should be addressed.

Associate Editor: Lenore Cowen

Received on June 24, 2020; revised on November 30, 2020; editorial decision on January 18, 2021; accepted on January 21, 2021

Abstract

Motivation: Even within well-studied organisms, many genes lack useful functional annotations. One way to generate such functional information is to infer biological relationships between genes/proteins, using a network of gene coexpression data that includes functional annotations. However, the lack of trustworthy functional annotations can impede the validation of such networks. Hence, there is a need for a principled method to construct gene coexpression networks that capture biological information and are structurally stable even in the absence of functional information.

Results: We introduce the concept of signed distance correlation as a measure of dependency between two variables, and apply it to generate gene coexpression networks. Distance correlation offers a more intuitive approach to network construction than commonly used methods, such as Pearson correlation and mutual information. We propose a framework to generate self-consistent networks using signed distance correlation purely from gene expression data, with no additional information. We analyse data from three different organisms to illustrate how networks generated with our method are more stable and capture more biological information compared to networks obtained from Pearson correlation or mutual information.

Contact: jdiaz@stats.ox.ac.uk or reinert@stats.ox.ac.uk

Availability and implementation: Code is available online (https://github.com/javier-pardodiaz/sdcorGCN).

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Gene expression data, while noisy, contains key information about biological processes (Kothapalli *et al.*, 2002). Such data are often represented as gene coexpression networks, where nodes are genes and edges represent correlations in their expression across multiple samples (Lee *et al.*, 2004). Representing gene coexpression as networks eases the study and visualization of the expression data (Magwene and Kim, 2004; Weirauch, 2011). One motivation behind creating these networks is that genes which are coexpressed across multiple samples are likely to have related functions (Hughes *et al.*, 2000; Makrodimitris *et al.*, 2020; Stuart *et al.*, 2003; van Noort *et al.*, 2003), allowing inference of gene function using *guilt by association* approaches (Wolfe *et al.*, 2005). This procedure is especially useful if the studied organism is poorly annotated. For example, *Rhizobium leguminosarum*, a soil bacteria important in agriculture that can infect plants of the legume family and provide them organic nitrogenous compound, has no functional information available for ~25% of its predicted genes. Most methods to generate and validate gene coexpression networks use exogenous biological information (such as gene ontologies and metabolic information) to select which edges need to or do not have to be present (e.g. Bar-Joseph *et al.*, 2003; Ihmels *et al.*, 2002; Ucar *et al.*, 2007). Therefore, the lack of reliable genomic functional information may hinder the construction of gene coexpression networks and the validation of their accuracy.

The most widely used methods to generate gene coexpression networks in the absence of exogenous information are based on the absolute value of the Pearson correlation of the expression of gene pairs across samples (Weirauch, 2011). After computing the correlation, there are two alternatives: (i) construct a fully connected weighted network, or (ii) impose a threshold to construct unweighted networks with edges connecting genes whose expression correlation is high enough. The former approach is widely used thanks to the R package WGCNA (Langfelder and Horvath, 2008), but results in noisy networks where gene relationships may not be

©The Author(s) 2021. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/bioinformatics/article/37/14/1982/6125359 by Bodleian Libraries of the University of Oxford user on 28 July 2022

easy to identify. The latter approach (e.g. George *et al.*, 2019) keeps the strongest relationships; however, it is not obvious which threshold value strikes the right balance. A threshold too low, results in overly dense networks that are difficult to analyse; a threshold that is too high risks discarding valuable information.

A natural way of studying gene expression data is to compare the expression of a gene in different samples to assess how it changes. For any two genes, the most intuitive approach should follow the same straightforward procedure: evaluate how the expression of each gene changes across the different samples and then compare the patterns of changes between the two genes. This same idea underpins the concept of *distance correlation* (Székely *et al.*, 2007). While Pearson correlation measures linear relationships, distance correlation measures the dependence, both linear and nonlinear, between two vectors, and provides a non-negative score that is zero if and only if the vectors are statistically independent (Székely *et al.*, 2007). Thus, distance correlations allow us to identify relationships between the expression of genes beyond linearity; this type of correlation has been successfully used in bioinformatics settings to predict miRNA-disease associations (Zhao *et al.*, 2018) and to generate gene regulatory networks from expression data (Ghanbari *et al.*, 2019; Guo *et al.*, 2014). Gene regulatory networks are different from gene coexpression networks because they are directed, much sparser and aim to identify regulatory pathways rather than functional associations. We give a detailed definition of distance correlation in Section 2 below.

Distance correlation is always non-negative; that is, it does not capture whether an association between the expression of two genes is positive or negative. Naturally, this information may be biologically relevant; to overcome this shortcoming, we introduce a *signed* distance correlation. After calculating the distance correlation between each pair of genes, we impose a sign, which corresponds to the sign of the Pearson correlation between the expression of the genes.

Here, we propose a method to construct gene coexpression networks using signed distance correlation as an intuitive alternative to networks from Pearson correlation, Spearman correlations and mutual information. To highlight the strengths of our approach, we construct and compare networks using the four methods. This work is to our knowledge the first work using distance correlation to construct gene coexpression networks.

We construct networks by including only edges between genes for which the signed distance correlation of their expression exceeds a threshold. We select the threshold based on the internal consistency of the networks using the R package COGENT (Bozhilova *et al.*, 2020) instead of using exogenous biological information known (or imposed) *a priori*. We evaluate our method in data from three different organisms. First, we generate an unweighted gene coexpression network for the bacteria *R.leguminosarum* from microarray data. We then analyse RNA-Seq data from the yeast *Saccharomyces cerevisiae*, and single-cell RNA-Seq data from human liver cells. The results of our analysis of yeast and human data can be found in the Supplementary Material.

We evaluate the biological information in our networks using the STRING database (Szklarczyk *et al.*, 2019), which is a protein–protein interaction database with scores for pairs of proteins. The higher the STRING score, the more likely they are to have a biologically meaningful functional interaction. Using STRING, we show that networks from signed distance correlation capture more biological information and are structurally more stable than networks based on Pearson or Spearman correlation or mutual information.

While we apply our method to gene expression data, our method to construct networks from signed distance correlations (in combination with COGENT) can be used in applications beyond bioinformatics.

Data and source code are available from https://github.com/javier-pardodiaz/sdcorGCN.

2 Materials and methods

The method we propose generates an unweighted coexpression network from gene expression data that may come from different sources, such as microarrays, RNA-Seq and single-cell RNA-Seq assays. Figure 1 illustrates the main steps of the method, which includes data pre-processing, computing correlations, and thresholding to create networks.

The input to the method is a $m \times n$ gene expression matrix M, where each of the m rows correspond to a gene, each of the n columns is a different sample, and the entries are the expression values of each gene in each sample.

2.1 Data

We analyse gene expression data from three different organisms: *R.leguminosarum*, Yeast (*S.cerevisiae*), and Human (*Homo sapiens*), obtained using different experimental techniques (microarrays, RNA-Seq and single-cell RNA-Seq). Below, we present our results on the *R.leguminosarum* dataset. The description and analysis of yeast and human datasets can be found in Supplementary Sections 3 and 4.

The *R.leguminosarum* bv. *viciae* 3841 data contains gene expression information observed under 18 different growth conditions. These data come from $n = 54$ microarray channels with 3 independent samples per condition (Karunakaran *et al.*, 2009; Pini *et al.*, 2017; Ramachandran *et al.*, 2011). The complete list of the conditions is in Supplementary Section 1. From the total 7263 genes in the current genome annotation (Young *et al.*, 2006), we remove genes that do not appear in all the microarrays or appear as pseudogenes, leaving $m = 7,077$ genes for which we calculate the mean expression within each microarray. These data are encoded in the $7,077 \times 54$ matrix M.

The data for all three organisms in the form of expression matrices (genes in rows and samples in columns) are available at https://github.com/javier-pardodiaz/sdcorGCN and http://opig.stats.ox.ac.uk/resources.

2.2 Pre-processing

Gene expression data is noisy and the raw values are only comparable within the same experiment due to their arbitrary scales. Therefore, gene expression data requires some pre-processing before we can use it to generate networks (Libralon *et al.*, 2009). We apply quantile normalization (Bolstad *et al.*, 2003) to the gene expression matrix M. This normalization step renders the distribution of the expression values in different samples (i.e. the columns of M) identical in their statistical properties, such as maximum value and quantiles. This normalization enables us to compare data from different experiments.

To avoid interference from low expression values in the quantile normalization, we ignore the 20% least expressed genes from each sample before the normalization step. After the quantile normalization, we set the ignored values to the lowest expression value in M to decrease the level of noise. In practice, we have observed that 20% offers a good balance between preserving as much information as possible, and weeding out noisy measurements. We denote the pre-processed expression data by M^*, and its i-th row by $M^*_{i,\cdot}$.

2.3 Computing correlations between the expression of the genes

Distance correlation is as a measure of association between random vectors that addresses some of the limitations of linear measures such as Pearson's (Székely *et al.*, 2007). To compute the distance correlation between the expression of two genes i and j, let the vectors $M^*_{i,\cdot} = (M^*_{i,1}, \dots M^*_{i,n})$ and $M^*_{j,\cdot} = (M^*_{j,1}, \dots, M^*_{j,n})$ contain their expression values across the n samples. For each gene $i \in \{1, \dots, m\}$, we calculate the $n \times n$ expression distance matrix $\tilde{Y}^{(i)}$ whose entries are the absolute differences between the expression values across samples:

$$\tilde{Y}^{(i)}_{h,k} = |M^*_{i,h} - M^*_{i,k}|, \tag{1}$$

where $h, k \in \{1, \dots, n\}$ are all the samples in the data. Then, we compute the double-centred expression distance matrix $Y^{(i)}$:

Downloaded from https://academic.oup.com/bioinformatics/article/37/14/1982/6125359 by Bodleian Libraries of the University of Oxford user on 28 July 2022

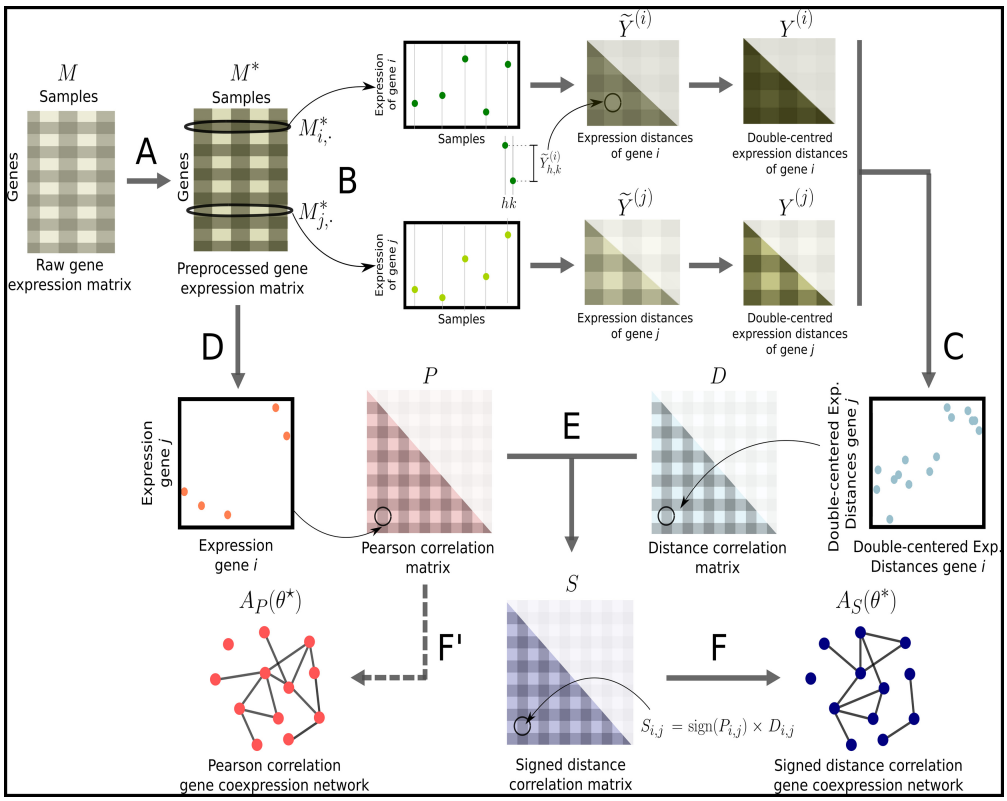


Fig. 1. Pipeline to construct networks from gene expression data using signed distance correlation. (**A**) We pre-process the input matrix M with raw gene expression data using quantile normalization and setting the lowest 20% values from each sample to the minimum value in M to obtain M^*. (**B**) We compute the expression distance matrices $\tilde{Y}^{(i)}$ and $\tilde{Y}^{(j)}$ for each gene $i, j \in \{1, \ldots, m\}$, and we double center them to obtain $Y^{(i)}$ and $Y^{(j)}$. (**C**) We compute the distance correlation matrix D, whose entries D_{ij} are the positive root of the Pearson correlation between $Y^{(i)}$ and $Y^{(j)}$, for every pair of genes. (**D**) We compute the Pearson correlation between each pair of rows in the M^* to obtain the Pearson correlation matrix P. (**E**) To construct the signed distance correlation matrix S, we multiply every distance correlation between the expression of two genes D_{ij} by the sign of their Pearson correlation $\text{sign}(P_{i,j})$. (**F**) Using COGENT (Bozhilova *et al.*, 2020), we find the optimal threshold θ^* that produces the most self-consistent network $A_S(\theta^*)$ from S. (**F′**) Analogously, we find the optimal threshold $\theta^\star$ to generate the network $A_P(\theta^\star)$ from P; this step is not part of the pipeline and only necessary to be able to compare Pearson and signed distance correlation networks.

$$Y^{(i)}_{h,k} = \tilde{Y}^{(i)}_{h,k} - \frac{1}{n}\sum_{k}\tilde{Y}^{(i)}_{h,k} - \frac{1}{n}\sum_{h}\tilde{Y}^{(i)}_{h,k} + \frac{1}{n^2}\sum_{h,k}\tilde{Y}^{(i)}_{h,k}. \tag{2}$$

where $a, b \in \{1, \ldots, n\}$. The distance covariance of the expression of genes i and j is:

$$\text{dcov}(i,j) = \frac{1}{n}\left(\sum_{h,k} Y^{(i)}_{h,k} Y^{(j)}_{h,k}\right)^{1/2}, \tag{3}$$

where $h, k \in \{1, \ldots, n\}$. Finally, the distance correlation between the expression of the genes is

$$\text{dcor}(i,j) = \frac{\text{dcov}(i,j)}{\sqrt{\text{dcov}(i,i)\text{dcov}(j,j)}}. \tag{4}$$

This expression is the non-negative square root of the Pearson correlation between $Y^{(i)}$ and $Y^{(j)}$. We store the pairwise distance correlation between the expression of all genes in the $m \times m$ symmetric matrix D with entries $D_{ij} = \text{dcor}(i,j)$.

The distance correlation in Equation (4) is always non-negative; however, the association between the expression of two genes can be either positive (i.e. both genes are expressed at the same time) or negative (i.e. one gene is expressed when the other is not). Thus, the sign of the association may contain crucial biological information that is lost if we only use distance correlation. We introduce sign into distance correlation by using the correlation matrix P that contains the Pearson correlation coefficients between the expression of each pair of genes:

$$P_{i,j} = \frac{\text{cov}(M^*_{i,\cdot}, M^*_{j,\cdot})}{\sigma_{M^*_{i,\cdot}}\sigma_{M^*_{j,\cdot}}}. \tag{5}$$

The matrix P also has size $m \times m$ but, unlike D, it may have negative values: values close to 1 mean strong positive correlation, values close to -1 mean strong negative correlation, and values around 0 mean no correlation. We generate a signed distance correlation matrix S whose values are the entries in D multiplied by the sign of the corresponding entries in P:

$$S_{i,j} = \text{sign}(P_{i,j})D_{i,j}. \tag{6}$$

For comparison purposes, we create the matrix I which contains the mutual information values for each pair of gene expression vectors from the preprocessed dataset M^*. To compute I, we use the R package minet (Meyer *et al.*, 2008). We select the empirical probability distribution as the entropy estimator, and use $\sqrt{n}$ bins with equal frequencies for the discretization.

2.4 Thresholding correlation matrices

Once we have a signed distance correlation S, we generate an unweighted network with adjacency matrix $A_S(\theta)$ by applying a threshold θ to S:

Downloaded from https://academic.oup.com/bioinformatics/article/37/14/1982/6125359 by Bodleian Libraries of the University of Oxford user on 28 July 2022

$$A_S(\theta)_{ij} = \begin{cases} 1 & \text{if } S_{i,j} \geq \theta, \\ 0 & \text{otherwise.} \end{cases} \quad (7)$$

The matrix $A_S(\theta)$ encodes an unweighted undirected network in which pairs of genes are connected if there is a strong and positive signed distance correlation in their expression. Naturally, different values of θ result in networks with different properties. We want to find the θ that minimizes the number of edges between genes that are not coexpressed, and maximizes the number of edges between genes that are coexpressed. For this purpose, we use the R package COGENT (Consistency of Gene Expression NeTworks) (Bozhilova *et al.*, 2020).

The main COGENT functions evaluate the internal consistency of a method to generate networks from a specific dataset. First, COGENT splits the gene expression data in two possibly overlapping groups of samples, and then constructs a network with the same node set from each group, G_1 and G_2. Then, COGENT measures the similarity between G_1 and G_2; the more similar the networks, the higher the internal consistency of the method. COGENT helps to find a threshold θ^* that results in the most internally consistent networks. The COGENT function getEdgeSimilarityCorrected provides a score of edge similarity between G_1 and G_2 and adjusts it so that results obtained for different values of θ are comparable. We use the semi-random density adjustment implemented in COGENT and select the function parameter that allows to keep the isolated nodes during the analysis.

The edge similarity between G_1 and G_2 is the Jaccard index of the set of edges. This index is adjusted using the randomized networks G_1^* and G_2^* from a a configuration-type model from the degree sequences d_1^* and d_2^*. The degree sequences d_1^* and d_2^* are random permutations of the degree sequences of G_1 and G_2. More precisely, the similarity between G_1 and G_2 is:

$$\text{sim}(G_1, G_2) = \frac{|E_1 \cap E_2| - \beta}{|E_1 \cup E_2| + \beta}, \quad (8)$$

where

$$\beta = \frac{|E_1| \sum_{(u,v)\in E_1} d_2^*(u) d_2^*(v)}{\sum_{u\in V_2} \sum_{\substack{v\in V_2 \\ v\neq u}} d_2^*(u) d_2^*(v)} + \frac{|E_2| \sum_{(u,v)\in E_2} d_1^*(u) d_1^*(v)}{\sum_{u\in V_1} \sum_{\substack{v\in V_1 \\ v\neq u}} d_1^*(u) d_1^*(v)},$$

and E_i, V_i are the set of edges and vertices for network G_i, $i = 1, 2$.

The value of β is the expected edge overlap between G_1 and G_2 if they were random networks. In general, β is higher for denser networks. The similarity $\text{sim}(G_1, G_2)$ in Eq. 8 is a value between $-1/3$ and 1; this value is high if G_1 is more similar to G_2 than to a randomization of G_2 and vice versa (Bozhilova *et al.*, 2020).

In our analysis, we create two random subsets of columns from M to obtain two $m \times \lfloor \frac{3n}{4} \rfloor$ matrices M_1 and M_2, such that half of the total number of samples n (i.e. the columns of M) are shared between both subsets, and 1/3 of their columns are different. Here, $\lfloor \frac{3n}{4} \rfloor$ denotes the greatest integer less or equal to $\frac{3n}{4}$. We choose this amount of overlap to evaluate how the relatively small change of about 1/3 of the data affects the final result. The choice of amount of overlap is user defined and may also depend on the research question. When the interest lies in clustering of genes, then one may like to choose an amount of overlap which leads to a moderately sparse network. Our previous *in silico* experiments show that if the proportion of samples from M shared between M_1 and M_2 is much smaller than 50%, the similarity between the obtained networks is low, and proportions much larger than 50% produce almost identical networks. We pre-process M_1 and M_2 and compute the signed distance correlation matrices S_1 and S_2 as outlined in Section 2.3. We test different values of θ, and for each of them, we obtain two unweighted networks and their similarity with COGENT. We repeat this whole process 25 times, every time with different subsets of columns of M. Finally, we compute the similarity score $s(\theta)$, which is the average of the similarity of the networks (in Eq. 8) over the 25 samples.

To favour signal over noise, we create a score that balances the similarity of the networks in $s(\theta)$ with the density of $A_S(\theta)$:

$$\text{Score}(\theta) = s(\theta) - \text{density}(A_S(\theta)). \quad (9)$$

We select the threshold θ^* that retrieves the highest $\text{Score}(\theta)$ and use this value to generate the unweighted gene coexpression network $A_S(\theta^*)$ from the signed distance correlation matrix S. We also calculate in the same manner the optimal thresholds $\theta^\star$ and $\theta^\diamond$ to construct the unweighted networks $A_P(\theta^\star)$ and $A_I(\theta^\diamond)$ from the Pearson correlation matrix P and the mutual information matrix I.

2.5 Network comparison

We compare gene coexpression networks obtained from thresholding the signed distance correlation matrix S, the Pearson correlation matrix P, and the MI matrix I. The density of a network may influence the amount of biological information it captures and hence we cannot compare networks with different densities. Instead, we generate additional networks from each matrix which match the different optimal edge densities. Letting the edge density of $A_S(\theta^*)$, $A_P(\theta^\star)$, and $A_I(\theta^\diamond)$ be d_S, d_P, and d_I, we compare the following nine networks:

- $NS(d_S)$: Network from S with edge density d_S (i.e. $A_S(\theta^*)$).
- $NP(d_P)$: Network from P with edge density d_P (i.e. $A_P(\theta^\star)$).
- $NI(d_I)$: Network from I with edge density d_I (i.e. $A_I(\theta^\diamond)$).
- $NP(d_S)$: Network from P with edge density d_S.
- $NI(d_S)$: Network from I with edge density d_S.
- $NS(d_P)$: Network from S with edge density d_P.
- $NI(d_P)$: Network from I with edge density d_P.
- $NS(d_I)$: Network from S with edge density d_I.
- $NP(d_I)$: Network from P with edge density d_I.

To construct the last six networks, we simply find a threshold manually that produces networks with density d_S, d_P, and d_I. Since for the *R.leguminosarum* dataset d_P is roughly the same as d_I (see Section 3), we use the network $NS(d_P)$ as a proxy for the network $NS(d_I)$ and omit networks $NP(d_I)$ and $NI(d_P)$.

We first evaluate the internal consistency of each network with COGENT. We also evaluate the biological information contained in the networks using STRING, a database of known and predicted protein–protein interactions (Szklarczyk *et al.*, 2019). STRING collects information from numerous sources, including experimental data, computational predictions and textmining. The association evidence in STRING is categorized into independent channels, weighted, and integrated to produce a confidence score C for all recorded protein interactions. Interactions with high C score are more likely to be true than those with a low score.

We work with three different sets of confidence scores:

- C: Total scores provided by STRING.
- $C^\dagger$: Scores that *only* consider coexpression information (coexpression channel combined with coexpression transferred channel).
- $C^\ddagger$: Scores that *exclude* coexpression information.

We provide details of how to retrieve $C^\dagger$ and $C^\ddagger$ in Supplementary Section 5.

For each network, we add the confidence score associated with each pair of connected genes and divide by the number of edges. We perform this operation independently for the three sets of confidence scores to obtain three aggregate confidence scores per network. These scores represent the amount of biological information that the networks capture. Then we compare these scores to the expected amount of biological information captured by chance. To do so, we generate three sets of 30 random networks with edge densities d_S, d_P, d_I, and evaluate their biological content following the approach detailed above. We compare the distribution of the aggregate C, $C^\dagger$ and $C^\ddagger$ scores from the random networks with density d_S with the scores from networks $NS(d_S)$, $NP(d_S)$, and $NI(d_S)$, the scores from the random networks with density d_P with $NS(d_P)$ and $NP(d_P)$, and the scores from the random networks with density d_P with $NS(d_I)$

Downloaded from https://academic.oup.com/bioinformatics/article/37/14/1982/6125359 by Bodleian Libraries of the University of Oxford user on 28 July 2022

and $NI(d_I)$. For the *R.leguminosarum* dataset, we omit the random networks with density d_I and compare $NI(d_I)$ with the network with density d_P due to the similarity in their densities. In Supplementary Section 2, we also perform the biological evaluation of the networks $NR(d_S)$ and $NR(d_P)$ with edge density d_S and d_P, respectively, obtained from a Spearman correlation matrix R.

3 Results

Here, we present our analysis of the *R.leguminosarum* dataset; the results for the yeast and human data are in Supplementary Sections 3 and 4.

3.1 Correlation matrices

The correlation matrices P (Pearson), D (distance), and S (signed distance) are all symmetric with $m = 7,077$ rows and columns. As we show in Table 1 and Figure 2, the distribution of the absolute values in P are different to those in D. In particular, the distribution of the values of P and S are different. Supplementary Figures S2 and S5 and Tables S4 and S7 contain the same analysis for the yeast and human data.

3.2 Gene coexpression networks

We first estimate, using COGENT (Bozhilova *et al.*, 2020), the optimal thresholds θ^*, $\theta^\star$, and $\theta^\diamond$ to construct the networks $NS(d_S)$, $NP(d_P)$, and $NI(d_I)$ from the matrices S, P, and I, respectively. To analyse and compare the networks, it is perhaps more intuitive to compare the networks using their edge density than with the threshold used to produce them. For this, we construct networks for a range of values of θ. For each density (and the θ that produced it), we evaluate its consistency score Score(θ) using Eq. 9. This score is related to the self-consistency of the network. Figure 3 shows the value of Score(θ) as a function of the density of the networks.

Table 1. Statistical summary of the correlation matrices from the *R.leguminosarum* data

Correlation	Min	1st q	Median	3rd q	Max	Mean
D	0.02	0.22	0.28	0.36	0.99	0.30
$\|P\|$	0.00	0.08	0.17	0.29	1.00	0.20
P	−0.86	−0.17	−0.01	0.17	1.00	0.00
S	−0.92	−0.27	−0.14	0.28	0.99	0.00

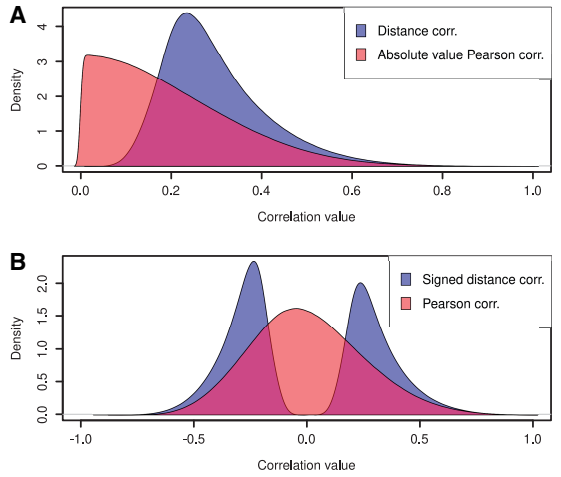


Fig. 2. Distribution of the entries of the correlation matrices from the *R.leguminosarum dataset*. (**A**) Distribution of distance correlations in D (blue), and the absolute of the Pearson correlations $|P|$ (red). (**B**) Distribution of signed distance correlation S (blue), and Pearson correlation P (red).

The highest score from the signed distance correlation networks $NS(d_S)$ is Score(θ^*) = 0.456, where $\theta^* = 0.62$, and density $d_S = 1.25\%$. The highest score from the Pearson correlation networks $NP(d_P)$ is Score($\theta^\star$) = 0.426, where $\theta^\star = 0.58$, and density $d_P = 1.63\%$. The highest score from the mutual information networks $NI(d_I)$ is Score($\theta^\diamond$) = 0.140, where $\theta^\diamond = 0.65$, and density $d_I = 1.65\%$. We also create the networks $NP(d_S)$ from P and $NI(d_I)$ from I to match the edge density of $NS(d_S)$, and the network $NS(d_P)$ from S to match the edge density of $NP(d_P)$ and $NI(d_I)$. Table 2 contains a statistical summary of the networks. The discrepancy in the number of edges in $NS(d_S)$ and $NI(d_S)$ is due to several pairs of genes having the same mutual information value. For both edge densities, the networks retrieved using our signed distance correlation matrix S ($NS(d_S)$ and $NS(d_P)$) have a smaller, denser, and with higher global clustering coefficient largest connected component (LCC) than the networks obtained using Pearson correlation ($NP(d_S)$ and $NP(d_P)$). The LCC of the networks obtained using mutual information ($NI(d_S)$ and $NI(d_I)$) are smaller and denser than those from density-matching networks but they show the lowest global clustering coefficient. See Supplementary Tables S5 and S8 for the gene coexpression networks for the yeast and human data.

3.3 Network evaluation

We perform two evaluations of the networks from signed distance correlation, Pearson correlation, and mutual information: self-consistency and biological content. We are interested in the self-consistency of the networks because it reflects their ability to cope with changes in the data used to generate them. If our network is more self-consistent, then we can have greater confidence in the biological conclusions we draw from it, even if the data is imperfect or noisy. The Jaccard similarity in Eq. 8 measures the similarity between networks generated from overlapping, non-identical subsets of a dataset using the same network construction method. The higher the similarity, the higher the internal consistency of the method and the more self-consistent the network obtained from applying it is. We measure the self-consistency Score(θ), obtained from computing the average Jaccard similarity of networks from different randomized subsets of data and subtracting the density of $A_S(\theta)$ ($A_P(\theta)$ and $A_I(\theta)$ for Pearson correlation and mutual information, respectively). As we show in Figure 3, the self-consistency scores of the networks from signed distance correlations are consistently higher than in the Pearson and mutual information networks over an interval of densities that includes the maxima for all three methods. The scores obtained for the mutual information networks are the lowest. From this analysis, we conclude that signed distance correlation networks are more self-consistent than networks based on Pearson correlation or mutual information. Even the optimal threshold for the Pearson matrix P produces a less self-consistent network $NP(d_P)$ than a signed distance correlation network with a matching edge density ($NS(d_P)$); the same applies for mutual information. See Supplementary Figures S3 and S6 for the same analysis on the yeast and human data, with similar results.

To evaluate the biological content of a network, we add the STRING scores of the edges using: all the information in STRING (C), only coexpression information ($C^\dagger$), and everything except coexpression information ($C^\ddagger$), and then divide by the number of edges in the network, as described in Section 2.5. Table 3 and Figure 4 (and Supplementary Fig. S1) show the results for all the networks, and the mean scores from random networks. In every case, the signed distance correlation networks contain more biological information than networks from Pearson correlation, those obtained using mutual information, and the randomized networks. See Supplementary Tables S6 and S9 and Figures S4 and S7 for a similar analysis on the human and yeast data, with similar results. The highest difference of the networks with the randomized networks occurs when we only use coexpression information. The $C^\dagger$ scores of $NS(d_S)$, $NP(d_S)$, and $NI(d_S)$ are 10, 8.5, and 8.7 times higher than the expected ones for random networks. This is not surprising

Downloaded from https://academic.oup.com/bioinformatics/article/37/14/1982/6125359 by Bodleian Libraries of the University of Oxford user on 28 July 2022

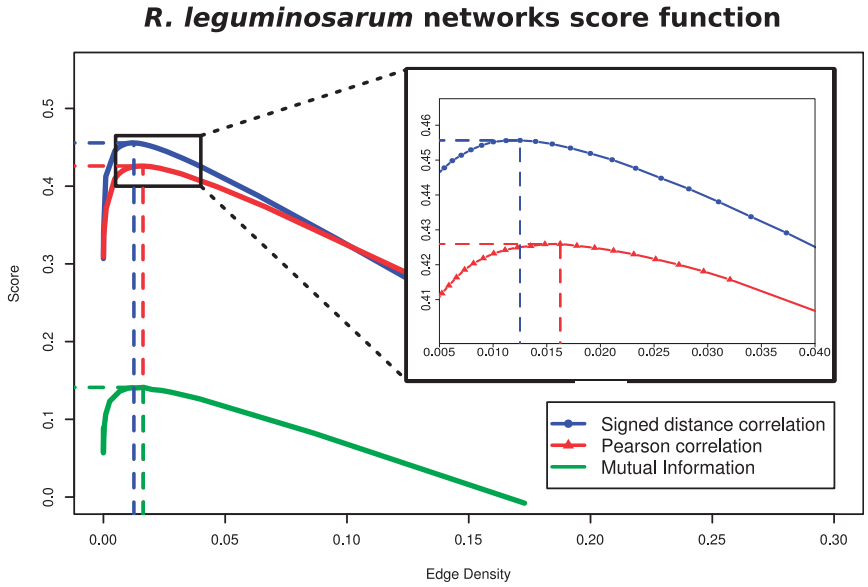


Fig. 3. Self-consistency scores of *R.leguminosarum* networks with edge densities between 0 and 0.3. The blue line with circles shows the scores of networks obtained using signed distance correlations; the red line with triangles shows the score of networks using Pearson correlations; the green line shows the scores of the networks obtained using MI. The dashed vertical lines indicate the density of the most self-consistent network for each type of correlation.

Table 2. Statistical summary of *R.leguminosarum* coexpression networks. LCC denotes largest connected component

Network	Number of edges	Number of vertices in LCC	Edge density LCC×100	Global clustering coeff. LCC
$NS(d_S)$	313 348	6431	1.52	0.570
$NS(d_P)$	406 977	6688	1.82	0.570
$NP(d_S)$	313 348	6697	1.40	0.544
$NP(d_P)$	406 977	6880	1.72	0.543
$NI(d_S)$	317 014	5993	1.77	0.279
$NI(d_I)$	414 140	6084	2.24	0.284

Table 3. Evaluation of the biological content of the networks with STRING

Network	All of STRING information (C)	Only coexpression information ($C^{\dagger}$)	All information except coexpression ($C^{\ddagger}$)
$NS(d_S)$	**29.61**	**9.02**	**26.55**
$NP(d_S)$	25.23	7.63	22.72
$NI(d_S)$	24.68	7.81	22.01
RE d_S	7.49 ± 0.08	0.90 ± 0.02	7.09 ± 0.08
$NS(d_P)$	**26.79**	**7.64**	**24.12**
$NP(d_P)$	23.32	6.61	21.08
$NI(d_I)$	22.39	6.60	20.07
RE d_P	7.54 ± 0.08	0.91 ± 0.03	7.13 ± 0.08

Note: RE indicates the expected (mean) result based on random networks with the indicated edge density and its SD. The values in bold correspond to the highest scores for each set of networks and confidence scores.

because we have built the networks using gene expression information. However, the coexpression data used to construct these networks is different to the data in STRING. The scores excluding the coexpression information $C^{\ddagger}$ are 3.7, 3.2, and 3.1 times higher than in the random networks. This is remarkable because this assessment is performed on data that is completely different than the data used to construct the networks. Hence, by applying our pipeline to gene expression data, we can identify new types of relationships between proteins (and genes). These results demonstrate the power of gene coexpression networks to predict functional interactions, especially when constructed using signed distance correlation.

The analysis for networks with d_P and d_I tells a similar story; the scores obtained by the signed distance correlation network $NS(d_P)$ are higher than the scores from the Pearson correlation network $NP(d_P)$ and the mutual information network $NS(d_I)$, despite d_P and d_I being the optimal edge density for Pearson correlation and mutual information. We highlight the fact that in absolute terms (i.e. not dividing the scores by the number of edges in the network) when we use only

Downloaded from https://academic.oup.com/bioinformatics/article/37/14/1982/6125359 by Bodleian Libraries of the University of Oxford user on 28 July 2022

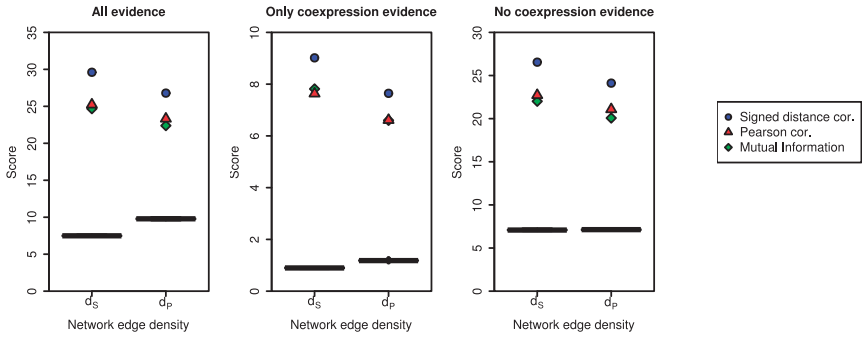


Fig. 4. Scores obtained for the *R.leguminosarum* gene coexpression networks using STRING. All panels show the score for the different networks in the *y* axis, and the network density on the *x* axis. The scores are the result of adding up the confidence scores with all evidence (C), only coexpression evidence ($C^{\dagger}$) and everything excluding coexpression ($C^{\ddagger}$) from STRING associated with the edges in the networks, each computed using different of information. The black box plots correspond to the scores obtained by 30 random networks. Blue circles, red triangles, and green diamonds represent signed distance correlation, Pearson correlation, and mutual information respectively.

coexpression information ($C^{\dagger}$), the score of $NS(d_S)$ is higher than the score of $NP(d_P)$, even though the former has fewer edges than the latter, and therefore its score is the result of adding fewer terms.

For the three datasets we analyse, signed distance correlation networks perform better than Pearson correlation and mutual information networks with matching densities, according to both evaluations and for all tested edge densities.

4 Discussion

In this work, we have introduced signed distance correlation, and presented a method to construct networks in a self-consistent way exclusively from gene expression data. This method has three main steps: data pre-processing, computing correlations, and thresholding. These steps combine well-established methods such as quantile normalization, and the use of COGENT (Bozhilova *et al.*, 2020) to identify the optimal threshold.

Distance correlation is an intuitive approach to study gene expression because it relies on the differences in the expression between samples. By incorporating signs into distance correlation, we can also differentiate between positive and negative relationships, and maintain the advantages of distance correlation.

We apply our method to data from *R.leguminosarum*, yeast, and human. In all cases, our method produces networks that are more self-consistent than using Pearson correlation and mutual information. The reason why self-consistency is so important is that it ensures that our results are robust to changes or noise in the data. Therefore, when we cannot assess the biological significance of a network directly, we can use its self-consistency as an indication of biological significance. Networks from signed distance correlation also capture more biological information than networks from Pearson correlation and mutual information, as shown by our evaluation using STRING. In the case of *R.leguminosarum*, the signed distance correlation network (using an optimal threshold θ^* found with COGENT) captures almost four times more biological information than random networks. The amount of captured biological information is less if we use Pearson or Spearman correlations (Supplementary Fig. S1) or mutual information to build the networks.

These results give a weak indication that self-consistent networks derived from biological data might capture more biological information than those with less stability. Therefore, if we cannot assess the biological significance of a network directly, measuring its self-consistency may serve as a proxy. The use of COGENT to select the threshold values requires a sufficiently strong signal in the association between the expression values of different genes. For example, using Euclidean distance, scaled by the square root of the sample size, did not yield sufficient signal, so that networks created by the subsampling method had no more overlap than expected at random. These networks also did not capture much biological information (data not shown).

We applied our method to construct, to our knowledge, the first gene coexpression network for *R.leguminosarum*. This network promises to reveal rich biological information that will illuminate our understanding of plant–bacteria interactions and nitrogen fixation, and it is therefore the starting point for further investigations of the biological mechanisms of this organism. In particular, we plan to identify groups of genes in *R.leguminosarum* which are highly connected in the network and associate them with specific biological processes. To do so, we will make use of community detection techniques and new experimental data. For the human liver dataset one could explore predicting disease-related biological information; see for example Song *et al.* (2019); Chen *et al.* (2018); Li *et al.* (2019).

Finally, we have showcased our method on gene expression datasets from different organisms obtained using different techniques: microarrays (*R.leguminosarum*), RNA-Seq (yeast), and single-cell RNA-Seq (human). However, the methods that we have developed are general, and can also be used to construct networks in a vast range of domains, such as, for example, economics (Wang *et al.*, 2018), neuroscience (Bernhardt *et al.*, 2011), climatology (Donges *et al.*, 2009), or any discipline where networks can be constructed from correlation data.

Acknowledgements

We thank Alison K. East and Florian Klimm for help and advice on how to analyse the datasets used in this manuscript. We thank Katharine Turner (ANU), Florian Klimm, and Malte D. Luecken for fruitful discussions. In addition, we acknowledge support from COST Action CA15109, Keble College Oxford and Keble Association. The authors would like to acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility in carrying out this work (http://dx.doi.org/10.5281/zenodo.22558), and Carlos Outerial for assistance in the computing.

Funding

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) (EP/R512333/1 to J.P.-D., M.B.-D., P.S.P., C.M.D. and G.R., EP/L016044/1 to L.V.B., C.M.D. and G.R.), the Biotechnology and Biological Sciences Research Council (BBSRC) (BB/T001801/1 to P.S.P. and G.R.), the COSTNET COST Action (CA15109 to G.R.) and e-Therapeutics plc. M.B.-D. acknowledges support from the Oxford-Emirates Data Science Lab.

Conflict of Interest: none declared.

Downloaded from https://academic.oup.com/bioinformatics/article/37/14/1982/6125359 by Bodleian Libraries of the University of Oxford user on 28 July 2022

References

Bar-Joseph,Z. *et al.* (2003) Computational discovery of gene modules and regulatory networks. *Nat. Biotechnol.*, **21**, 1337–1342.

Bernhardt,B.C. *et al.* (2011) Graph-theoretical analysis reveals disrupted small-world organization of cortical thickness correlation networks in temporal lobe epilepsy. *Cereb. Cortex*, **21**, 2147–2157.

Bolstad,B.M. *et al.* (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Bozhilova,L.V. *et al.* (2020) COGENT: evaluating the consistency of gene co-expression networks. *Bioinformatics*, btaa787.

Chen,X. *et al.* (2018) MDHGI: matrix decomposition and heterogeneous graph inference for mirna-disease association prediction. *PLoS Comput. Biol.*, **14**, e1006418.

Donges,J.F. *et al.* (2009) Complex networks in climate dynamics. *Eur. Phys. J. Spec. Top.*, **174**, 157–179.

George,G. *et al.* (2019) Gene co-expression network analysis for identifying genetic markers in Parkinson's disease-a three-way comparative approach. *Genomics*, **111**, 819–830.

Ghanbari,M. *et al.* (2019) The distance precision matrix: computing networks from non-linear relationships. *Bioinformatics*, **35**, 1009–1017.

Guo,X. *et al.* (2014) Inferring nonlinear gene regulatory networks from gene expression data based on distance correlation. *PLoS One*, **9**, e87446.

Hughes,T.R. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.

Ihmels,J. *et al.* (2002) Revealing modular organization in the yeast transcriptional network. *Nat. Genet.*, **31**, 370–377.

Karunakaran,R. *et al.* (2009) Transcriptomic analysis of *Rhizobium leguminosarum biovar viciae* in symbiosis with host plants *Pisum sativum* and *Vicia cracca*. *J. Bacteriol.*, **191**, 4002–4014.

Kothapalli,R. *et al.* (2002) Microarray results: how accurate are they? *BMC Bioinformatics*, **3**, 22.

Langfelder,P. and Horvath,S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.

Lee,H.K. *et al.* (2004) Coexpression analysis of human genes across many microarray data sets. *Genome Res.*, **14**, 1085–1094.

Li,H. *et al.* (2019) A novel human microbe-disease association prediction method based on the bidirectional weighted network. *Front. Microbiol.*, **10**, 676.

Libralon,G.L. *et al.* (2009) Pre-processing for noise detection in gene expression classification data. *J. Braz. Comput. Soc.*, **15**, 3–11.

Magwene,P.M. and Kim,J. (2004) Estimating genomic coexpression networks using first-order conditional independence. *Genome Biol.*, **5**, R100.

Makrodimitris,S. *et al.* (2020) Metric learning on expression data for gene function prediction. *Bioinformatics*, **36**, 1182–1190.

Meyer,P.E. *et al.* (2008) minet: ar/bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinformatics*, **9**, 461.

Pini,F. *et al.* (2017) Lux bacterial biosensors for in vivo spatiotemporal mapping of root secretion. *Plant Physiol.*, **174**, 1289–1306.

Ramachandran,V.K. *et al.* (2011) Adaptation of *Rhizobium leguminosarum* to pea, alfalfa and sugar beet rhizospheres investigated by comparative transcriptomics. *Genome Biol.*, **12**, R106.

Song,F. *et al.* (2019) mies: predicting the essentiality of mirnas with machine learning and sequence features. *Bioinformatics*, **35**, 1053–1054.

Stuart,J.M. *et al.* (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**, 249–255.

Székely,G.J. *et al.* (2007) Measuring and testing dependence by correlation of distances. *Ann. Stat.*, **35**, 2769–2794.

Szklarczyk,D. *et al.* (2019) STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.*, **47**, D607–D613.

Ucar,D. *et al.* (2007) Construction of a reference gene association network from multiple profiling data: application to data analysis. *Bioinformatics*, **23**, 2716–2724.

van Noort,V. *et al.* (2003) Predicting gene function by conserved co-expression. *Trends Genet.*, **19**, 238–242.

Wang,G.-J. *et al.* (2018) Correlation structure and evolution of world stock markets: evidence from Pearson and partial correlation-based networks. *Comput. Econ.*, **51**, 607–635.

Weirauch,M.T. (2011) Gene coexpression networks for the analysis of DNA microarray data. *Appl. Stat. Netw. Biol.*, **1**, 215–250.

Wolfe,C.J. *et al.* (2005) Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. *BMC Bioinformatics*, **6**, 227.

Young,J.P.W. *et al.* (2006) The genome of *Rhizobium leguminosarum* has recognizable core and accessory components. *Genome Biol.*, **7**, R34.

Zhao,H. *et al.* (2018) Prediction of microRNA-disease associations based on distance correlation set. *BMC Bioinformatics*, **19**, 141.

Downloaded from https://academic.oup.com/bioinformatics/article/37/14/1982/6125359 by Bodleian Libraries of the University of Oxford user on 28 July 2022

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Robust gene coexpression networks using signed distance correlation
Publication Status	Published
Publication Details	Javier Pardo-Diaz, Lyuba V Bozhilova, Mariano Beguerisse-Díaz, Philip S Poole, Charlotte M Deane, Gesine Reinert, Robust gene coexpression networks using signed distance correlation, *Bioinformatics*, Volume 37, Issue 14, 15 July 2021, Pages 1982–1989, https://doi.org/10.1093/bioinformatics/btab041

Student Confirmation

Student Name:	Javier Pardo Díaz (Javier Pardo-Diaz)		
Contribution to the Paper	Conceptualization (equal), data curation (lead), formal analysis, investigation, methodology(lead), software, validation (lead), visualization (lead), writing—original draft (lead), and writing—review and editing (equal) COGENT was conceptualised and implemented by Dr Lyuba V. Bozhilova. I developed density adjustment using random samples from the configuration model in Section 2.4.		
Signature		Date	19th September 2022

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Professor Gesine Reinert		
Supervisor comments This is an accurate description.		
Signature 	Date	September 20, 2022

This completed form should be included in the thesis, at the end of the relevant chapter.

Bioinformatics, 37(13), 2021, 1928–1929
doi: 10.1093/bioinformatics/btaa787
Advance Access Publication Date: 15 September 2020
Applications Note

OXFORD

Gene expression

COGENT: evaluating the consistency of gene co-expression networks

Lyuba V. Bozhilova , Javier Pardo-Diaz , Gesine Reinert and Charlotte M. Deane *

Department of Statistics, University of Oxford, Oxford OX1 3LB, United Kingdom

*To whom correspondence should be addressed.

Associate Editor: Cowen Lenore
Received on June 25, 2020; revised on August 28, 2020; editorial decision on September 2, 2020; accepted on September 3, 2020

Abstract

Summary: Gene co-expression networks can be constructed in multiple different ways, both in the use of different measures of co-expression, and in the thresholds applied to the calculated co-expression values, from any given dataset. It is often not clear which co-expression network construction method should be preferred. COGENT provides a set of tools designed to aid the choice of network construction method without the need for any external validation data.

Availability and implementation: https://github.com/lbozhilova/COGENT.

Contact: deane@stats.ox.ac.uk

Supplementary information: Supplementary information is available at *Bioinformatics* online.

1 Introduction

Gene expression data are a powerful resource for understanding genetic function under different conditions. A common way of exploring this data is through gene co-expression networks (Lee *et al.*, 2004). In these networks, genes are represented by nodes and highly co-expressed gene pairs are connected by edges. Such networks have been used in many ways, including for gene function prediction and the identification of disease- or tissue-relevant gene modules (van Dam *et al.*, 2017).

Gene expression data typically take the form of a matrix, in which rows correspond to genes and columns correspond to samples. Network construction commonly consists of three steps—the data are pre-processed, a measure of co-expression is calculated for every pair of genes and a score cut-off is applied. Different approaches to data pre-processing and normalization exist (Abbas-Aghababazadeh *et al.*, 2018; Park *et al.*, 2003). Further, after normalization co-expression can be calculated in a number of different ways—e.g. via a correlation coefficient or mutual information. A score cut-off is then usually imposed in order to identify gene pairs which are highly co-expressed. Alternatively, weighted networks can be analysed, in which edge weights correspond to levels of co-expression (Langfelder and Horvath, 2008).

There are many available methods for network construction which can be applied to the same dataset, and which then lead to different networks. Enrichment analysis or comparison to orthogonal data, such as protein interaction data, is commonly used for network selection and validation. However, for many species and datasets poor or non-existent functional annotation makes this type of validation difficult. It is therefore often not clear which of the available network construction methods should be prioritized (De Smet and Marchal, 2010).

Here, we introduce COGENT (COnsistency of Gene Expression NeTworks), an R package designed to aid the choice of a network construction pipeline without the need for annotation or external data. COGENT can be used to choose between competing co-expression measures, as well as to inform score cut-off choice. While designed for gene expression data, COGENT can be applied to other cases where network construction relies on similarity profiling, e.g. microbiome or synthetic lethality data.

2 Software description

2.1 Method overview

Consistent co-occurrence of two gene products in the cell points towards a functional relationship between the genes. Gene product abundance, as well as co-occurrence, is a continuous-time phenomenon, which is experimentally observed at discrete time points or samples. The construction of a gene co-expression network can therefore be thought of as an estimation problem—we aim to infer general co-expression patterns from a limited set of data points. One way of investigating the success of such a procedure is through resampling. Networks constructed from a subset of all available samples will be noisier than the network constructed from the full dataset. However, they should still resemble each other: if subsetting the data results in networks with little to no overlap, then the network construction procedure may be too sensitive to noise in the data.

COGENT evaluates network construction methods through iterative resampling. At each step, the gene expression samples are split into two possibly overlapping sets of equal size. The same network construction function $f(\cdot)$ is applied to both sets in order to obtain two gene co-expression networks $G_1 = (V, E_1)$ and

©The Author(s) 2020. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/bioinformatics/article/37/13/1928/5906022 by Oxford University user on 07 October 2022

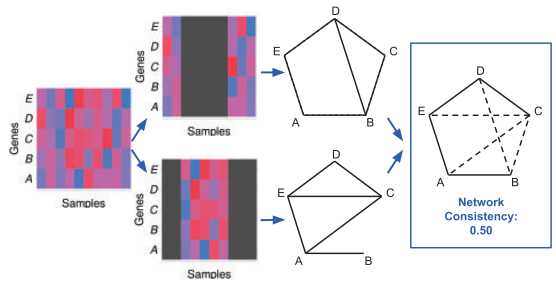


Fig. 1. COGENT workflow schematic. In this example, the input data are a gene expression matrix with rows corresponding to genes—in this case A, B, C, D, E—and columns corresponding to samples (far left). First, the expression matrix columns are randomly split into two possibly overlapping groups of equal size (left). Then, a network is constructed from each of the sample groups (right). Finally, the two resulting networks are compared and the consistency between them is calculated (far right). In this example, the two networks have six edges each, and overlap at four of these edges. One measure of their consistency is the Jaccard index between their edge sets (see Supplementary Section S4), which in this case is 0.50

$G_2 = (V, E_2)$. COGENT then calculates several measures of consistency between these two networks. An example workflow for calculating consistency is shown in Figure 1; see Supplementary Section S2 for workflow details and Supplementary Section S4 for different consistency measures.

The entire procedure is repeated multiple times in order to obtain robust results. Network construction methods which result in highly similar pairs of networks $G_1 \approx G_2$ are considered to be consistent. When two or more competing methods are considered, the method exhibiting higher internal consistency should be preferred.

2.2 Edge set consistency

Consistency in COGENT is measured through a network comparison step at each iteration. The two networks G_1 and G_2 are considered to be similar if their edge sets are similar ($E_1 \approx E_2$). We measure agreement between E_1 and E_2 using a (weighted) Jaccard index to produce two measures of edge set consistency—global and local similarity (Kao and Porter, 2018; see Supplementary Section S4.1).

Since global similarity scales with network density, a density adjustment is required when methods resulting in different network densities are compared. This is particularly important when choosing a score cut-off. Density adjustment in COGENT is carried out through comparison to random networks generated using a configuration model (see Supplementary Section S4.2).

2.3 Node metric consistency

Density-independent network comparison can also be performed by calculating a node metric such as degree or betweenness for all nodes in each of the two networks, and then comparing the obtained metric values. If the aim of downstream network analysis is e.g. to identify genes with many co-expression partners, the degree is a natural node metric to use.

At each COGENT iteration, a node metric set by the user can be applied to G_1 and G_2, resulting in two node metric vectors d_1 and d_2, respectively. These two can be compared in three different ways: via a correlation coefficient, rank k-similarity (Bozhilova *et al.*, 2019; Trajanovski *et al.*, 2013) and Euclidean distance; see Supplementary Section S4.3 for details.

3 Application

Global and local similarity and adjusted edge consistency, as well as node metric comparisons can all be used to evaluate network consistency. By iteratively resampling the data and measuring network consistency, COGENT can be used to prioritize different network construction pipelines, as well as to inform co-expression cut-offs without the need for external validation data. A full worked example can be found in Supplementary Section S5 and in the COGENT tutorial. COGENT has also been used to assess signed distance correlation as a measure of gene co-expression (Pardo-Diaz *et al.*, 2021). This application further shows that network construction methods prioritized by COGENT also capture more protein–protein interaction data than methods which were not prioritized.

While originally developed for gene expression data, COGENT can also be used for the inspection of other data types. For example, it can be applied to microbiome data in order to identify symbiotic organisms, or to synthetic lethality data in order to identify genetic interactions with high confidence.

Funding

This work was supported by the EPSRC [EP/L016044/1 to L.V.B. and C.M.D., EP/R512333/1 to J.P.D., G.R. and C.M.D.], the BBSRC [BB/T001801/1 to G.R.], COST [CA15109 to G.R.] and e-Therapeutics plc.

Conflict of Interest: none declared.

References

Abbas-Aghababazadeh,F. *et al.* (2018) Comparison of normalization approaches for gene expression studies completed with high-throughput sequencing. *PLoS One*, **13**, e0206312.

Bozhilova,L.V. *et al.* (2019) Measuring rank robustness in scored protein interaction networks. *BMC Bioinform.*, **20**, 446.

De Smet,R. and Marchal,K. (2010) Advantages and limitations of current network inference methods. *Nat. Rev. Microbiol.*, **8**, 717.

Kao,T.-C. and Porter,M.A. (2018) Layer communities in multiplex networks. *J. Stat. Phys.*, **173**, 1286–1302.

Langfelder,P. and Horvath,S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinform.*, **9**, 559.

Lee,H.K. *et al.* (2004) Coexpression analysis of human genes across many microarray data sets. *Genome Res.*, **14**, 1085–1094.

Pardo-Diaz,J. *et al.* (2021) Robust gene coexpression networks using signed distance correlation. *Bioinformatics*, doi: 10.1093/bioinformatics/btab041.

Park,T. *et al.* (2003) Evaluation of normalization methods for microarray data. *BMC Bioinform.*, **4**, 33.

Trajanovski,S. *et al.* (2013) Robustness envelopes of networks. *J. Complex Netw.*, **1**, 44–62.

van Dam,S. *et al.* (2017) Gene co-expression analysis for functional classification and gene–disease predictions. *Brief. Bioinform.*, **19**, 575–592.

Downloaded from https://academic.oup.com/bioinformatics/article/37/13/1928/5906022 by Oxford University user on 07 October 2022

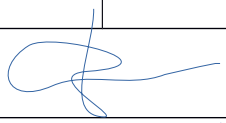
Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

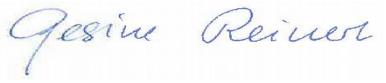
Title of Paper	COGENT: evaluating the consistency of gene co-expression networks
Publication Status	Published
Publication Details	Lyuba V. Bozhilova, Javier Pardo-Diaz, Gesine Reinert, Charlotte M. Deane, COGENT: evaluating the consistency of gene co-expression networks, *Bioinformatics*, Volume 37, Issue 13, 1 July 2021, Pages 1928–1929, https://doi.org/10.1093/bioinformatics/btaa787

Student Confirmation

Student Name:	Javier Pardo Díaz (Javier Pardo-Diaz)		
Contribution to the Paper	Conceptualization (secondary), methodology(secondary), writing—review and editing (secondary) COGENT was conceptualised and implemented by Dr Lyuba V. Bozhilova. I developed the density adjustment using random samples from the configuration model in Section 2.2 and contributed by suggesting applying COGENT to threshold gene coexpression networks.		
Signature		Date	19th September 2022

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Professor Gesine Reinert		
Supervisor comments This is an accurate description.		
Signature 	Date	September 20, 2022

This completed form should be included in the thesis, at the end of the relevant chapter.

3

Construction of weighted and thresholded gene coexpression networks

Network Science **10** (2): 131–145, 2022
doi:10.1017/nws.2022.13

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Generating weighted and thresholded gene coexpression networks using signed distance correlation

Javier Pardo-Diaz[1], Philip S. Poole[2], Mariano Beguerisse-Díaz[3], Charlotte M. Deane[1] and Gesine Reinert[1]*

[1]Department of Statistics, University of Oxford, Oxford OX1 3LB, UK, [2]Department of Plant Sciences, University of Oxford, Oxford OX1 3RB, UK and [3]Mathematical Institute, University of Oxford, Oxford OX2 6GG, UK
*Corresponding author. Email: reinert@stats.ox.ac.uk

Action Editor: Christoph Stadtfeld

Abstract

Even within well-studied organisms, many genes lack useful functional annotations. One way to generate such functional information is to infer biological relationships between genes or proteins, using a network of gene coexpression data that includes functional annotations. Signed distance correlation has proved useful for the construction of unweighted gene coexpression networks. However, transforming correlation values into unweighted networks may lead to a loss of important biological information related to the intensity of the correlation. Here, we introduce a principled method to construct *weighted* gene coexpression networks using signed distance correlation. These networks contain weighted edges only between those pairs of genes whose correlation value is higher than a given threshold. We analyze data from different organisms and find that networks generated with our method based on signed distance correlation are more stable and capture more biological information compared to networks obtained from Pearson correlation. Moreover, we show that signed distance correlation networks capture more biological information than unweighted networks based on the same metric. While we use biological data sets to illustrate the method, the approach is general and can be used to construct networks in other domains. Code and data are available on https://github.com/javier-pardodiaz/sdcorGCN.

Keywords: gene expression; robustness; weighted networks; correlation

1. Introduction

Gene expression is the process by which genes in different organisms are *activated* to produce proteins when they are needed to carry out their function. Data relating to gene expression data contains key information about intracellular biological processes (Kothapalli et al., 2002). Gene coexpression datasets typically describe the expression level of different genes across different samples often taken under different experimental conditions. Such data are frequently represented as gene coexpression networks, with nodes representing genes and edges representing correlations in expression between pairs of genes across multiple samples (Lee et al., 2004). Representing gene coexpression as networks helps in the study and visualization of the expression data and the exploitation of the structure of interactions between genes at a whole-system level (Weirauch, 2011; Magwene & Kim, 2004). One motivation behind creating these networks is that genes which are highly coexpressed across multiple samples are likely to have related functions (Hughes et al., 2000; Stuart et al., 2003; van Noort et al., 2003; Makrodimitris et al., 2020), allowing inference of gene function using *guilt by association* approaches (Wolfe *et al.*, 2005). This procedure is especially useful if the studied organism is poorly annotated. However, the lack of reliable genomic information can hinder the validation of the accuracy of the network models generated from gene

© The Author(s), 2022. Published by Cambridge University Press.

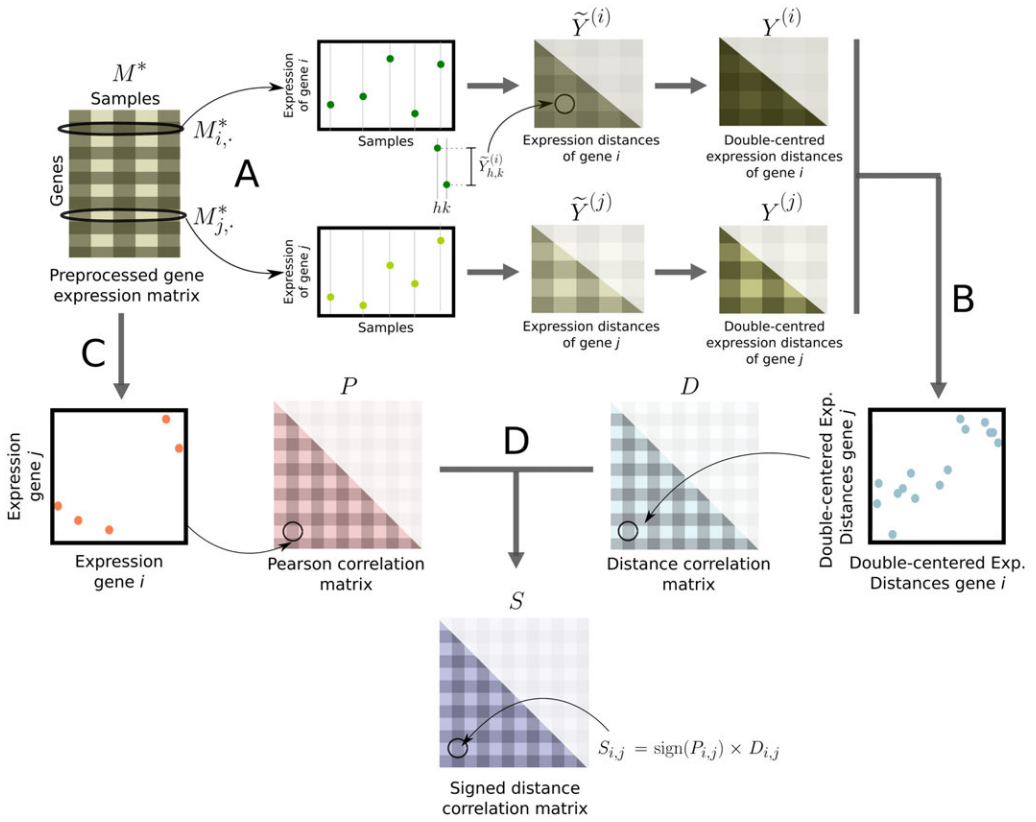


Figure 1. Pipeline to construct networks from gene expression data using signed distance correlation, adapted from Pardo-Diaz et al. (2021). **A**: We compute the expression distance matrices $\widetilde{Y}^{(i)}$ and $\widetilde{Y}^{(j)}$ for each gene $i, j \in \{1, \ldots, m\}$, and we double center them to obtain $Y^{(i)}$ and $Y^{(j)}$. **B**: We compute the distance correlation matrix D, whose entries $D_{i,j}$ are the positive root of the Pearson correlation between $Y^{(i)}$ and $Y^{(j)}$, for every pair of genes. **C**: We compute the Pearson correlation between each pair of rows in the M^* to obtain the Pearson correlation matrix P. **D**: To construct the signed distance correlation matrix S, we multiply every distance correlation between the expression of two genes $D_{i,j}$ by the sign of their Pearson correlation $\text{sign}(P_{i,j})$.

expression data. Noise in the data makes it difficult to distinguish genes that are expressed at a low level from those not expressed. Therefore, there is a need for network construction pipelines that are robust to experimental error, yet capture most of the information in the data.

Multiple strategies to construct a gene coexpression networks from gene expression data are available (Bar-Joseph et al., 2003; Ihmels et al., 2002; Ucar et al., 2007; Langfelder & Horvath, 2008; George et al., 2019). A recent novel approach uses signed distance correlation (Pardo-Diaz et al., 2021). This approach identifies coexpression relationships between genes and produces robust networks that capture more biological information than those obtained using alternative metrics such as Pearson correlation, Spearman correlation, and mutual information (Pardo-Diaz et al., 2021). Signed distance correlation is based on distance correlation (Székely et al., 2007), a measure that evaluates an association between the pattern of changes in the samples, and provides a non-negative score that is zero if and only if the expression vectors are statistically independent. Because it is unsigned, distance correlation values do not permit the differentiation of genes with the same expression pattern and genes expressed at opposite times. Signed distance correlation overcomes this problem by combining the distance correlation value with a sign that indicates the direction of the association. The sign corresponds to the sign of the Pearson correlation between the expression of the two genes across the samples in the dataset. Figure 1 illustrates how signed distance correlation is obtained.

https://doi.org/10.1017/nws.2022.13 Published online by Cambridge University Press

To our knowledge, signed distance correlation has been used in this field only to generate unweighted networks, using the R package COGENT (Bozhilova et al., 2020). COGENT aids the selection of a robust network construction method without the need for any external validation data. The main functions in this package iteratively split the gene expression data into two different sets, construct a network from each of them and then evaluate their similarity; the higher the similarity of the networks, the more robust the construction method. The functions implemented in COGENT to measure the similarity between the two generated networks are suitable to analyze only unweighted networks and complete weighted networks.

COGENT assists the selection of the optimal threshold value so that only pairs of genes for which the correlation value of their expression exceeds the threshold are connected in the network. The use of a threshold to construct networks results in a relatively sparse unweighted network. However, unweighted networks obtained through thresholding weighted edges may ignore important information about the strength of the correlation. Avoiding the use of a threshold results in complete weighted networks in which the weights of the edges are proportional to the correlation values. In this type of networks, it may be difficult to distinguish the signal from the noise. Thus, there is a need for a method that eases the signal and noise separation while keeping information about the strength of the correlation.

In this paper, we present a method to construct weighted and thresholded gene coexpression networks in which the sparsity can be controlled by assigning weighted edges only to those pairs of genes with an expression correlation higher than a given threshold. The weights of the edges correspond to the correlation values. To select the threshold value, we have extended COGENT, including a comparison methodology that allows us to evaluate the similarity between weighted and thresholded networks.

Using this extension, we construct and compare networks constructed from gene expression data using Pearson or signed distance correlation. For ease of comparison and breadth, we analyze the three datasets presented in Pardo-Diaz et al. (2021) as they do not only correspond to different organisms (bacteria, yeast and human) but also are derived from three different experimental techniques: microarrays, RNA-Seq and single-cell RNA-Seq. When comparing the networks, we evaluate the robustness measurement from COGENT and the amount of biological information they capture according to STRING, a protein–protein interaction database with scores for pairs of proteins (Szklarczyk et al., 2019). The higher the STRING score for a protein pair, the more likely the pair is to have a biologically meaningful functional relationship. Using STRING, we show that for our datasets, networks constructed using signed distance correlation capture even more biological information and are structurally more stable than networks based on Pearson correlation.

We also compare the resulting weighted and thresholded networks based on signed distance correlation to the unweighted networks obtained using the same metric as presented in Pardo-Diaz et al. (2021). In this comparison, the weighted and thresholded networks capture more biological information than the corresponding unweighted networks, according to STRING.

While we apply our method to gene expression data, our method to construct networks from signed distance correlations (in combination with COGENT) can be used in applications beyond gene expression and beyond bioinformatics.

Data and source code are available from `https://github.com/javier-pardodiaz/sdcorGCN` and `http://opig.stats.ox.ac.uk/resources`.

2. Methodology

2.1 Datasets and pre-processing

We analyze the three datasets employed in Pardo-Diaz et al. (2021):

- RL3841: A collection of 54 microarrays measuring the expression of 7,077 genes of the bacterium *Rhizobium leguminosarum* bv. *viciae* 3841.

https://doi.org/10.1017/nws.2022.13 Published online by Cambridge University Press

- Yeast: A dataset obtained using RNA-Seq of yeast (*Saccharomyces cerevisiae*) expressing pathways designed to increase ATP or GTP consumption. We use all the raw-counts for experiment E-MTAB-5174 in Expression Atlas (Petryszak et al., 2016) and remove the genes with zero expression variance. The final dataset which we feed into our pipeline includes the expression of 6,930 genes across 209 samples.
- Human liver: A dataset obtained using single-cell RNA-Seq of human liver cells (Klimm et al., 2020). The original dataset measures the expression of 15,353 genes in 1,622 cells.

The three datasets in the form of expression matrices (genes in rows and samples in columns) are publicly available online. We denote the expression matrices by M.

As detailed in Pardo-Diaz et al. (2021), for the RL3841 and yeast datasets, we apply quantile normalization (Bolstad et al., 2003) to the gene expression matrix M. This normalization enables us to compare data from different experiments. To avoid interference from low expression values in the quantile normalization, we ignore the 20% least expressed genes from each sample before the normalization step. After the quantile normalization, we set the ignored values to the lowest expression value in M to decrease the level of noise.

For the human liver dataset, as described in Pardo-Diaz et al. (2021), we follow a different approach due to differences in the data—the gene expression levels in the dataset correspond to the expression of the genes in different cells instead of to different samples—and the organism—while both *R. leguminosarum* and *S. cerevisiae* are unicellular organisms, humans are not. Hence, a considerable proportion of genes are not expressed in the studied cells. For this dataset, we quantile-normalise the data (Bolstad et al., 2003) to make the measurements in the different cells comparable. Afterwards, as in Prieto et al. (2008), we identify the *non-changing genes*. These genes are those for which the difference between its highest and lowest expression value ("expression difference") is lower than the median of all the expression differences calculated for each gene, and for which the mean expression signal between samples is lower than the median of all the expression signals calculated for each gene. After removing the "non-changing genes", we obtain an already quantile-normalised dataset with information for 8,585 genes.

In all three datasets, we denote the pre-processed gene expression matrix by M^*. This matrix has n rows (genes) and m columns (samples).

2.2 Correlation matrices

From each pre-processed gene expression matrix M^*, we compute two correlation matrices S and P, both of them with dimensions $n \times n$. The matrix S contains the signed distance correlation values for the expression of each pair of genes, whereas the matrix P includes the pairwise Pearson correlation values. The matrix S is the result of assigning the sign of the values in the matrix P to the distance correlation values (Székely et al., 2007) between the expression of the genes. Assigning a sign to the (unsigned) distance correlation values allows us to differentiate positive and negative correlation values. The pipeline followed to construct the expression matrices is described in Pardo-Diaz et al. (2021) and depicted in Figure 1.

2.3 Weighted and thresholded network construction

In each of our three datasets, we use the gene correlation matrices S and P to obtain weighted and thresholded networks. These networks contain edges only between those pairs of genes that show a correlation of their expression greater than a given threshold. Unlike in unweighted networks, we assign a different weight to each edge. The weight of the edge corresponds to the value of the correlation of the expression between the two genes. The weighted and thresholded networks obtained using the signed distance correlation matrix S with threshold ϕ have an adjacency matrix $B_S(\phi)$:

https://doi.org/10.1017/nws.2022.13 Published online by Cambridge University Press

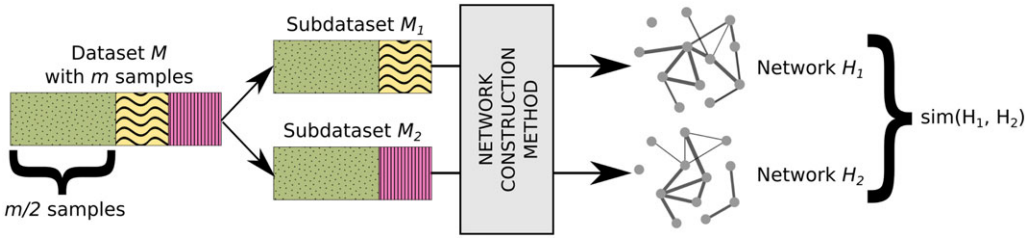


Figure 2. Illustration of a COGENT iteration. At each iteration, from the full dataset M with all m samples we generate two subdatasets M_1 and M_2. Each of the subdatasets contains half of the samples included in M (represented in green with dots in the image). The rest of the samples in M are equally divided between M_1 (yellow with waves) and M_2 (pink with stripes), resulting in the two datasets having the same number of samples. Afterwards, from each dataset, we construct a gene coexpression network (network H_1 from subdataset M_1 and network H_2 from subdataset M_2) using the same network construction method, which might include pre-processing steps. Lastly, we obtain a similarity score based on the overlap between the networks H_1 and H_2.

$$B_S(\phi)_{i,j} = \begin{cases} S_{i,j} & \text{if } S_{i,j} \geq \phi \text{ and } i \neq j, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

We construct the weighted and thresholded networks from the Pearson correlation matrix P in a similar way.

We use an extension of COGENT (Bozhilova et al., 2020) to find a threshold value ϕ^* that results in a network with optimal self-consistency. The main idea behind COGENT is that the more similar two networks obtained using the same network construction method and from different overlapping subsets of the same dataset are, the more self-consistent the employed method is: despite the changes in the dataset, the topology of the resulting networks is similar. Heuristically, the networks obtained using a self-consistent construction method are self-consistent themselves since small changes in the datasets will not affect their structure. Our contribution to COGENT relies on the addition of a method that allows assessing the similarity between weighted and thresholded networks (Equation (2)).

As in Pardo-Diaz et al. (2021), we run 25 COGENT iterations in which the samples (columns) in M are grouped into two overlapping sets M_1 and M_2. Both sets share half of the total number of samples, and differ in 1/3 from each other. At each iteration, we generate 25 pairs of correlation matrices S_1 and S_2 and test different threshold values. For each iteration and threshold value, the correlation matrices S_1 and S_2 are thresholded (Equation (1)) and turned into networks H_1 and H_2. The processes taking place in each of the COGENT iterations to obtain the networks H_1 and H_2 are depicted in Figure 2. To calculate a similarity between the two networks H_1 and H_2, we use an adjusted weighted Jaccard index that permits the comparison between networks with different edge densities:

$$\text{sim}(H_1, H_2) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\min(H_1(i,j), H_2(i,j)) - \gamma}{\sum_{i=1}^{n}\sum_{j=1}^{n}\max(H_1(i,j), H_2(i,j)) + \gamma} \tag{2}$$

where

$$\gamma = \frac{1}{2}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\min(H_1(i,j), R_2(i,j)) + \sum_{i=1}^{n}\sum_{j=1}^{n}\min(H_2(i,j), R_1(i,j))\right), \tag{3}$$

R_1 and R_2 are randomizations of the adjacency matrices H_1 and H_2, respectively, and n is the number of nodes in the network. To generate the randomizations, we first permute the rows in the correlation matrices and then shuffle the columns using the same permutation. The networks R_1 and R_2 have the same topological properties as H_1 and H_2 and allow us to estimate the expected

https://doi.org/10.1017/nws.2022.13 Published online by Cambridge University Press

similarity between the analysed networks and random networks with the same edge weight distribution. Without this correction, the networks derived from low threshold values would have the highest similarity values.

The similarity score associated with each threshold $s(\phi)$ is the average of the similarity of 25 pairs of networks (Equation (2)). Then, to favor signal over noise, we prioritise those networks with a low sum of edge weights and obtain the Score(ϕ) value associated to each threshold ϕ:

$$\text{Score}(\phi) = s(\phi) - \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} B_S(\phi)_{i,j}}{n(n-1)} \tag{4}$$

where n denotes the number of nodes in the network (i.e. the number of genes). After computing the scores for all tested thresholds, we select the threshold value ϕ^* that results in the highest score Score(ϕ^*). We denote the optimal weighted and thresholded network obtained using signed distance correlation as $WS(w_S)$. We denote the optimal sum of edge weights for the signed correlation by w_S.

Following the same pipeline, we select the unique threshold value $\phi^\star$ that results in the optimal network $WP(w_P)$ (sum of edge weights = w_P) obtained using the Pearson correlation matrix P.

2.4 STRING evaluation of weighted and thresholded networks

We assess the amount of biological information contained in the just generated weighted and thresholded networks $WS(w_S)$ and $WP(w_P)$ using STRING, a database of known and predicted protein—protein interactions (Szklarczyk et al., 2019). STRING collects information from numerous sources, including experimental data, computational predictions, and textmining. The association evidence in STRING is categorized into independent channels, weighted, and integrated to produce a confidence score C for all recorded protein interactions. Interactions with a high C score are more likely to be true than those with a low score. To evaluate the networks, we employ the same set of confidence scores as in Pardo-Diaz et al. (2021):

- C: Total scores provided by STRING
- $C^\dagger$: Scores that *only* consider coexpression information
- $C^\ddagger$: Scores that *exclude* coexpression information.

We expect the overlap of our networks with the scores in $C^\dagger$ to be higher than the overlap with the scores in $C^\ddagger$ since our input datasets contain gene expression information. Nevertheless, both $C^\dagger$ and $C^\ddagger$ are interesting to analyse: the first one might be an indicator of how well the networks capture coexpression relationships whereas the latter can be used as an indicator of how well coexpression can predict other types of relationships.

For each network and confidence score set, we compute the dot product of the edge weights and the STRING confidence score. Afterwards, we divide the result by the sum of the weights of all edges in the network. Equation (5) indicates how we obtain the STRING score for a weighted network with adjacency matrix $B_S(\phi)$ and confidence score set C:

$$\text{STRING Score}(B_S(\phi), C) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} B_S(\phi)_{i,j}\, C_{i,j}}{\sum_{i=1}^{n}\sum_{j=1}^{n} B_S(\phi)_{i,j}} \tag{5}$$

where n denotes the number of nodes in the network, $B_S(\phi)_{i,j}$ is the the entry for genes i and j in $B_S(\phi)$, and $C_{i,j}$ refers to the confidence score for the genes i and j in the confidence score set C.

Despite being normalised for the sum of edge weights, the score might be biased towards networks with fewer and heavier edges. For this reason, and in order to keep the comparison between

https://doi.org/10.1017/nws.2022.13 Published online by Cambridge University Press

the networks obtained using Pearson and distance correlation the fairest possible, we construct and evaluate six additional networks:

- $WS(w_P)$: Network from S with the same sum of weights of edges as $WP(w_P)$
- $WS(e_P)$: Network from S with the same number of edges as $WP(w_P)$
- $WS(a_P)$: Network from S with the same average edge weight as $WP(w_P)$
- $WP(w_S)$: Network from P with the same sum of weights of edges as $WS(w_S)$
- $WP(e_S)$: Network from P with the same number of edges as $WS(w_S)$
- $WP(a_S)$: Network from P with the same average edge weights as $WS(w_S)$.

We also generate two sets of 30 random networks to compare the results obtained in the STRING evaluation with those expected by chance. The sets of random networks have the same number of edges and distribution of edge weights as $WS(w_S)$ and $WP(w_P)$, respectively. We generate the networks by assigning the edges of the original networks to randomly picked pairs of vertices as a Bernoulli graph with a fixed number of edges and weights, and compute their STRING scores.

2.5 Comparison of weighted and thresholded networks and unweighted networks

Next, we assess whether the weighted and thresholded gene coexpression networks we obtain capture more biological information than unweighted coexpression networks. We compare our networks with the optimal unweighted signed distance correlation networks from Pardo-Diaz et al. (2021), denoted by $NS(d_S)$. We analyse and compare four networks:

- $WS(w_S)$: Optimal weighted network from S
- $NS(w_S)$: Unweighted network with the same edges as $WS(w_S)$
- $NS(d_S)$: Optimal unweighted network from S
- $WS(d_S)$: Weighted network with the same edges as $NS(d_S)$.

We focus only on signed distance correlation networks because unweighted and weighted signed distance correlation networks are both more self-consistent and capture more biological information than their matching Pearson networks (see Results section).

Unweighted networks, by definition, do not have weights associated with their edges; in order to be able to compare the networks using the scoring function from Equation (5), we assign to edges of the networks $NS(w_S)$ and $NS(d_S)$ as weight the average edge weight of the networks $WS(w_S)$ and $WS(d_S)$, respectively. All the edges in an unweighted network thus have the same weight.

3. Results

3.1 Network construction and COGENT evaluation

We evaluate how the score retrieved using COGENT changes across different thresholds in each dataset and for each correlation matrix. This score depends on the similarity of the networks constructed at each COGENT iteration. The similarity is adjusted using the overlap expected between each of the networks and a network with the same edge weights distributed randomly, and then adjusted to prioritise sparser networks (see Equations (2) and (4)). Figure 3 illustrates how the signed distance and Pearson scores change for different edge weight values in the three datasets. In all the datasets (RL3841, yeast, and human liver), the curve shows a similar shape and there is an edge weight value for which the scores reach their maxima. The thresholds associated with those edge weights are the optimal thresholds ϕ^* and $\phi^\star$ which we choose when constructing the networks $WS(w_S)$ and $WP(w_P)$. The score, threshold, and summaries of these networks are shown in Table 1. In all cases, the highest scores obtained using signed distance correlation are

https://doi.org/10.1017/nws.2022.13 Published online by Cambridge University Press

Table 1. Metrics and COGENT score for the optimal networks of the studied datasets. "Corr" indicates the correlation matrix employed and "Thr" denotes the optimal threshold

RL3841 dataset						
Network	Corr	COGENT score	Thr	Number of Edges	Sum of weights	Average weight
$WS(w_S)$	S	0.46	0.60	388,374	263,662	0.6789
$WP(w_P)$	P	0.43	0.55	531,374	342,785	0.6451
Yeast dataset						
Network	Corr	COGENT score	Thr	Number of Edges	Sum of weights	Average weight
$WS(w_S)$	S	0.76	0.82	405,576	355,227	0.8759
$WP(w_P)$	P	0.73	0.81	350,549	303,310	0.8652
Human liver dataset						
Network	Corr	COGENT score	Thr	Number of Edges	Sum of weights	Average weight
$WS(w_S)$	S	0.88	0.44	3,320	1734.58	0.5225
$WP(w_P)$	P	0.77	0.44	3,047	1567.93	0.5146

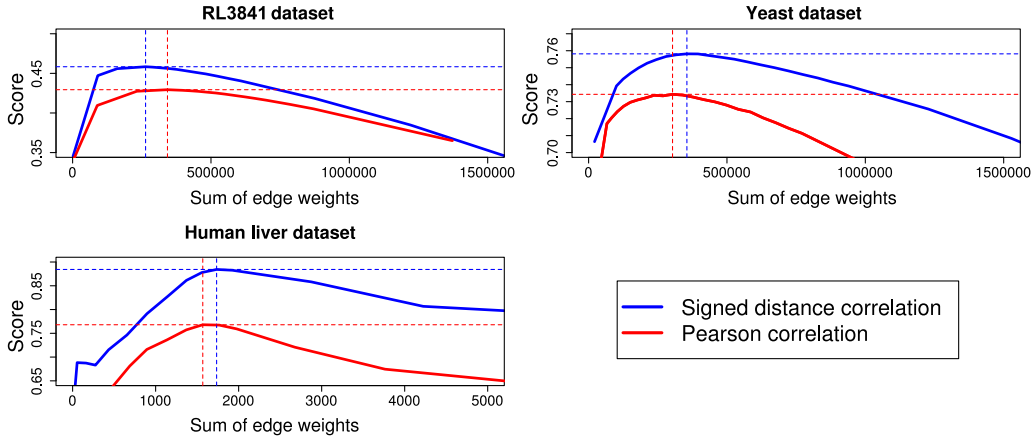


Figure 3. Self-consistency scores for different thresholds and datasets. The blue and red lines show the scores (Equation (4)) of networks obtained using signed distance and Pearson correlations, respectively, to generate weighted and thresholded networks. The dashed vertical and horizontal lines indicate the optimal sums of edge weights and the highest scores, respectively. Scores obtained by signed distance correlation are almost uniformly higher than scores from Pearson correlation.

higher than those for Pearson correlation. Therefore, weighted and thresholded gene coexpression networks based on signed distance correlation can be more self-consistent than those based on Pearson correlation.

3.2 STRING evaluation

We assess the amount of biological information contained in the networks following the methodology described in the Methods section. Tables 2–4 present the networks constructed using the different datasets, their metrics, and their STRING scores. The optimal signed distance correlation networks $WS(w_S)$ retrieve a higher score than all alternatives based on Pearson correlation for all the studied datasets. The optimal Pearson correlation networks $WP(w_P)$ retrieve a lower score

https://doi.org/10.1017/nws.2022.13 Published online by Cambridge University Press

Table 2. Metrics and STRING scores for the networks from the RL3841 dataset. RE indicates the expected result and its standard deviation. C, $C^{\dagger}$, and $C^{\ddagger}$ indicate the scores obtained for each of the different sets of confidence from STRING. For each set of networks, the highest score values are indicated in bold

Network	Corr	Number of edges	Sum of weights	Average weight	C	$C^{\dagger}$	$C^{\ddagger}$
$WS(w_S)$	S	388,374	263,662	0.6789	**28.66**	**8.70**	**25.72**
$WP(e_S)$	P	388,374	261,723	0.6739	24.49	7.31	22.08
$WP(w_S)$	P	391,691	263,662	0.6731	24.43	7.28	22.03
$WP(a_S)$	P	367,002	249,153	0.6789	24.86	7.52	22.39
RE w_S	S	388,374	263,662	0.6789	7.51 $\pm$ 0.08	0.90 $\pm$ 0.02	7.11 $\pm$ 0.08
$WP(w_P)$	P	531,374	342,785	0.6451	22.39	6.26	20.26
$WS(e_P)$	S	531,374	347,174	0.6534	25.64	7.27	23.12
$WS(w_P)$	S	523,675	342,785	0.6546	**25.78**	**7.33**	**23.23**
$WS(a_P)$	S	585,747	377,861	0.6451	24.78	6.88	22.37
RE w_P	P	531,374	342,785	0.6451	7.53 $\pm$ 0.06	0.91 $\pm$ 0.02	7.12 $\pm$ 0.05

Table 3. Metrics and STRING scores for the networks from the Yeast dataset. RE indicates the expected result and its standard deviation. C, $C^{\dagger}$, and $C^{\ddagger}$ indicate the scores obtained for each of the different sets of confidence from STRING. For each set of networks, the highest score values are indicated in bold

Network	Corr	Number of edges	Sum of weights	Average weight	C	$C^{\dagger}$	$C^{\ddagger}$
$WS(w_S)$	S	405,576	355,227	0.8759	106.40	75.61	62.25
$WP(e_S)$	P	405,576	347,541	0.8569	105.69	75.60	61.58
$WP(w_S)$	P	415,223	355,227	0.8556	104.48	74.52	61.06
$WP(a_S)$	P	286,663	251,077	0.8759	**124.81**	**92.58**	**69.84**
RE w_S	S	405,576	355,227	0.8759	13.94 $\pm$ 0.14	4.77 $\pm$ 0.07	11.83 $\pm$ 0.13
$WP(w_P)$	P	350,549	303,310	0.8652	113.36	86.36	64.90
$WS(e_P)$	S	350,549	309,784	0.8837	114.12	82.30	65.67
$WS(w_P)$	S	342,774	303,310	0.8849	**115.34**	**83.36**	**66.21**
$WS(a_P)$	S	486,198	420,679	0.8652	97.45	67.98	58.16
RE w_P	P	350,549	303,310	0.8652	13.96 $\pm$ 0.13	4.77 $\pm$ 0.06	11.84 $\pm$ 0.13

than most of their (signed distance) competitors. The only exception to this trend is the yeast network $WS(a_P)$ (signed distance correlation network with the average edge weight of the optimal Pearson correlation network) which returns a lower score than the network $WP(w_P)$ (optimal Pearson correlation network). These results suggest that using signed distance correlation for generating weighted and thresholded networks from gene expression data can yield better results than using Pearson correlation. Figure 4 shows the STRING scores obtained by the networks with the sum of edge weights w_S and w_P for the three datasets and the different sets of evidence.

We observe that for both correlation matrices, a higher average edge weight (and a lower number of edges) implies a higher STRING score. Figure 5 shows the high positive correlation between

https://doi.org/10.1017/nws.2022.13 Published online by Cambridge University Press

Table 4. Metrics and STRING scores for the networks from the Human liver dataset. RE indicates the expected result and its standard deviation. C, $C^{\dagger}$, and $C^{\ddagger}$ indicate the scores obtained for each of the different sets of confidence from STRING. For each set of networks, the highest score values are indicated in bold

Network	Corr	Number of edges	Sum of weights	Average weight	C	$C^{\dagger}$	$C^{\ddagger}$
$WS(w_S)$	S	3,320	1,734.58	0.5225	**214.99**	**188.10**	**206.18**
$WP(e_S)$	P	3,320	1,686.44	0.5080	199.80	173.09	191.20
$WP(w_S)$	P	3,433	1,734.56	0.5053	196.29	169.61	187.86
$WP(a_S)$	P	2,728	142.28	0.5225	210.24	184.12	201.34
RE w_S	S	3,320	1,734.58	0.5225	6.90 $\pm$ 1.01	2.00 $\pm$ 0.29	6.06 $\pm$ 0.10
$WP(w_P)$	P	3,047	1,567.93	0.5146	202.26	176.67	193.58
$WS(e_P)$	S	3,047	1,612.62	0.5292	219.16	192.66	210.57
$WS(w_P)$	S	2,949	1,567.93	0.5317	**220.92**	**194.64**	**212.32**
$WS(a_P)$	S	3,639	1,872.58	0.5146	210.79	182.78	201.78
RE w_P	P	3,047	1,567.93	0.5146	6.80 $\pm$ 1.00	2.00 $\pm$ 0.39	5.95 $\pm$ 0.92

the STRING score using the information set C and the average edge weight of the networks for the RL3841 dataset. Still, often signed distance correlation networks retrieve a higher STRING score than Pearson networks even when their average edge weight is lower. For example, for the RL3841 dataset, the only case in which a Pearson correlation network retrieves a higher STRING score than a signed distance correlation network is when comparing the network $WS(a_P)$ (signed distance correlation network with the average edge weight of the optimal Pearson correlation network, with score 24.78) versus the network $WP(a_S)$ (Pearson correlation network with the average edge weight of the optimal signed distance correlation network, with score 24.86). However, this comparison lacks interest since their average edge weights (0.6789 and 0.6451) are quite different.

The random networks obtain a lower STRING score than Pearson and signed distance correlation networks (see Tables 2–4 and Figure 4). We find the highest difference between the random expected STRING scores and those retrieved by the networks when using only coexpression information ($C^{\dagger}$). The scores obtained by the networks $WS(w_S)$ and $WP(w_P)$ are 9.63 and 6.89 times higher than the score expected by chance in the case of the RL3841 dataset. The smallest differences correspond to the exclusion of the coexpression information ($C^{\ddagger}$), where the values are 3.62 ($WS(w_S)$) and 2.84 ($WP(w_P)$) times higher than the randomly expected values for the RL3841 dataset.

3.3 Comparison with unweighted networks

Next, we compare the amount of biological information captured by weighted and thresholded networks, and unweighted networks, constructed using signed distance correlation. We compare the two groups of networks with the same sets of edges as described in the Methods section: the optimal set of edges for the weighted and thresholded network and the optimal set for the unweighted network in Pardo-Diaz et al. (2021). The weights in the unweighted networks are set to the average weight in the matching weighted and thresholded networks. The results for the STRING evaluation for the different datasets are shown in Table 5. For the three datasets, the weighted and thresholded networks can capture more of the biological information and should therefore be preferred. For the human liver dataset, we present only a set of two networks since the optimal threshold for the weighted and the unweighted networks is the same and therefore they have the same edges.

https://doi.org/10.1017/nws.2022.13 Published online by Cambridge University Press

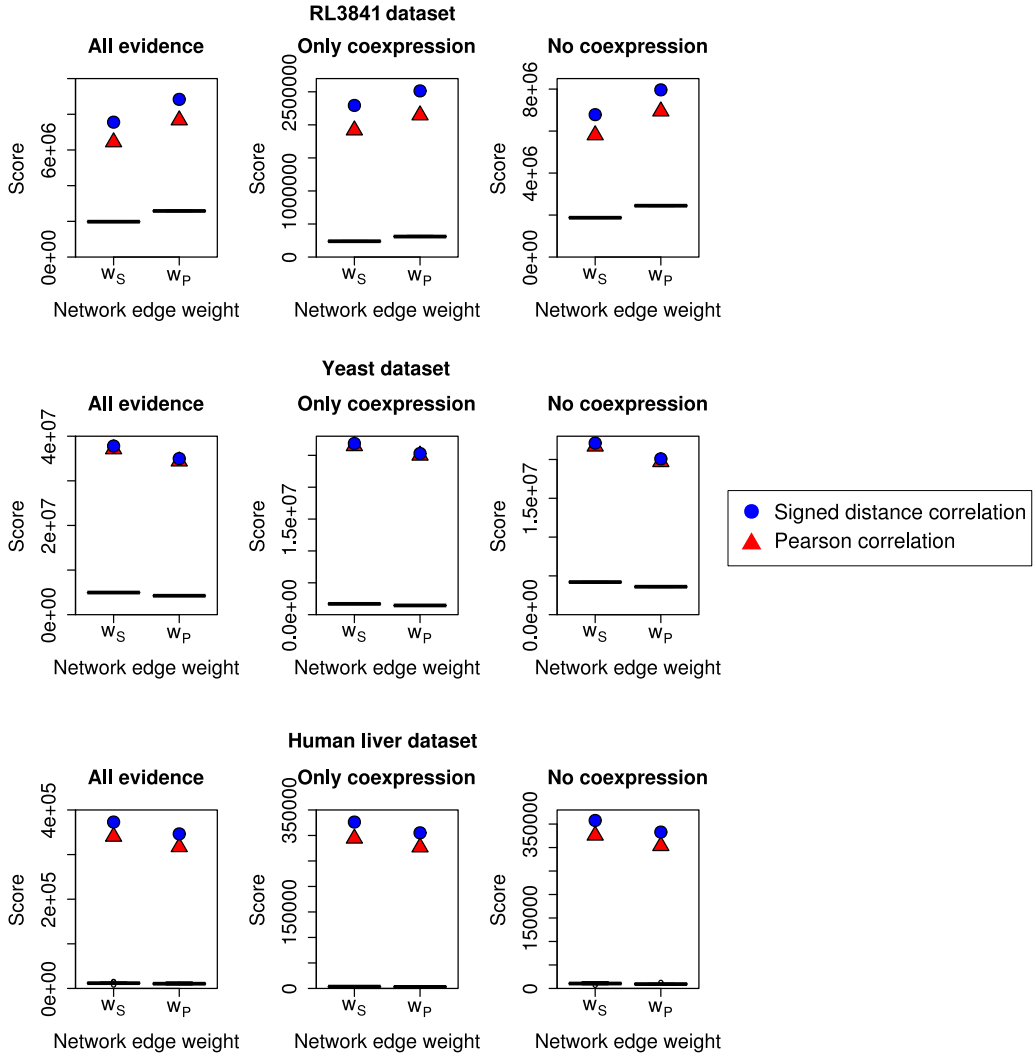


Figure 4. Scores obtained for the gene coexpression networks obtained from the three datasets using STRING. All panels show the score for the different networks on the y-axis, and the sum of weights on the *x*-axis. The scores are the result of the dot product of the weights of the edges in the networks and the confidence scores with all evidence (C), only coexpression evidence ($C^{\dagger}$) and everything excluding coexpression ($C^{\ddagger}$) from STRING associated with them, divided by the sum of edge weights. The black box plots correspond to the scores obtained by 30 random networks. Blue circles and red triangles represent signed distance correlation and Pearson correlation, respectively.

4. Discussion and conclusions

The method we propose combines the intuitiveness of the unweighted networks, as there are edges connecting only those pairs of genes with a high correlation in their expression, with the fine-tuning provided by assigning different weight values to different edges. This combination allows differentiating genes which are highly coexpressed from genes that, even if they are coexpressed, are not highly associated. These characteristics can help to study and analyze the networks.

We select the optimal threshold value to generate the weighted and thresholded networks using an extension of COGENT (Bozhilova et al., 2020) that allows assessing the similarity of this type of networks. For each correlation matrix, we select the threshold value that results in a high self-consistency and prioritizes signal over noise. We also normalized the correlation values that are

https://doi.org/10.1017/nws.2022.13 Published online by Cambridge University Press

Table 5. STRING scores for unweighted and weighted and thresholded networks for the three datasets using signed distance correlation. C, $C^{\dagger}$, and $C^{\ddagger}$ indicate the scores obtained for each of the different sets of confidence from STRING. Types "W" and "U" denote whether the networks are weighted or unweighted, respectively. Each pair of networks have the same edges; for each pair of networks, the highest score values are indicated in bold

RL3841 dataset							
Network	Type	Number of edges	Sum of weights	Average weight	C	$C^{\dagger}$	$C^{\ddagger}$
$WS(w_S)$	W	388,374	263,662	0.6789	**28.66**	**8.70**	**25.72**
$NS(w_S)$	U	388,374	263,662	0.6789	27.31	7.87	24.57
$WS(d_S)$	W	313,348	217,914	0.6954	**30.97**	**9.88**	**27.70**
$NS(d_S)$	U	313,348	217,914	0.6954	29.61	9.02	26.55
Yeast dataset							
Network	Type	Number of edges	Sum of weights	Average weight	C	$C^{\dagger}$	$C^{\ddagger}$
$WS(w_S)$	W	405,576	355,227	0.8759	**106.40**	**75.61**	**62.25**
$NS(w_S)$	U	405,576	355,227	0.8759	103.49	72.96	61.12
$WS(d_S)$	W	314,771	279,879	0.8892	**120.12**	**87.59**	**68.19**
$NS(d_S)$	U	314,771	279,879	0.8892	117.36	85.03	67.14
Human liver dataset							
Network	Type	Number of edges	Sum of weights	Average weight	C	$C^{\dagger}$	$C^{\ddagger}$
$WS(w_S)$	W	3,320	1,734.58	0.5225	**214.99**	**188.10**	**206.18**
$NS(d_S)$	U	3,320	1,734.58	0.5225	212.02	185.35	203.55

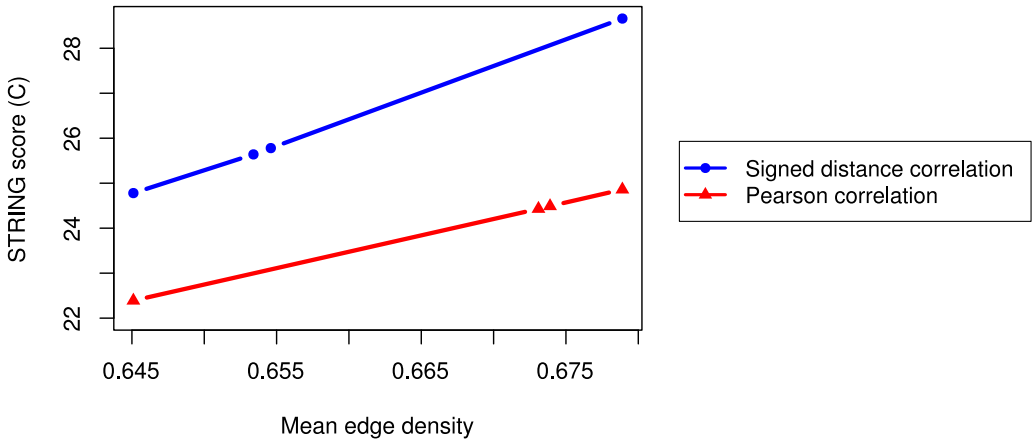


Figure 5. Distribution of STRING scores using the set of information C and average edge weight of the RL3841 networks. Blue and circles are for signed distance correlation; red and triangles are for Pearson correlation.

higher than the threshold to lie between zero and one, using these normalized values as, as follows: edges connecting genes with an expression correlation less or equal to the threshold had a weight of zero; edges connecting genes with an expression correlation equal to one had a weight of one; edges connecting genes with a expression correlation between the threshold and one had weights between zero and one. We evaluated the self-consistency of the obtained network as described in Section 2.3. We observed that this normalization step resulted in a similar self-consistency.

https://doi.org/10.1017/nws.2022.13 Published online by Cambridge University Press

For the signed distance correlation network, the optimal threshold was 0.52, resulting in a score of 0.4582. This value is lower than the one obtained without the normalization: 0.4584. For the Pearson correlation networks, we also obtained higher values when not using the normalization. In light of these results, we do not include this normalization step in our pipeline.

The correlation matrices S and P differentiate positive from negative correlations. However, the threshold value we select to construct the networks is always greater than zero and therefore the weights of the edges in the network are always positive. As discussed in Pardo-Diaz et al. (2021), we construct networks that include only positive correlations since negative correlations do not always imply a functional relationship. Nevertheless, a signed network with weights associated with its edges might include valuable information since the sign of the weights allow to differentiate positive and negative associations. Exploring a modification of our network construction pipeline of this type might provide an improvement.

For the three studied datasets, the use of signed distance correlation to generate weighted and thresholded networks results in a higher self-consistency than the use of Pearson correlation. Regarding the STRING evaluation, most times, the signed distance correlation networks also obtain a higher score than their competitors constructed using Pearson correlation. The only exception to this trend is when comparing networks with the same average edge weight in the yeast dataset. However, overall, the presented results suggest that the networks constructed using signed distance correlation capture more biological information.

As shown in Table 5, our weighted and thresholded networks capture more biological information than unweighted networks. We extract the same conclusion from the analysis of the three datasets. This result is in line with what we expect since the use of edge weights representing the strength of the correlation between the expression of the pairs of genes results in an increase in the amount of information contained in the network.

The threshold values that we select to generate the weighted and thresholded networks from the three datasets are very similar to those that we use to construct the unweighted networks. This fact suggests that independently of the construction methodology we follow, the barrier between signal and noise is in the same range of correlation values for a given correlation matrix.

We use our method to construct a novel weighted and thresholded gene coexpression network for *R. leguminosarum*. This network promises to reveal rich biological information and it is, therefore, the starting point for further investigations of the biological mechanisms of this organism. In particular, we plan to identify groups of genes in *R. leguminosarum* which are highly connected in the network, and associate them with specific biological processes. To do so, we will make use of community detection techniques and new experimental data. For the human liver dataset one could similarly validate communities by exploring predicting disease-related biological information; see for example Song et al. (2019), Chen et al. (2018), Li et al. (2019).

Finally, we have showcased our method on gene expression datasets from different organisms obtained using different techniques: microarrays (*R. leguminosarum*), RNA-Seq (yeast), and single-cell RNA-Seq (human). However, the methods that we have developed are general, and can also be used to construct networks in a vast range of domains, such as, for example, economics (Wang et al., 2018), neuroscience (Bernhardt et al., 2011), climatology (Donges et al., 2009), and indeed any discipline where networks are constructed from correlation data.

Acknowledgments. We thank Alison K East and Florian Klimm for their help and advice on how to analyze the datasets used in this manuscript. We thank Katharine Turner (ANU), Lyuba V. Bozhilova, Florian Klimm, and Malte D Luecken for fruitful discussions. In addition, we acknowledge support from COST Action CA15109, Keble College Oxford and the Keble Association, as well as the use of the University of Oxford Advanced Research Computing (ARC) facility in carrying out this work (http://dx.doi.org/10.5281/zenodo.22558).

Funding. This work is supported by the Engineering and Physical Sciences Research Council (EPSRC) [EP/R512333/1 to JPD, MBD, PSP, CMD, and GR, EP/T018445/1 to GR], the Biotechnology and Biological Sciences Research Council (BBSRC)

https://doi.org/10.1017/nws.2022.13 Published online by Cambridge University Press

[BB/T001801/1 to PSP and GR], the COSTNET COST Action [CA15109 to GR], and e-Therapeutics plc. MBD acknowledges support from the Oxford-Emirates Data Science Lab.

Competing interests. None.

References

Bar-Joseph, Z., Gerber, G. K., Lee, T. I., Rinaldi, N. J., Yoo, J. Y., Robert, F., ... Young, R. A. (2003). Computational discovery of gene modules and regulatory networks. *Nature Biotechnology*, *21*(11), 1337–1342.

Bernhardt, B. C., Chen, Z., He, Y., Evans, A. C., & Bernasconi, N. (2011). Graph-theoretical analysis reveals disrupted small-world organization of cortical thickness correlation networks in temporal lobe epilepsy. *Cerebral Cortex*, *21*(9), 2147–2157.

Bolstad, B. M., Irizarry, R. A., Åstrand, M., & Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, *19*(2), 185–193.

Bozhilova, L. V., Pardo-Diaz, J., Reinert, G., & Deane, C. M. (2020). COGENT: Evaluating the consistency of gene co-expression networks. *Bioinformatics*, 09. btaa787.

Chen, X., Yin, J., Qu, J., & Huang, L. (2018). MDHGI: Matrix decomposition and heterogeneous graph inference for miRNA-disease association prediction. *PLoS Computational Biology*, *14*(8), e1006418.

Donges, J. F., Zou, Y., Marwan, N., & Kurths, J. (2009). Complex networks in climate dynamics. *The European Physical Journal Special Topics*, *174*(1), 157–179.

George, G., Singh, S., Lokappa, S. B., & Varkey, J. (2019). Gene co-expression network analysis for identifying genetic markers in Parkinson's disease-a three-way comparative approach. *Genomics*, *111*(4), 819–830.

Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., ... Friend, S. H. (2000). Functional discovery via a compendium of expression profiles. *Cell*, *102*(1), 109–126.

Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y., & Barkai, N. (2002). Revealing modular organization in the yeast transcriptional network. *Nature Genetics*, *31*(4), 370–377.

Klimm, F., Toledo, E. M., Monfeuga, T., Zhang, F., Deane, C. M., & Reinert, G. (2020). Functional module detection through integration of single-cell RNA sequencing data with protein–protein interaction networks. *BMC Genomics*, *21*(1), 1–10.

Kothapalli, R., Yoder, S. J., Mane, S., & Loughran, T. P. (2002). Microarray results: How accurate are they? *BMC Bioinformatics*, *3*(1), 22.

Langfelder, P., & Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics*, *9*(1), 559.

Lee, H. K., Hsu, A. K., Sajdak, J., Qin, J., & Pavlidis, P. (2004). Coexpression analysis of human genes across many microarray data sets. *Genome Research*, *14*(6), 1085–1094.

Li, H., Wang, Y., Jiang, J., Zhao, H., Feng, X., Zhao, B., & Wang, L. (2019). A novel human microbe-disease association prediction method based on the bidirectional weighted network. *Frontiers in Microbiology*, *10*, 676.

Magwene, P. M., & Kim, J. (2004). Estimating genomic coexpression networks using first-order conditional independence. *Genome Biology*, *5*(12), R100.

Makrodimitris, S., Reinders, M. J. T., & van Ham, R. C. H. J. (2020). Metric learning on expression data for gene function prediction. *Bioinformatics*, *36*(4), 1182–1190.

Pardo-Diaz, J., Bozhilova, L. V., Beguerisse-Daz, M., Poole, P. S., Deane, C. M., & Reinert, G. (2021). Robust gene coexpression networks using signed distance correlation. *Bioinformatics*, 02. btab041.

Petryszak, R., Keays, M., Tang, Y. A., Fonseca, N. A., Barrera, E., Burdett, T., ... Brazma, A. (2016). Expression atlas update–an integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Research*, *44*(D1), D746–D752.

Prieto, C., Risueño, A., Fontanillo, C., & De Las Rivas, J. (2008). Human gene coexpression landscape: Confident network derived from tissue transcriptomic profiles. *PloS One*, *3*(12), e3911.

Song, F., Cui, C., Gao, L., & Cui, Q. (2019). miES: Predicting the essentiality of miRNAs with machine learning and sequence features. *Bioinformatics*, *35*(6), 1053–1054.

Stuart, J. M., Segal, E., Koller, D., & Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science*, *302*(5643), 249–255.

Székely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, *35*(6), 2769–2794.

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., ... von Mering, C. (2019). STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, *47*(D1), D607–D613.

Ucar, D., Neuhaus, I., Ross-MacDonald, P., Tilford, C., Parthasarathy, S., Siemers, N., & Ji, R.-R. (2007). Construction of a reference gene association network from multiple profiling data: application to data analysis. *Bioinformatics*, *23*(20), 2716–2724.

https://doi.org/10.1017/nws.2022.13 Published online by Cambridge University Press

van Noort, V., Snel, B., & Huynen, M. A. (2003). Predicting gene function by conserved co-expression. *TRENDS in Genetics*, *19*(5), 238–242.

Wang, G.-J., Xie, C., & Stanley, H. E. (2018). Correlation structure and evolution of world stock markets: Evidence from Pearson and partial correlation-based networks. *Computational Economics*, *51*(3), 607–635.

Weirauch, M. T. (2011). Gene coexpression networks for the analysis of DNA microarray data. *Applied Statistics for Network Biology: Methods in Systems Biology*, *1*, 215–250.

Wolfe, C. J., Kohane, I. S., & Butte, A. J. (2005). Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. *BMC Bioinformatics*, *6*(1), 227.

A preliminary version of this paper was presented at the Ninth International Conference on Complex Networks and their Applications (COMPLEX NETWORKS 2020).

Cite this article: Pardo-Diaz J., Poole P.S., Beguerisse-Díaz M., Deane C.M. and Reinert G. (2022). Generating weighted and thresholded gene coexpression networks using signed distance correlation. *Network Science* **10**, 131–145. https://doi.org/10.1017/nws.2022.13

https://doi.org/10.1017/nws.2022.13 Published online by Cambridge University Press

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

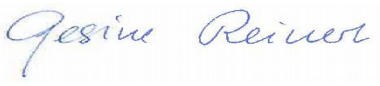
Title of Paper	Generating weighted and thresholded gene coexpression networks using signed distance correlation
Publication Status	Published
Publication Details	Pardo-Diaz, J., Poole, P., Beguerisse-Díaz, M., Deane, C., & Reinert, G. (2022). Generating weighted and thresholded gene coexpression networks using signed distance correlation. *Network Science, 10*(2), 131-145. doi:10.1017/nws.2022.13

Student Confirmation

Student Name:	Javier Pardo Díaz (Javier Pardo-Diaz)		
Contribution to the Paper	Conceptualisation (lead), data curation (lead), formal analysis, investigation, methodology, software, validation, visualization, writing—original draft, and writing—review and editing (equal).		
Signature		Date	19th September 2022

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Professor Gesine Reinert		
Supervisor comments This is an accurate description.		
Signature 	Date	September 20, 2022

This completed form should be included in the thesis, at the end of the relevant chapter.

4

Analysis and evaluation of gene coexpression networks for *Rhizobium leguminosarum*

The Supplementary Information for the publication integrated into this chapter is available in Appendix D.

The colour version of Figures 4, 5, and 6 are available after the paper. For easier readability, we recommend looking at those figures instead of the black and white figures included in the paper.

JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 29, Number 7, 2022
© Mary Ann Liebert, Inc.
Pp. 752–768
DOI: 10.1089/cmb.2021.0600

Open camera or QR reader and scan code to access this article and other resources online.

Extracting Information from Gene Coexpression Networks of *Rhizobium leguminosarum*

JAVIER PARDO-DIAZ,[1,2] MARIANO BEGUERISSE-DÍAZ,[3] PHILIP S. POOLE,[2] CHARLOTTE M. DEANE,[1] and GESINE REINERT[1]

Downloaded by 192.76.8.89 from www.liebertpub.com at 07/28/22. For personal use only.

ABSTRACT

Nitrogen uptake in legumes is facilitated by bacteria such as *Rhizobium leguminosarum*. For this bacterium, gene expression data are available, but functional gene annotation is less well developed than for other model organisms. More annotations could lead to a better understanding of the pathways for growth, plant colonization, and nitrogen fixation in *R. leguminosarum*. In this study, we present a pipeline that combines novel scores from gene coexpression network analysis in a principled way to identify the genes that are associated with certain growth conditions or highly coexpressed with a predefined set of genes of interest. This association may lead to putative functional annotation or to a prioritized list of genes for further study.

Keywords: functional annotation, gene co-expression network analysis, *Rhizobium leguminosarum*.

1. INTRODUCTION

Rhizobium leguminosarum is a bacterium that fixes atmospheric nitrogen when associated with legumes (e.g., peas, beans, lentils). *R. leguminosarum* transforms molecular nitrogen in the air into ammonia, which can be assimilated by plants. Nitrogen fixation improves the growth of plants, as nitrogen is one of the limiting factors during the growth process (Gutiérrez, 2012). Bacteria such as *R. leguminosarum* grow in the soil in the absence of their legume host (Downie, 2010; *free-living conditions*) and when sensing the plant, they approach its root hairs and attach. The bacteria are eventually released into the developing nodule, where they differentiate to a nitrogen-fixing state called *bacteroid* (Downie, 2010). Thus, a better understanding of the metabolic and regulatory pathways in *R. leguminosarum* for growth, plant colonization, and nitrogen fixation may lead to improved farming methods of legumes; see also Waterman (2016).

Currently, functional gene annotation in *R. leguminosarum* is far from complete: We know the precise function for only around 25% of the genes in the genome, and we have no functional annotation at all for another 25%; the remaining genes have limited functional annotation. To facilitate the identification of

Departments of [1]Statistics, and [2]Plant Sciences, University of Oxford, Oxford, United Kingdom.
[3]Department of the Mathematical Institute, University of Oxford, Oxford, United Kingdom.

functional groups of genes, based on gene expression data we aim at finding *R. leguminosarum* genes (*new candidate genes*) that are highly connected to a pre-selected group of genes (*seed list*), or to genes that have a high expression under a particular experimental condition. Our objectives are to assign the new candidate genes a putative molecular function and to unravel which signals trigger their expression.

As a main tool we use gene coexpression network analysis. Gene coexpression can be assessed via correlation measures of the expression across multiple samples; see, for example, Wang et al. (2014) and Pardo-Diaz et al. (2021b). Gene coexpression networks are networks where nodes are genes and edges indicate (high) coexpression of these genes (Lee et al., 2004). Representing gene coexpression as networks eases the study and visualization of the expression data (Magwene and Kim, 2004; Weirauch, 2011) and helps exploit the structure of gene interactions at a whole-system level.

One motivation behind creating these networks is that genes that are highly coexpressed across multiple samples are likely to have related functions (Hughes et al., 2000; Stuart et al., 2003; van Noort et al., 2003; Makrodimitris et al., 2020), allowing the inference of gene function using *guilt by association* approaches (Wolfe et al., 2005), in particular if the studied organism is poorly annotated.

Gene coexpression networks can be weighted (so that there is a *weight* value associated with each edge that represents the level of correlation of the expression) or unweighted. According to Pardo-Diaz et al. (2021a), keeping the correlation values of the gene expression as weights only if they are higher than a chosen threshold results in thresholded and weighted networks that can capture more biological information than unweighted networks, while reducing noise.

In this study, we only analyze thresholded and weighted networks that are constructed by using signed distance correlation, following the pipeline described in Pardo-Diaz et al. (2021a), with the threshold value chosen using COnsistency of Gene Expression NeTworks (COGENT) (Bozhilova et al., 2021). Networks based on signed distance correlation can be more robust and capture more biological information than those constructed using Pearson correlation, Spearman correlation, or Mutual Information (Pardo-Diaz et al., 2021a,b). This study builds on Pardo-Diaz et al. (2021a) by taking the networks from Pardo-Diaz et al. (2021a) as starting points to extract biological signals from the gene expression data.

In the literature, there are different approaches to extract information from gene networks in general and gene coexpression networks in particular (Thalamuthu et al., 2006). Here, we identify candidate genes for annotation in two different *R. leguminosarum* coexpression networks. In Cowen et al. (2017), an excellent survey is provided about the methods for gene function prediction that are based on propagating information about available functional annotation through the network.

This study uses these ideas by introducing a score that quantifies such propagation via PageRank, but it also introduces a score-based community detection as well as a score based on edge weights of nearest neighbors. We also provide a novel way to combine these scores to enhance the signals from the scores. We assess how the results depend on the different techniques and approaches employed, as well as on the input network.

Our main findings are that all three scores can successfully retrieve genes that are highly coexpressed with those in the seed list by using a set of control genes, and that a combination of these scores adds considerable signals. We also present these scores as a tool to compare different experimental conditions. Although the methods in this study are used to explore the gene coexpression networks for *R. leguminosarum*, they could be applied to other (weighted or unweighted) coexpression networks.

The article is structured as follows. In Section 2, we recall some basic network notions, the problem of community detection in networks, and personalized PageRank. Section 3 details the three scores we use in our analysis; it also includes our methodology to evaluate them, combine them, and compare them across different experimental set-ups. In Section 4, we present the results of the evaluation of the scores, and then we illustrate with two examples how to apply our approach. Lastly, in Section 5, we summarize our findings, compare them with the information available in the literature, and present directions for future work.

The data, the code, and the main outputs are available online at https://github.com/javier-pardodiaz/NetAnalysis.

2. BACKGROUND

In this study, a network is simple and undirected, on a set V of N nodes with adjacency matrix A, whose elements are edge weights; $A_{i,j}=0$ indicates that there is no edge between nodes i and j. The *degree* of node

Downloaded by 192.76.8.89 from www.liebertpub.com at 07/28/22. For personal use only.

i is the number of nodes j such that $A_{i,j} > 0$ and is denoted by k_i, for $i = 1, \ldots, N$. The number of edges is then $m = \frac{1}{2}\sum_{i=1}^{N} k_i$. The *strength* of node i is $S_i = \sum_{j;j\neq i} A_{i,j}$; for an unweighted network, the strength equals the degree.

2.1. Community detection

Detecting communities allows non-trivial internal network organizations to be unveiled at a coarse grain level (Yang et al., 2016). Communities are usually understood to be groups of nodes that are connected ''more densely'' to each other than to the rest of the network. Some community detection methods are based on the maximization of the *modularity*, that is, the difference between the number of edges in communities and the expected number under a configuration model (Newman and Girvan, 2004). For a given partition $P = \{P_1, \ldots, P_k\}$ of the set of nodes, for node i we set $\sigma_i = k$ if $i \in P_k$.

Reichardt and Bornholdt (2006) add a resolution parameter (γ) to the modularity to yield as quality function $Q(P) = \sum_{i,j}\left(A_{i,j} - \gamma\frac{k_i k_j}{2m}\right)\delta(\sigma_i, \sigma_j)$, where $\delta(\sigma_i, \sigma_j) = 1$ if i and j are in the same set in the partition P, and 0 otherwise. High values of the resolution parameter γ result in small communities; low γ values lead to large communities. A popular algorithm for maximizing $Q(P)$ is the Leiden algorithm (Traag et al., 2019). The modularity function in this algorithm is piecewise constant in γ; the algorithm also provides the values of γ at which the modularity function changes. These values are interpreted as ''optimal'' choices for γ.

2.2. PageRank and personalized PageRank

A typical objective when studying a network is to identify which nodes are the most *important* ones (those with the highest centrality). The PageRank algorithm, originally developed to rank web-pages (Brin and Page, 1998), tackles this problem. This algorithm produces a probability distribution over the N nodes of a directed network G that represents the fraction of time that a random walker spends in each node in an infinitely long random walk.

At each step, the random walker can, with probability α, choose to follow an outgoing edge from the current node (chosen with probability proportional to its weight), or, with probability $1-\alpha$, ''teleport'' to any node chosen from a distribution v. This algorithm can also be applied to undirected networks. For simplicity assume that there are no isolated nodes in the undirected network, and that all node strengths are positive, $s_i > 0$ for all $i = 1, \ldots, N$. Let $x(t) = (x_1(t), \ldots, x_N(t)) \geq 0$ be the vector with $x_i(t)$ denoting the probability that node i receives a random walker at time t (so that $\sum_{i=1}^{N} x_i(t) = 1$, for all t), which evolves according to

$$x(t) = \alpha\left(D^{-1}\right)^T x(t-1) + (1-\alpha)v. \tag{1}$$

The matrix D is the diagonal matrix of the strength of the nodes, so that $D_{i,i} = s_i > 0$ is the strength of node i and $D_{i,j} = 0$ if $i \neq j$. The PageRank vector π is the solution of equation (1) when $t \to \infty$, that is $\pi = \alpha(D^{-1})^T\pi + (1-\alpha)v$; see Xing and Ghorbani (2004).

Typically the teleportation distribution v is chosen to be uniform (i.e., $v_i = \frac{1}{N}$ for all i); however, *personalized PageRank* Gleich (2015) manipulates v to bias teleported walkers to a specific node or group of nodes. For example if $v_i = 0.9$ and $v_j = \frac{0.1}{N-1}$ for $j \neq i$, teleported walkers will go to node i with probability 0.9. This has the effect that nodes near node i will have a higher personalized PageRank than with uniform teleportation.

3. METHODS

3.1. Data

We analyze two different *R. leguminosarum* networks (*standard* and *complete*), both of them obtained applying the approach described in Pardo-Diaz et al. (2021a) but using two different microarray datasets, each of which contains gene expression values. The *standard* dataset includes three replicates of 18 different conditions, making a total of 54 microarrays; these conditions include free-living bacteria, bacteria from the soil in which legumes are grown (rhizosphere), and bacteroids. What we call the *complete* dataset includes a total of 87 microarrays from 35 different conditions; however, not all conditions have replicates.

Downloaded by 192.76.8.89 from www.liebertpub.com at 07/28/22. For personal use only.

Both networks have the same set of 7077 nodes, representing genes. These expression data are published in Karunakaran et al. (2009), Ramachandran et al. (2011), Pini et al. (2017), Prell et al. (2009), Terpolilli et al. (2016), and Garcia-Fraile et al. (2015) but they have never been analyzed jointly. The list of microarrays and the summary statistics of both networks are provided in Section SI.A of the Supplementary Data.

For the analysis in Section 4.4, we use a subset of microarrays of the published data. These microarrays include expression values for 10 conditions of interest (sample channel) and also ''control'' conditions (control channel). The expression values from the sample channel of all these microarrays are included (among other data) in our ''complete'' dataset. These different conditions, as well as the control conditions in the microarray, are given in Table 1.

We generate an expression matrix in which the rows are all the genes in the network and the columns are the expression values of both channels of each microarray. On the columns of this matrix we perform quantile normalization (Bolstad et al., 2003) so that the values in the different columns are comparable. We note that instead of using microarray data one could use RNA-Seq data. If using transcripts per million (Li and Dewey, 2011) as expression units, there is no need to quantile-normalize the expression data since the values for the different samples are already comparable.

3.2. Identification of new candidate genes

To study the networks, we present three complementary scores to identify *new candidate genes* that are highly coexpressed with a given set of genes, or with genes that are highly expressed under a particular condition. For each of the scores, we include two approaches, depending on the input.

The first approach identifies genes that are highly coexpressed with a predefined selection of genes we want to study (*seed genes*). This predefined selection of genes could include genes involved in a particular process, genes with a particular expression pattern, or just a gene that is studied as part of an exploratory analysis. We denote this preselected set of genes by L. The seed genes may not be straightforward to select. For example, for selecting genes that are highly expressed in a particular experimental condition, a typical method is to choose all genes whose expression exceeds a pre-set threshold, but an objective method for setting such a threshold is lacking. Different threshold choices can lead to different results.

To overcome this problem, the second approach retrieves scores that reflect the level of coexpression of genes with genes that are highly expressed in a given condition, taking the expression values of all the genes in that condition as an input. We denote the input expression vector by E and the expression of gene i by $E(i)$. An advantage of using these scores instead of simply selecting the genes that show the highest expression, or the highest increase in expression between conditions, is that we can extract more information from the data.

For example, one gene might increase its expression only slightly between two given conditions but could be highly coexpressed with a gene that shows a high increase in its expression. Using our methods, we can capture this gene with a relatively low expression that would not have been picked otherwise. This second approach might also be useful when using other types of continuous values instead of expression data.

Table 1. Experimental Conditions for Both Channels in the Microarrays (Sample and Control) Used to Study the Gene Expression in Different Growth Media

Abbreviation	*Condition of interest*	*Abbreviation*	*Control condition*	*Type*
Pyr	Pyruvate NH_4	Glu	Glucose NH_4	OA
Suc	Succinate NH_4	Glu	Glucose NH_4	OA
OAC	Acetate NH_4	Glu	Glucose NH_4	OA
ACAC	Acetoacetate NH_4	Glu	Glucose NH_4	OA
PCA	Protocatechuate NH_4	Pyr	Pyruvate NH_4	OA
For	Formate NH_4	Pyr	Pyruvate NH_4	OA
HBA	Hydroxybenzoate NH_4	Pyr	Pyruvate NH_4	OA
Ino	Inositol NH_4	Glu	Glucose NH_4	O
Ara	Arabinose NH_4	Glu	Glucose NH_4	S
Gal	Galactose NH_4	Glu	Glucose NH_4	S

All of these conditions are free-living. The growth conditions are OA for organic acid, O for other, and S for Sugar.

Downloaded by 192.76.8.89 from www.liebertpub.com at 07/28/22. For personal use only.

We use three ways of obtaining scores for every gene in the network. The interpretation for each of the scores is that the higher the score, the more connected the gene is with the genes in the list L, or with the genes that have high expression values. The three scoring systems are as follows.

3.2.1. Community detection co-occurrence score. We perform community detection analysis on the networks using the Leiden algorithm (Traag et al., 2019) to optimize the network partitions. Different values in the resolution parameter γ lead to different partitions, but, as a function of γ, the quality function $Q(P)$ has only finitely many local maxima. These local maxima are estimated by using the Optimiser object included in the Leiden module, and they result in a set of partitions. For a given resolution value, we then select the community that contains the highest number of seed genes.

This approach is based on the assumption that highly coexpressed genes tend to be at the same community, for an illustration that supports this assumption (see Section 4.1). To combine the output of the community detection algorithm at different resolutions, we propose a *co-occurrence score* that evaluates each gene on the frequency of it belonging to communities enriched in seed genes. Let R be the set of resolutions for which partitions are considered, let $C_r^{(i)}$ be the set of genes belonging to the same community as gene i at resolution $r \in R$, and let $\tau(i)$ be the indicator function that equals 1 if i is in the list L, and 0 otherwise. Then, we define the community detection co-occurrence score $S_{CD}^{(i)}$ for gene i as

$$S_{CD}^{(i)} = \sum_{r \in R} \frac{|C_r^{(i)} \in L| - \tau(i)}{(|C_r^{(i)}| - \tau(i))|L|}. \tag{2}$$

Thus, genes that tend to belong to small communities with a high number of seed genes across most of the partitions will have a high score. The addition of τ adjusts the co-occurrence score for genes in the seed list L, to help avoid over-estimation. When including gene expression (or any other continuous value) in our analysis, we modify the score function $S_{CD}^{(i)}$ to $Z_{CD}^{(i)}$ so that $Z_{CD}^{(i)}$ gives a high score to those genes that belong to communities that are enriched in highly expressed genes. We define

$$Z_{CD}^{(i)} = \frac{1000}{\sum_{v \in V} E(v)} \sum_{r \in R} \frac{\sum_{j \in C_r{}^{(i)}} E(j)}{|C_r{}^{(i)}|}, \tag{3}$$

where $E(j)$ denotes the expression value of gene j.

3.2.2. Personalized PageRank score. In each of the networks we analyze, we compute the PageRank vector π (Brin and Page, 1998). This vector provides a centrality measurement $\pi^{(i)}$ for each node i in the network. These values represent the fraction of time that a random walker spends in each node in an infinitely long random walk across the network (see Section 2.2). At each step, we use $\alpha = 0.85$ so that the teleporting probability is $1 - 0.85 = 0.15$ (the standard parameter value, according to Brin and Page [1998]). An extensive discussion about the choices of teleporting probabilities in different types of networks can be found in Gleich (2015); often, the standard parameter provides a good starting point for exploration.

In addition, for each gene $l \in L$ in the seed list we compute its personalized PageRank vector π_l (Gleich, 2015; see Section 2.2), with the personalization that teleportation takes place only to gene l itself. We set the entry $\pi_l^{(l)}$ to zero so that in the final score the genes in the seed list can be compared with the other genes in the network. Then, we compute the average of the personalized PageRank vectors of the genes included in the seed list;

$$\rho_L{}^{(i)} = \frac{1}{|L| - \tau(i)} \sum_{l \in L} \pi_l{}^{(i)}.$$

The vector ρ_L offers a score for each of the genes in the network, including those in the seed list. We expect genes that are highly connected to those in the seed to have higher values in ρ_L than in π. For this reason, we define our PageRank score for gene i as the ratio

$$S_{PR}^{(i)} = \rho_L{}^{(i)} / \pi^{(i)}.$$

Similar to the co-occurrence score, we may like to take gene expression into account instead. To calculate a PageRank vector $\rho_{L,E}$ that takes gene expression into account, we use a personalized PageRank

Downloaded by 192.76.8.89 from www.liebertpub.com at 07/28/22. For personal use only.

in which, when teleporting, the probability to teleport to a gene k is equals to its gene expression value $E(k)$ divided by the sum of the expression of all the genes in the network, so that genes with a low expression will be visited less often than those with a high expression. With the resulting personalized PageRank vectors $\rho_{L,E}$ we set

$$Z_{PR}^{(i)} = \rho_{L,E}^{(i)} / \pi^{(i)}.$$

3.2.3. Edge weight score. The scores in Sections 3.2.1 and 3.2.2 are relatively computer-intensive. We also introduce a straightforward edge-based score. For each gene $i \notin L$ we calculate its edge weight score $S_{EW}^{(i)}$ as

$$S_{EW}^{(i)} = \frac{1}{|L|} \sum_{l \in L} A_{i,l},$$

the average weight of the edges connecting i to the genes in the seed list. For a seed list gene $i \in L$, we set

$$S_{EW}^{(i)} = \frac{1}{|L|-1} \sum_{l \in L,\, l \neq i} A_{i,l}.$$

This score favors genes with a high degree or strength. For using gene expression instead of a seed list, we introduce the score

$$Z_{EW}^{(i)} = \frac{1}{N-1} \sum_{v \in V} A_{i,v}\ E(v).$$

3.3. Evaluation of the scores using control genes

To evaluate the performance of the different scores for studying seed lists from Sections 3.2.1–3.2.3, we use some *control genes*. These genes are 49 highly coexpressed genes that code for ribosomal proteins, according to the genome annotation in Young et al. (2006). The average signed distance correlation (Pardo-Diaz et al., 2021b) of the expression of the control genes is 0.79 when using the standard dataset and 0.70 when using the complete one.

The control genes are listed in Supplementary Table S4 in the Supplementary Data. According to Wheatley et al. (2020), most of these genes are essential for bacterial growth in all three studied conditions, namely free-living, rhizosphere, and bacteroid. The control genes show a high expression in free-living conditions and a lower expression in the rhizosphere and bacteroid stages. This difference in their expression might relate to the fact that the growth of bacteria and the synthesis of proteins decreases during the two latter stages (especially in bacteroids, which do not divide) and therefore ribosomes are less needed.

To evaluate the scores, we take a random subset of five control genes and use them as seed list (*control seed list*). Then, we score all the genes in the coexpression network by using the three different scores from Sections 3.2.1–3.2.3. The higher the scores, the more related the genes are deemed to be to the selected five genes. As all the control genes are coexpressed and functionally related, we expect the remaining 44 control genes to retrieve high score values.

We compare the rank of these 44 genes for the different scores with their ranks when sampling all the genes in the network uniformly at random. If the scores are informative, then the rank of the 44 genes should be larger using the control seed list compared with ranking them randomly. We use the scores as a classifier of all genes, classifying whether or not they belong to the list of 49 ribosomal proteins, so that there are exactly 44 true positives. We repeat this simulation process 25 times, at each iteration taking a potentially different random subset of five genes as control seed list. To evaluate the results, we use the area under the receiver operating characteristics curve (AUROC).

3.4. Combination of the scores

Given the scores from Sections 3.2.1–3.2.3, we select genes based on their individual scores as well as their performance across all three scores. Specifically, we select a set of genes based on the rule ''among the top M for at least one of the three scores, and, of those, we select the genes which are among the top T for all three scores, or among the top S for at least one of the scores.'' Here, we assume that $T \geq M \geq S$.

Downloaded by 192.76.8.89 from www.liebertpub.com at 07/28/22. For personal use only.

If we were just to select the top S genes for each of K scores, we would expect to select roughly KS genes. To assess the effect of selecting also genes that are among the top M for any of the K scores, and which are among the top T for all K scores, let W denote the number of such genes selected. Using a Poisson approximation, we can approximate the probability of selecting at least as many genes as W if the ranks were allocated at random for each score. Let

$$\lambda = \lambda(M, T; K, N) = N\left\{\left(\frac{T}{N}\right)^K - \left(\frac{T-M}{N}\right)^K\right\}. \quad (4)$$

Then, for $m \geq 3\lambda$,

$$\mathbb{P}(W \geq m) \leq 2\frac{m+1}{(m+1-\lambda)}\frac{1}{\sqrt{2\pi m}}\exp\left\{-\frac{(m-\lambda)^2}{2(m+\lambda)}\right\}. \quad (5)$$

More details can be found in Section SI.D of the Supplementary Data.

3.5. Joint analysis for different experimental conditions

In this study, we analyze *R. leguminosarum* gene expression across 10 different experimental conditions, as detailed in Table 1. We are not interested in the most highly expressed genes under the different conditions, as most of them will be the same constitutive genes that are highly expressed under all conditions. Instead, we focus on those genes that are upregulated when compared with a reference condition; in this study, we use growth in glucose as a reference condition.

To obtain the genes with high scores for the different conditions compared with their scores in the glucose reference condition, we use the quantile-normalized expression matrix presented in Section 3.1. We first calculate the Z_{CD}, Z_{PR}, and Z_{EW} scores for all the genes in the network, as detailed in Sections 3.2.1–3.2.3, for each of the 10 conditions of interest and for the 10 control conditions, by using the expression values from the sample and control channels in the microarrays, respectively. This procedure yields 20 different sets of three scores for each gene (3 scores, 10 microarrays, 2 channels per microarray).

If for a microarray the control condition is the same condition as the one that we compare against (here, glucose), we just divide the scores for the condition of interest by the scores for the control condition. Otherwise, we use the data from a second microarray that has as condition of interest the control condition of the first microarray: We divide the product of the scores for both conditions of interest by the product of the scores for both control conditions. We, thus, obtain a set of three scores for each gene for each of the 10 conditions. We compare the genes with high scores across the different conditions.

Applying the procedure from Section 3.4 to the just obtained scores results in a set of potentially relevant genes for each individual condition; these sets may differ in sizes for different experimental conditions. Now we study the intersection of the sets of genes obtained for different conditions. For this task for simplicity we leave out the pre-filtering using M and instead, for each condition we take the genes that are either in the top S for at least one of the scores, or in the top T for all of the scores. We denote the union of the genes that are selected for the different conditions by Q.

We use $S = 5$ and $T = 20$, which makes Q likely to be relatively small. To understand the behavior of genes in Q across conditions, we now introduce a possibly more lenient threshold, U. For each gene in Q, for each condition we record whether or not it is among the top U genes for at least one of the scores. Thus, every gene will have at least one entry, namely the condition that qualified the gene to be part of the union Q. It may, however, have relatively high scores for some other conditions as well, which is what this procedure addresses. The pipeline for this procedure is shown in Figure 1. This analysis results in a matrix of genes and conditions, which is the main output of the pipeline in this study. The resulting pattern of genes and conditions could then be examined in further detail, for example using in vitro experiments.

Summarizing, in this study we present two methods to identify new candidate genes: one based on the use of seed gene scores (S_{XX}) and another based on the use of gene expression scores (Z_{XX}). Next, in Section 4.2, we evaluate the scores. Then we illustrate their use on *R. leguminorasum* in two ways. The first way is the use of all the ribosomal genes as seed list (Section 4.3). The second way is the use of the expression across different growth conditions (Section 4.4). These two approaches will turn out to identify different sets of genes for further study and they, thus, complement each other.

Downloaded by 192.76.8.89 from www.liebertpub.com at 07/28/22. For personal use only.

Downloaded by 192.76.8.89 from www.liebertpub.com at 07/28/22. For personal use only.

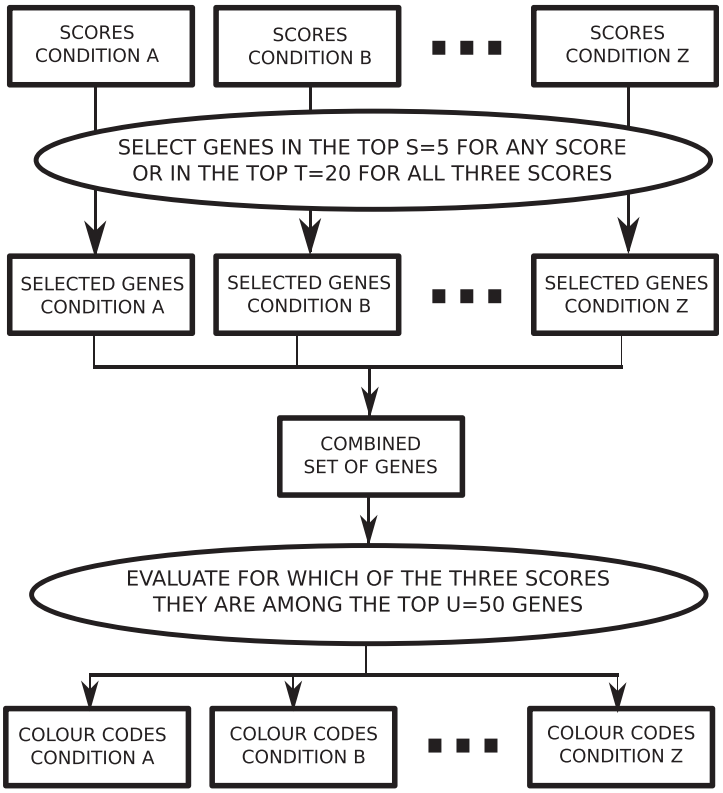


FIG. 1. The pipeline for combining the scores for different conditions for retrieving genes for further investigation. The first step is to select for each condition those genes that are in the top $S=5$ for any of the three scores or in the top $T=20$ for all three of them. The second step consists of evaluating at each condition which of the three scores the selected genes are among the top $U=20$ scores. The output is a coded matrix in which each of the genes in the combined set of genes has a value for each condition (see Fig. 6 for reference).

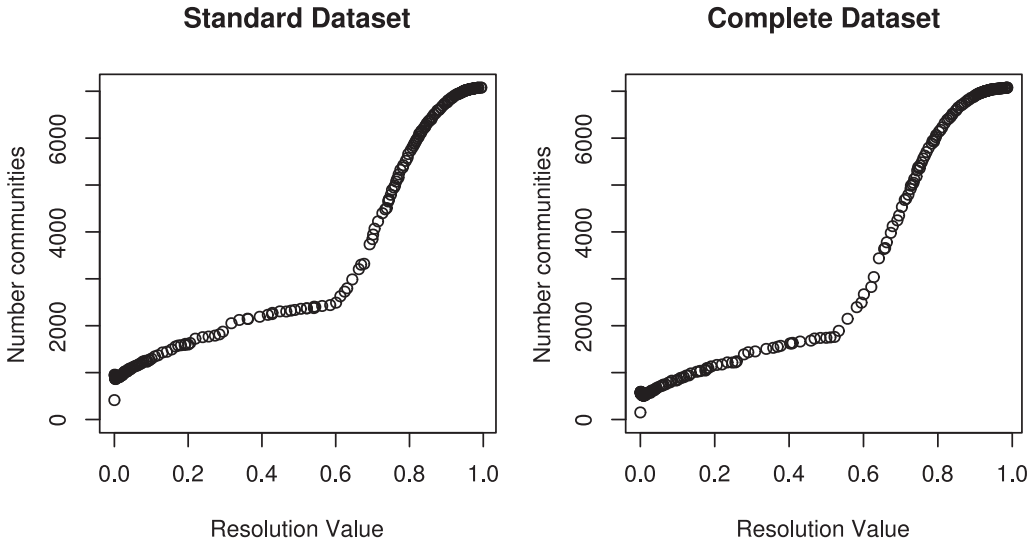


FIG. 2. The number of communities in partitions depends on the value of the resolution value. The behavior in the standard dataset and in the complete dataset show a similar nonlinear dependence on the resolution value.

4. RESULTS

4.1. Community detection results

Community detection as described in Section 2.1 yields 226 partitions for the standard dataset and 241 partitions for the complete dataset. Figure 2 presents the number of communities that we retrieve at each partition.

As an illustration that communities include functionally related genes and that the outcomes may depend on the chosen partition, Figure 3 shows that when using ribosomal genes as seed genes, a community of 48 genes contains a large number of 19 ribosomal genes, which is much higher than the number of genes expected by chance.

Moreover, Figure 3 shows that communities with different sizes can result in a different subset of ribosomal genes. Hence, a principled way of combining the results of community detection outputs is required, further motivating our score from Section 3.2.1.

Downloaded by 192.76.8.89 from www.liebertpub.com at 07/28/22. For personal use only.

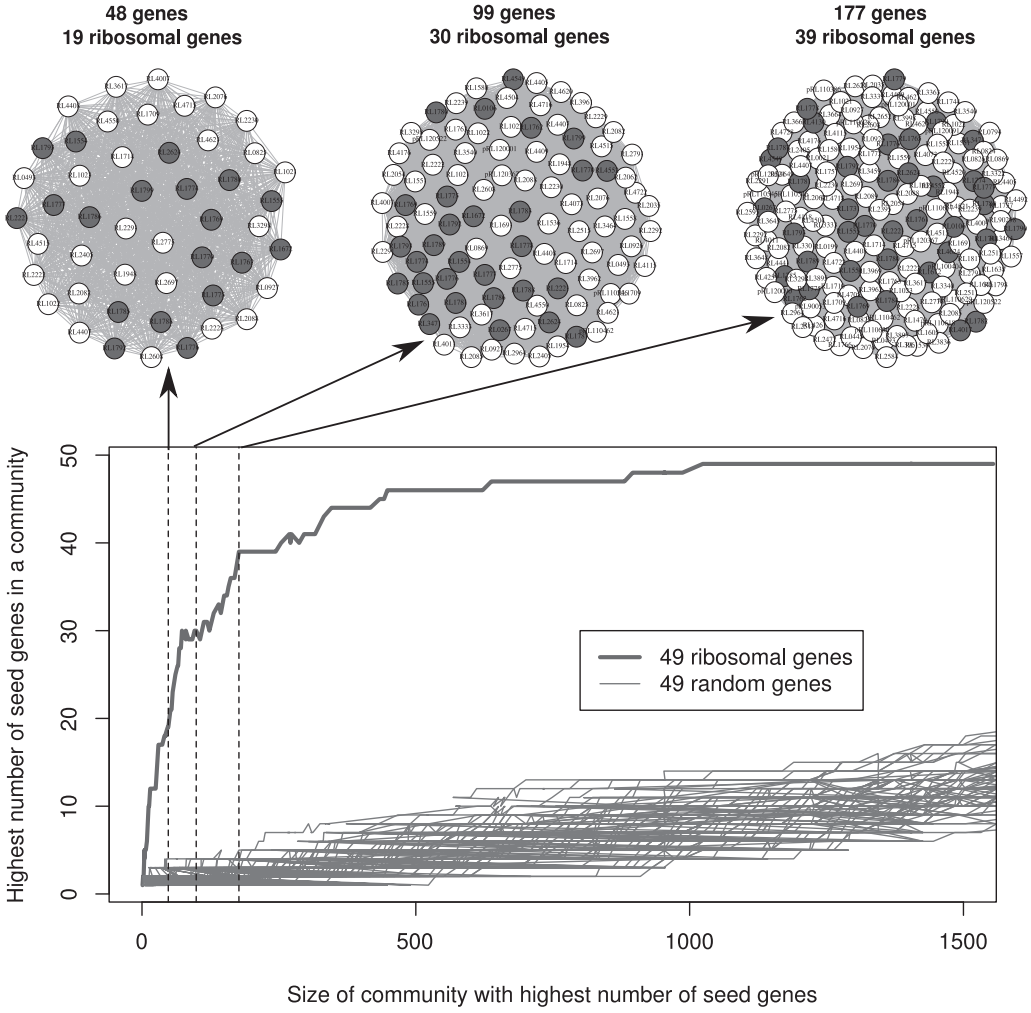


FIG. 3. Top: Communities with the highest number of ribosomal genes (filled) for different partitions. Bottom: Plot representing for different partitions the size of the community with the highest number of ribosomal genes and the number of ribosomal genes in those communities.

4.2. Performance of the seed list scores

As detailed in Section 3.3, we use different subsets of the control genes as control seed lists to evaluate the performance of the scores S_{CD}, S_{PR}, and S_{EW}. We score all the genes in the networks by using the scores and rank the genes; then, we compare the rank of the remaining control genes (those not included in the control seed list) with their ranks when sampling all the genes in the network uniformly at random. If the scores were unrelated to the functionality of the genes, then all orderings would be equally likely, and control genes would not have higher scores on average than other genes.

Using the score as a classifier of whether or not a gene is a control gene (a ribosomal gene in this case), we assess its performance through the AUROC. We expect to observe several ''not ribosomal genes'' with high scores since genes coexpressed with the ribosomal genes do not necessarily need to be ribosomal genes themselves; nevertheless, we expect all ribosomal genes to have a high score. Only the 44 ribosomal genes that are not in the control seed list are counted as true positives.

The upper panels in Figure 4 show the relationship between the true positive rate and the false positive rate for the studied networks (standard and complete) when using control seed lists of size five. The boxplots reflect the distribution of the AUROC. We observe that the AUROC for the three scores is higher than the one obtained when ordering the genes randomly, which indicates that the remaining control genes have higher scores than expected.

We obtain the same results when we include a different number of genes in the control seed list (Supplementary Fig. S1 in the Supplementary Data). These results suggest that all the three scores capture some signal and are appropriate for retrieving genes functionally related to those included in the seed list.

Figure 4 indicates that, for both networks, when including five control genes in the control seed list, the AUROC for the edge weight scores is significantly higher than for the other two scores (ANOVA tests with p values < 0.05). However, we do not observe this trend when using other control seed list sizes instead

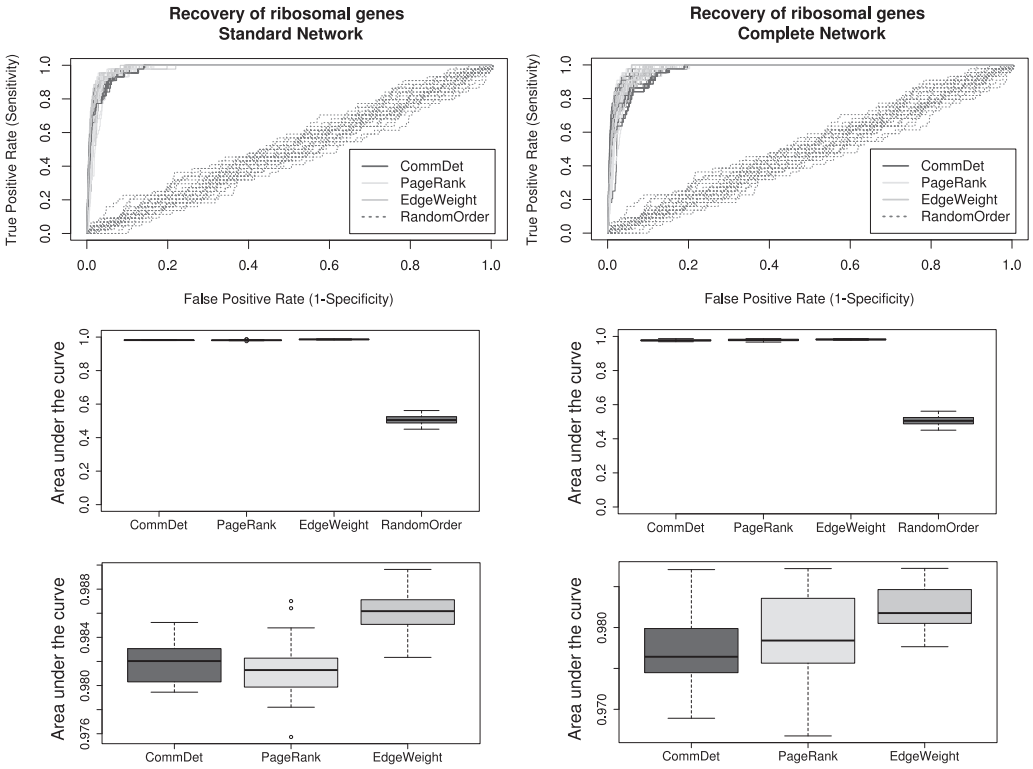


FIG. 4. Scores obtained using the control seed list. All scores perform significantly better than random. There is no clear ranking between the scores.

Downloaded by 192.76.8.89 from www.liebertpub.com at 07/28/22. For personal use only.

of five (Supplementary Fig. S1 in the Supplementary Data); we note that for Supplementary Figure S1 we only perform 10 instead of 25 iterations. The only significant difference we observe is that the edge weight score performs better than the PageRank score when the size of the control seed list is 25 and using the standard network. Therefore, we do not have strong evidence to favor one score above the others.

We also observe that for most cases the AUROC for the standard network is higher than for the complete network. We discuss this point further in Section 5.

4.3. Scoring genes based on the ribosomal genes

We study which genes in the network are highly coexpressed with the ribosomal genes listed in Supplementary Table S4 in the Supplementary Data. For this purpose, we use all the ribosomal genes as a seed list (in contrast to Section 4.2 where we use subsets of the ribosomal genes as control seed lists), and we retrieve the scores S_{CD}, S_{PR}, and S_{EW}. Supplementary Figure S2 in the Supplementary Data shows that there is clearly some, but not a perfect, correlation between the scores; we observe Spearman's rank correlation coefficients between 0.65 and 0.88 when comparing the different scores against each other and Spearman's rank correlation coefficients between 0.79 and 0.91 when comparing the scores for the different datasets. A Friedman test of the null hypothesis that there is no association between the different rankings is rejected with p value 2.2×10^{-16}, providing very strong evidence that there is an association between the rankings. For the standard network and the community detection co-occurrence score, we observe that out of the top 48 genes, 33 are ribosomal genes, whereas, if considering the different partitions separately, according to Figure 3. in a community of 48 genes there are only 19 ribosomal genes. This shows the potential of combining the different resolutions.

The next step in our analysis is selecting those genes with a high ranking for one score or, with a possibly slightly lower ranking across all of the scores. We exclude the ribosomal genes from these ranks. First, we select those genes in the top $S = 10$ for any of the three scores, resulting in 26 and 25 genes for the complete and standard networks, respectively. This count is very much in line with what we would expect from independent counts, indicating that different scores might provide different information when attending to the highest scoring genes.

Then, we concentrate on those genes that are in the top $M = 50$ for at least one of the rankings and in the top $T = 100$ for all of them and assess whether the number of genes that are selected this way is unusually large, using the Poisson approximation from Eq. (5). Their expected count is 0.01771514, which is calculated using Eq. (4). For this category we observe $W = 17$ in the standard network and $W = 26$ in the complete network. Proposition Eq. (1) in the Section SI.D of the Supplementary Data, which applies because $17 > 3 \times 0.01771514$, gives that the probability of observing a count of 17 or higher if the rankings were random would be less than 4.05×10^{-5}, which is highly significant. Thus, there is strong evidence that the choice of combining the rankings by picking not only genes that are highly ranked in one of them, but also genes that are somewhat less highly ranked, but across all rankings, adds considerable signals.

Overall, for our analysis of the standard as well as the complete network we select the genes that are in the top $M = 50$ for at least one of the rankings and in the top $T = 100$ for all of them, and those in the top $S = 10$ for any of the scores. The final count we get is 35 genes for the standard network and 39 genes for the complete network; 23 genes are shared between the two networks. The selected genes are detailed in Supplementary Table S5 in the Supplementary Data.

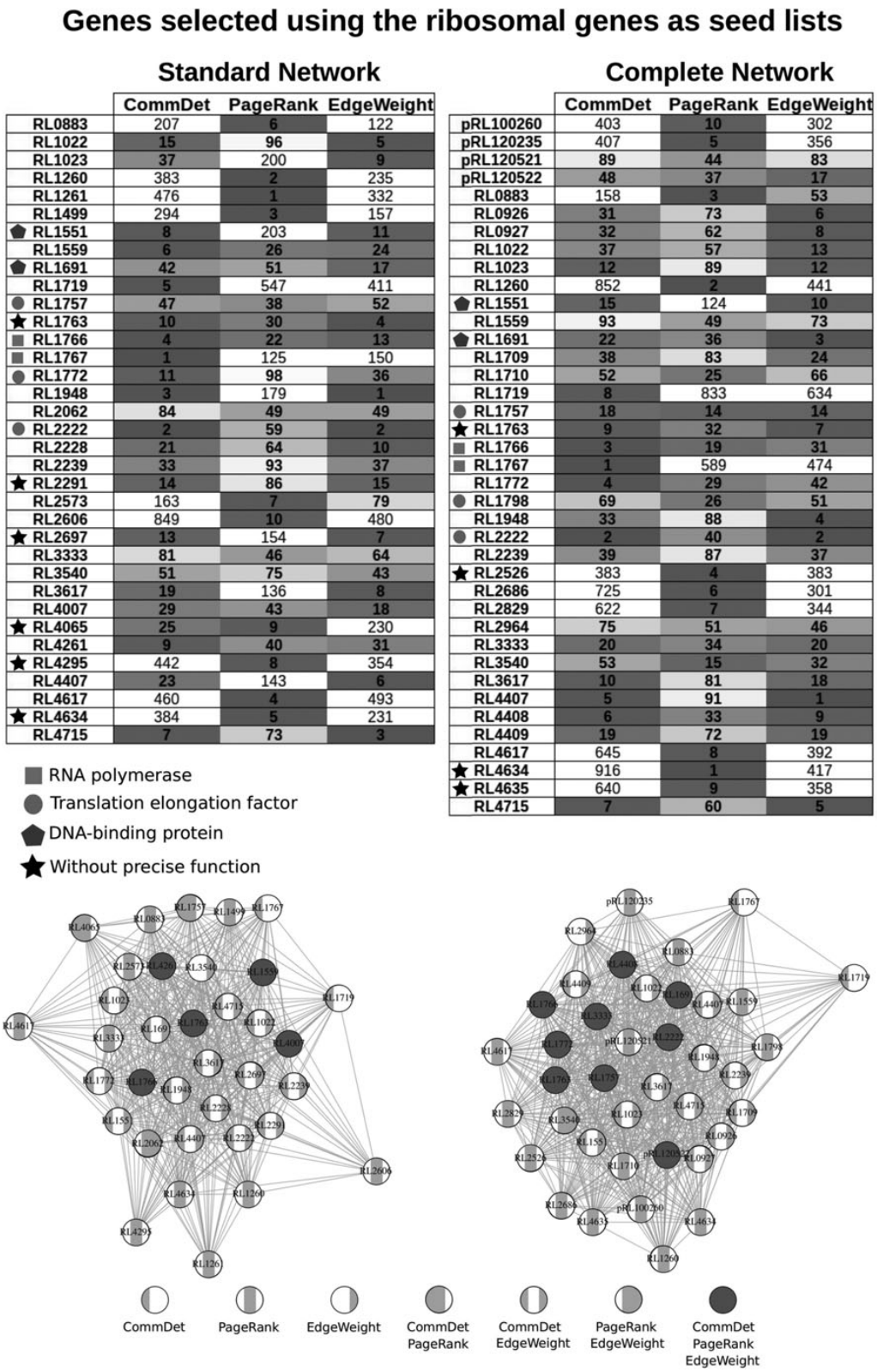
The average weight of the edges connecting the selected genes with the ribosomal genes is 0.79 (0.73 if including the 131 missing edges) for the standard network and 0.72 (0.68 if including the 121 missing edges) for the complete one. The upper panels in Figure 5 show the selected genes for the two networks and their ranks. Genes are colored for a given score if their rank is below 100 and white otherwise; the more intense the color, the lower the rank.

We observe that the rank with the smallest overlap with the other two is the one derived from the PageRank score, suggesting that this score provides some different information. The bottom images represent induced subgraphs of the networks containing the selected genes. The nodes are colored based on the scores for which the genes are in the top $T = 100$: only S_{CD} (red), only S_{PR} (yellow), only S_{EW} (blue), S_{CD} and S_{PR} (orange), S_{CD} and S_{EW} (purple), S_{PR} and S_{EW} (green), or all of them (gray). These network plots show that the genes are highly connected.

Most of the selected genes are essential or defective for free-living, rhizosphere, and bacteroid conditions, according to Wheatley et al. (2017). Out of the selected genes we find genes that code for DNA-

Downloaded by 192.76.8.89 from www.liebertpub.com at 07/28/22. For personal use only.

Downloaded by 192.76.8.89 from www.liebertpub.com at 07/28/22. For personal use only.



	CommDet	PageRank	EdgeWeight
RL0883	207	6	122
RL1022	15	96	5
RL1023	37	200	9
RL1260	383	2	235
RL1261	476	1	332
RL1499	294	3	157
RL1551	8	203	11
RL1559	6	26	24
RL1691	42	51	17
RL1719	5	547	411
RL1757	47	38	52
RL1763	10	30	4
RL1766	4	22	13
RL1767	1	125	150
RL1772	11	98	36
RL1948	3	179	1
RL2062	84	49	49
RL2222	2	59	2
RL2228	21	64	10
RL2239	33	93	37
RL2291	14	86	15
RL2573	163	7	79
RL2606	849	10	480
RL2697	13	154	7
RL3333	81	46	64
RL3540	51	75	43
RL3617	19	136	8
RL4007	29	43	18
RL4065	25	9	230
RL4261	9	40	31
RL4295	442	8	354
RL4407	23	143	6
RL4617	460	4	493
RL4634	384	5	231
RL4715	7	73	3

	CommDet	PageRank	EdgeWeight
pRL100260	403	10	302
pRL120235	407	5	356
pRL120521	89	44	83
pRL120522	48	37	17
RL0883	158	3	53
RL0926	31	73	6
RL0927	32	62	8
RL1022	37	57	13
RL1023	12	89	12
RL1260	852	2	441
RL1551	15	124	10
RL1559	93	49	73
RL1691	22	36	3
RL1709	38	83	24
RL1710	52	25	66
RL1719	8	833	634
RL1757	18	14	14
RL1763	9	32	7
RL1766	3	19	31
RL1767	1	589	474
RL1772	4	29	42
RL1798	69	26	51
RL1948	33	88	4
RL2222	2	40	2
RL2239	39	87	37
RL2526	383	4	383
RL2686	725	6	301
RL2829	622	7	344
RL2964	75	51	46
RL3333	20	34	20
RL3540	53	15	32
RL3617	10	81	18
RL4407	5	91	1
RL4408	6	33	9
RL4409	19	72	19
RL4617	645	8	392
RL4634	916	1	417
RL4635	640	9	358
RL4715	7	60	5

FIG. 5. Top: The genes that are selected using $M = 50$, $S = 10$, and $T = 100$, for the standard network and the complete network. The shade reflects the rank. The symbols next to the gene names indicate functional annotation. Bottom: the induced subgraphs using only these sets of genes. The stripes indicate which scores the genes are in the top $T = 100$.

binding proteins (RL1551 and RL1691), translation elongation factors (RL1757, RL1772, and RL2222), and subunits of the RNA polymerase (RL1766, RL1767, and RL1798). All these proteins are involved in cell growth and protein synthesis and therefore expected to be coexpressed with ribosomal genes.

In the selection, we also find several genes without functional annotation (RL1763, RL2291, and RL4065) and genes (RL2526, RL2697, RL4295, RL4634, and RL4635) for which only a general functional annotation is available; for example, a gene may be known to be an oxidoreductase but we do not know which specific process they take part in. From their association with ribosomal genes, we conjecture that these genes might be related to protein synthesis, too.

4.4. Study of different growth conditions

In this subsection, we analyze the genes with high Z_{CD} and Z_{PR}, and Z_{WE} scores for the conditions of interest, detailed in Table 1, compared with growth in glucose, following the approach detailed in Section 3.5. To select the genes included in Q we use $T = 20$ and $S = 5$. We retrieve 94 and 80 genes for the standard and the complete network, respectively.

These genes are detailed in Supplementary Table S6 in the Supplementary Data and illustrated in Figure 6. In this figure, they are shade-based on the set of scores for which they are among the top $U = 50$ scores: only Z_{CD}, only Z_{PR}, only Z_{EW}, Z_{CD} and Z_{PR}, Z_{CD} and Z_{EW}, Z_{PR} and Z_{EW}, or all of them (black). We observe that the score that overlaps the least with the others is the Edge Weight score.

We observe that the sets of genes selected for the different organic acids (Pyr, Suc, OAC, ACAC, PCA, For, and HBA, as in Table 1) are more similar to each other than to those for the other conditions. This is unsurprising since these seven conditions are relatively similar to each other.

Moreover, RL0037 (phosphoenolpyruvate carboxykinase [pckA]) has high scores (among the top $U = 50$ genes for at least one of the scores) in all the studied conditions except inositol for the complete dataset; for the standard dataset, RL0037 appears for all the conditions except for inositol, galactose, and formate. Similarly, RL4012 (fructose-biphosphate aldolase [fbaA]) appears for all the conditions for the complete dataset. Both enzymes, pckA and fbaA are involved in the synthesis of glucose through the

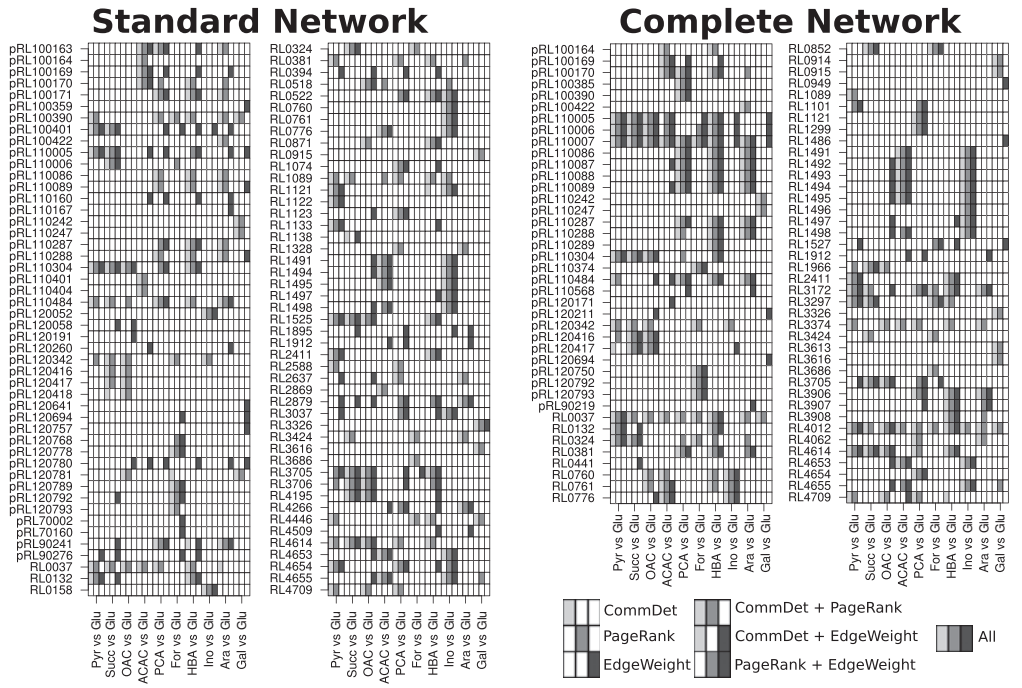


FIG. 6. Genes included in Q when $T = 20$ and $S = 5$. The three entries corresponding to each condition indicate which scores the genes are in the top $U = 50$ for the different conditions. OA, O, and S denote organic acid, other, and sugar, respectively.

Downloaded by 192.76.8.89 from www.liebertpub.com at 07/28/22. For personal use only.

gluconeogenesis pathway. As the conditions are compared against glucose, it is not surprising to retrieve RL0037 and RL4012, since gluconeogenesis will not be active in bacteria grown in glucose. We do not see the same effect for other gluconeogenesis enzymes since they also catalyze the glycolysis reactions.

A group of three genes involved in alanine catabolism: pRL120416 (alanine racemase), pRL120417 (D-amino acid dehydrogenase), and RL1966 (alanine dehydrogenase) are among the top $U = 50$ genes for the three scores for succinate, acetate, and formate; RL1966 is only selected for the complete network. These three acids, when metabolized by the cell, are transformed into acetyl coenzyme A (acetyl-CoA). To avoid the accumulation of acetyl-CoA, acetyl-CoA is used to synthesize alanine. We do not observe these genes for pyruvate, because growth in pyruvate does not result in an excess of acetyl-CoA.

Further, the genes involved in the catabolism of inositol (RL1491, RL1494–RL1498) are among the top $U = 50$ for all three scores and both datasets. Also, the genes related to the metabolism of protechatechuate (pRL110086–pRL110089) are identified for protechatechuate, as well as for hydoxybenzoate and arabinose, for the complete dataset. These findings suggest some similarities between these three conditions.

There is an operon (pRL110005–pRL110007) that for the complete dataset is selected by at least two scores for all the organic acids except protocatechuate. This operon includes a transmembrane protein, an anti-sigma factor, and an RNA polymerase extracytoplasmic function (ECF) factor, suggesting that it is involved in a regulatory process that might be related to growth in this condition.

Lastly, there are some genes such as RL0132, RL0852, and RL3172 for which no functional annotation is currently available. We conjecture that these genes might be related to the growth in the conditions for which they are selected.

5. DISCUSSION

Summarizing, this study introduces a principled pipeline to extract biological information from gene coexpression networks. This method is an exploratory tool to analyze and visualize gene coexpression networks, which can help to identify new putative functional annotations for some genes.

We introduce three scores to retrieve genes that are highly coexpressed with a pre-selected set of genes (given by a seed list), or with genes that are highly expressed in a particular experimental condition. In both cases, the output is a set of three scores for each gene in the network. These scores are used to rank the genes for each score separately, as well as for a combination of scores. We also present a method to compare these scores across different conditions. The flexibility of choice of the parameters (S, M, T, U) allows the user to tune the size and requirements of the final set of genes.

We illustrate two ways of how to apply our approach to real gene expression data, focusing on two networks for the nitrogen-fixing bacteria *R. leguminosarum*. These two networks are constructed from different (and overlapping) sets of microarrays following the pipeline described in Pardo-Diaz et al. (2021a).

In the first application, we use all 49 ribosomal genes as a seed list to identify genes that are highly coexpressed with this set. Applying our approach, the final selection includes a high number of genes that are involved in protein synthesis and that are associated with cell growth. In the second application, we study and compare the genes that are selected under different growth media, illustrating that our method does not only help to identify the genes that are strongly connected with those that are highly expressed, but it also allows to compare the selected genes across conditions. Comparing across conditions may yield insights into which biological processes are shared across different experimental conditions; in this example, these are mainly metabolic processes.

Most of our results are in line with those included in Karunakaran et al. (2009), namely the findings related to gluconeogenesis, and alanine and inositol catabolism. Nevertheless, to our knowledge, in the literature there is no reference to the operon pRL110005–pRL110007 that we mention in the results section, nor to the genes RL0132, RL0852, and RL3172. Thus, our pipeline suggests potential functional annotations for some genes that currently do not have detailed functional annotation and that have not been related to cell growth in the conditions we study.

These results evince that our principled pipeline can help to increase our biological insights while agreeing with previously published data and illustrate that using gene coexpression networks from transcriptomic data can offer an advantage compared with only looking at expression levels, as was the approach taken, for example in Karunakaran et al. (2009).

Downloaded by 192.76.8.89 from www.liebertpub.com at 07/28/22. For personal use only.

Finally, there are two issues to flag. First, we have shown that the three scores can successfully retrieve genes that are highly coexpressed with those in the seed list by using a set of control genes. Here, the control genes are ribosomal genes that are involved in protein synthesis; they have a high strength (418.57 and 490.54 for the standard and complete network, respectively) compared with the average node strength in the network (74.51 and 117.47). This difference in strength may be advantageous for our scoring methods. Ideally, a list of control genes would be relatively large but comparable in strength to the full list of genes. Unfortunately, for *R. leguminosarum* we could not identify such an alternative list of control genes.

Second, throughout the article we use two datasets. For both of them, the results are similar but there are some noticeable differences. When analyzing the AUROC for the different control seed lists, often the standard network retrieves a higher AUROC. This behavior might relate to the fact that the standard dataset contains a higher proportion of bacteroid and rhizosphere samples, where ribosomal genes are down-regulated, which may result in a higher correlation between them. The differences when analyzing the growth conditions might be related to the fact that the standard dataset contains fewer samples; in particular, some specific conditions such as growth in protocatechuate and hydroxybenzoate are not included. Therefore, different networks might offer slightly different information.

In future work, we will use the approach in this study to better understand the metabolic and regulatory changes in *R. leguminosarum* when infecting the plant. For this purpose, we will use new RNA-Seq datasets from rhizosphere and bacteroid samples.

Downloaded by 192.76.8.89 from www.liebertpub.com at 07/28/22. For personal use only.

AUTHORS' CONTRIBUTIONS

J.P.D.: Conceptualization (equal), data curation (lead), formal analysis (lead), investigation (lead), methodology (lead), project administration (equal), software, validation (lead), visualization (lead), writing—original draft (lead), and writing—review and editing (equal). M.B.D.: Conceptualization (equal), methodology, project administration (equal), funding acquisition (equal), and supervision. P.S.P.: Conceptualization (equal), data curation, project administration (equal), resources (equal), funding acquisition (equal), and supervision. C.M.D.: Conceptualization (equal), methodology, project administration (equal), resources (equal), and supervision. G.R.: Conceptualization (equal), formal analysis, investigation, methodology, project administration (equal), resources (equal), validation, visualization, writing—original draft, writing—review and editing (equal), funding acquisition (equal), and supervision (lead).

ACKNOWLEDGMENTS

The authors would like to thank the anonymous referees for their helpful suggestions.

AUTHOR DISCLOSURE STATEMENT

The authors declare they have no conflicting financial interests.

FUNDING INFORMATION

This work is supported by the Engineering and Physical Sciences Research Council (EP/R512333/1 to J.P.D., M.B.D., P.S.P., and G.R.; EP/T018445/1 and EP/R018472/1 to G.R.), as well as by the Biotechnology and Biological Sciences Research Council (BB/T001801/1 to P.S.P. and G.R.). M.B.D. acknowledges support from the Oxford-Emirates Data Science Lab.

SUPPLEMENTARY MATERIAL

Supplementary Data
Supplementary Figure S1
Supplementary Figure S2

Supplementary Table S1
Supplementary Table S2
Supplementary Table S3
Supplementary Table S4
Supplementary Table S5
Supplementary Table S6

REFERENCES

Bolstad, B.M., Irizarry, R.A., Åstrand, M., et al. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*. 19, 185–193.

Bozhilova, L.V., Pardo-Diaz, J., Reinert, G., et al. 2021. COGENT: Evaluating the consistency of gene co-expression networks. *Bioinformatics.* 37, 1928–1929.

Brin, S., and Page, L. 1998. The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.* 30, 107–117.

Cowen, L., Ideker, T., Raphael, B.J., *et al*. 2017. Network propagation: A universal amplifier of genetic associations. *Nat. Rev. Genet*. 18, 551–562.

Downie, J.A. 2010. The roles of extracellular proteins, polysaccharides and signals in the interactions of rhizobia with legume roots. *FEMS Microbiol. Rev*. 34, 150–170.

Garcia-Fraile, P., Seaman, J.C., Karunakaran, R., et al. 2015. Arabinose and protocatechuate catabolism genes are important for growth of *Rhizobium leguminosarum biovar viciae* in the pea rhizosphere. *Plant Soil*. 390, 251–264.

Gleich, D.F. 2015. PageRank beyond the web. *Siam Rev*. 57, 321–363.

Gutiérrez, R.A. 2012. Systems biology for enhanced plant nitrogen nutrition. *Science*. 336, 1673–1675.

Hughes, T.R., Marton, M.J., Jones, A.R., et al. 2000. Functional discovery via a compendium of expression profiles. *Cell*. 102, 109–126.

Karunakaran, R., Ramachandran, V., Seaman, J., et al. 2009. Transcriptomic analysis of *Rhizobium leguminosarum biovar viciae* in symbiosis with host plants *Pisum sativum* and *Vicia cracca*. *J. Bacteriol*. 191, 4002–4014.

Lee, H.K., Hsu, A.K., Sajdak, J., et al. 2004. Coexpression analysis of human genes across many microarray data sets. *Genome Res*. 14, 1085–1094.

Li, B., and Dewey, C.N. 2011. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. 12, 1–16.

Magwene, P.M., and Kim, J. 2004. Estimating genomic coexpression networks using first-order conditional independence. *Genome Biol*. 5, R100.

Makrodimitris, S., Reinders, M.J., and van Ham, R.C. 2020. Metric learning on expression data for gene function prediction. *Bioinformatics*. 36, 1182–1190.

Newman, M.E., and Girvan, M. 2004. Finding and evaluating community structure in networks. *Phys. Rev. E*. 69, 026113.

Pardo-Diaz, J., Bozhilova, L.V., Beguerisse-Díaz, M., et al. 2021b. Robust gene coexpression networks using signed distance correlation. *Bioinformatics*. 37, 1982–1989.

Pardo-Diaz, J., Poole, P., Beguerisse-Díaz, M., et al. 2021a. Generating weighted and thresholded gene coexpression networks using signed distance correlation. *bioRxiv*, 2021.11.15.46827. To appear in *Network Science*.

Pini, F., East, A.K., Appia-Ayme, C., et al. 2017. Lux bacterial biosensors for in vivo spatiotemporal mapping of root secretion. *Plant Physiol*. 174, 1289–1306.

Prell, J., Bourdès, A., Karunakaran, R., et al. 2009. Pathway of γ-aminobutyrate metabolism in *Rhizobium leguminosarum 3841* and its role in symbiosis. *J. Bacteriol*. 191, 2177–2186.

Ramachandran, V.K., East, A.K., Karunakaran, R., et al. 2011. Adaptation of *Rhizobium leguminosarum* to pea, alfalfa and sugar beet rhizospheres investigated by comparative transcriptomics. *Genome Biol*. 12, R106.

Reichardt, J., and Bornholdt, S. 2006. Statistical mechanics of community detection. *Phys. Rev. E*. 74, 016110.

Stuart, J.M., Segal, E., Koller, D., et al. 2003. A gene-coexpression network for global discovery of conserved genetic modules. *Science*. 302, 249–255.

Terpolilli, J.J., Masakapalli, S.K., Karunakaran, R., et al. 2016. Lipogenesis and redox balance in nitrogen-fixing pea bacteroids. *J. Bacteriol*. 198, 2864–2875.

Thalamuthu, A., Mukhopadhyay, I., Zheng, X., et al. 2006. Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics*. 22, 2405–2412.

Traag, V.A., Waltman, L., and van Eck, N.J. 2019. From Louvain to Leiden: Guaranteeing well-connected communities. *Sci. Rep*. 9, 1–12.

van Noort, V., Snel, B., and Huynen, M.A. 2003. Predicting gene function by conserved co-expression. *Trends Genet*. 19, 238–242.

Downloaded by 192.76.8.89 from www.liebertpub.com at 07/28/22. For personal use only.

Wang, Y.R., Waterman, M.S., and Huang, H. 2014. Gene coexpression measures in large heterogeneous samples using count statistics. *Proc. Natl. Acad. Sci. USA*. 111, 16371–16376.

Waterman, M. 2016. *Getting Outside: A Far-western Childhood*. CreateSpace Independent Publishing Platform, USA.

Weirauch, M.T. 2011. Gene coexpression networks for the analysis of DNA microarray data. *Appl. Stat. Netw. Biol. Methods Syst. Biol*. 1, 215–250.

Wheatley, R.M., Ford, B.L., Li, L., et al. 2020. Lifestyle adaptations of rhizobium from rhizosphere to symbiosis. *Proc. Natl. Acad. Sci. USA*. 117, 23823–23834.

Wheatley, R.M., Ramachandran, V.K., Geddes, B.A., et al. 2017. Role of O2 in the growth of *Rhizobium leguminosarum bv. viciae 3841* on glucose and succinate. *J. Bacteriol*. 199, e00572-16.

Wolfe, C.J., Kohane, I.S., and Butte, A.J. 2005. Systematic survey reveals general applicability of ''guilt-by-association'' within gene coexpression networks. *BMC Bioinformatics.* 6, 227.

Xing, W., and Ghorbani, A. 2004. Weighted PageRank algorithm, pp. 305–314. Proceedings. Second Annual Conference on Communication Networks and Services Research. IEEE, May 19–21, 2004, Fredericton, N.B., Canada.

Yang, Z., Algesheimer, R., and Tessone, C.J. 2016. A comparative analysis of community detection algorithms on artificial networks. *Sci. Rep*. 6, 30750.

Young, J.P.W., Crossman, L.C., Johnston, A.W., et al. 2006. The genome of *Rhizobium leguminosarum* has recognizable core and accessory components. *Genome Biol*. 7, R34.

Address correspondence to:
Mr. Javier Pardo-Diaz
Department of Statistics
University of Oxford
Oxford OX1 3LB
United Kingdom

E-mail: jdiaz@stats.ox.ac.uk

Downloaded by 192.76.8.89 from www.liebertpub.com at 07/28/22. For personal use only.

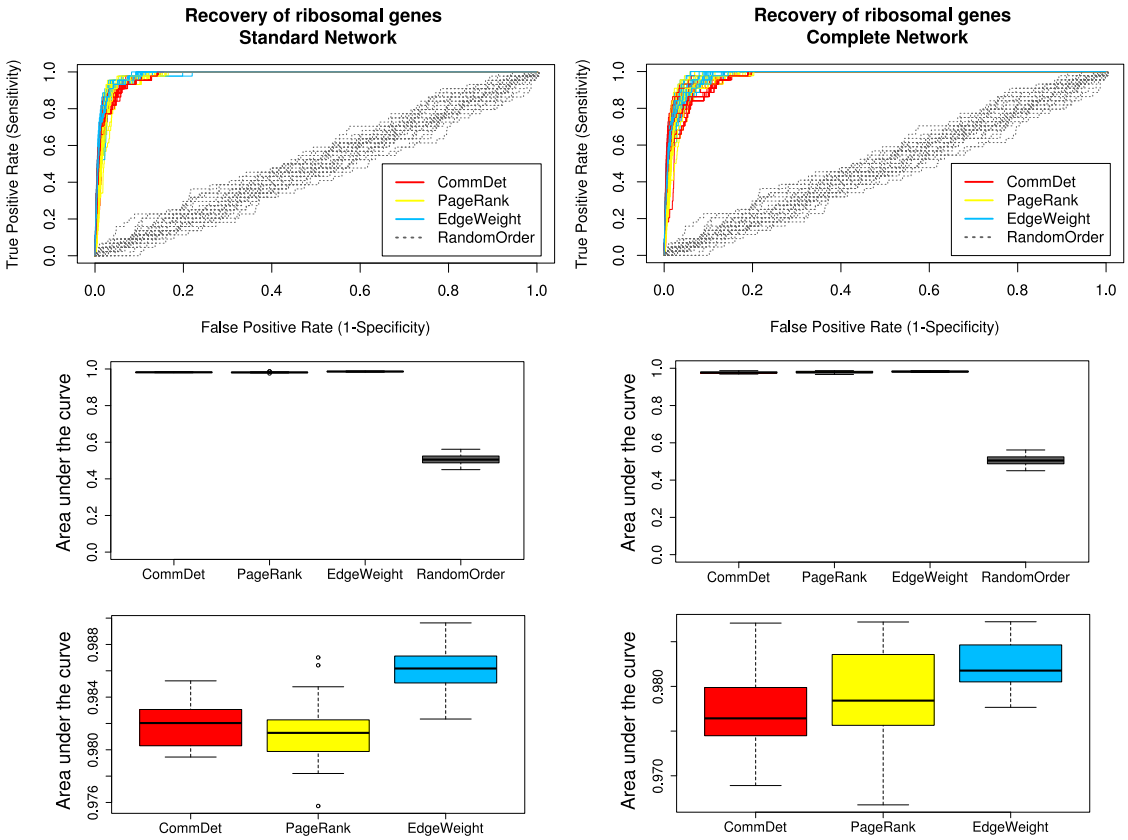


Figure 4.1: Scores obtained using the control seed list. All scores perform significantly better than random. There is no clear ranking between the scores.

Genes selected using the ribosomal genes as seed lists

Standard Network

	CommDet	PageRank	EdgeWeight
RL0883	207	6	122
RL1022	15	96	5
RL1023	37	200	9
RL1260	383	2	235
RL1261	476	1	332
RL1499	294	3	157
RL1551	8	203	11
RL1559	6	26	24
RL1691	42	51	17
RL1719	5	547	411
RL1757	47	38	52
RL1763	10	30	4
RL1766	4	22	13
RL1767	1	125	150
RL1772	11	98	36
RL1948	3	179	1
RL2062	84	49	49
RL2222	2	59	2
RL2228	21	64	10
RL2239	33	93	37
RL2291	14	86	15
RL2573	163	7	79
RL2606	849	10	480
RL2697	13	154	7
RL3333	81	46	64
RL3540	51	75	43
RL3617	19	136	8
RL4007	29	43	18
RL4065	25	9	230
RL4261	9	40	31
RL4295	442	8	354
RL4407	23	143	6
RL4617	460	4	493
RL4634	384	5	231
RL4715	7	73	3

Complete Network

	CommDet	PageRank	EdgeWeight
pRL100260	403	10	302
pRL120235	407	5	356
pRL120521	89	44	83
pRL120522	48	37	17
RL0883	158	3	53
RL0926	31	73	6
RL0927	32	62	8
RL1022	37	57	13
RL1023	12	89	12
RL1260	852	2	441
RL1551	15	124	10
RL1559	93	49	73
RL1691	22	36	3
RL1709	38	83	24
RL1710	52	25	66
RL1719	8	833	634
RL1757	18	14	14
RL1763	9	32	7
RL1766	3	19	31
RL1767	1	589	474
RL1772	4	29	42
RL1798	69	26	51
RL1948	33	88	4
RL2222	2	40	2
RL2239	39	87	37
RL2526	383	4	383
RL2686	725	6	301
RL2829	622	7	344
RL2964	75	51	46
RL3333	20	34	20
RL3540	53	15	32
RL3617	10	81	18
RL4407	5	91	1
RL4408	6	33	9
RL4409	19	72	19
RL4617	645	8	392
RL4634	916	1	417
RL4635	640	9	358
RL4715	7	60	5

RNA polymerase
Translation elongation factor
DNA-binding protein
Without precise function

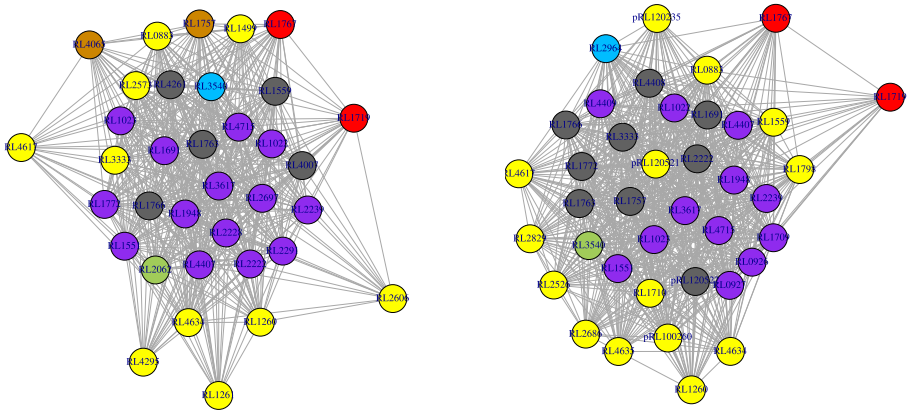


Figure 4.2: Top: the genes which are selected using $M = 50, S = 20$, and $T = 100$, for the standard network and the complete network. The colours refer to the score used, and the shade reflects the rank. The symbols next to the gene names indicate functional annotation. **Bottom:** the induced subgraphs using only these sets of genes. The colour shows for which scores the genes are in the top $M = 50$

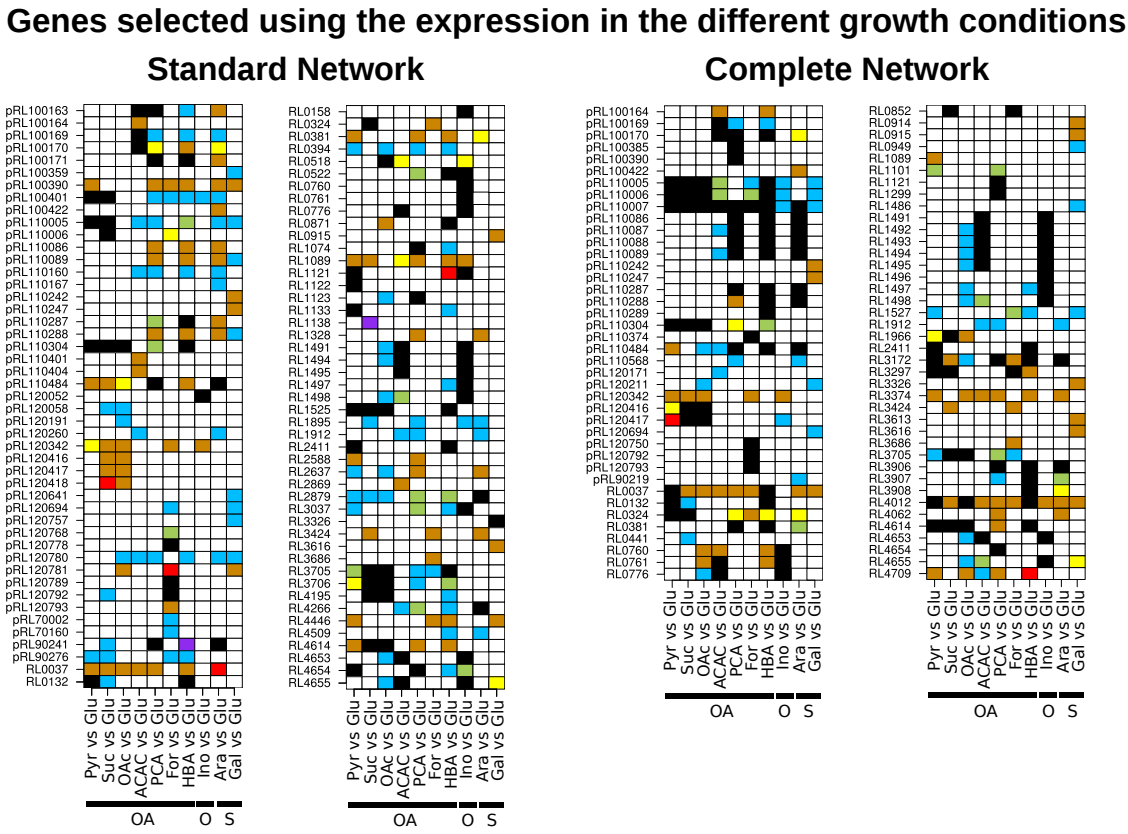


Figure 4.3: Genes included in Q when $T = 20$ and $S = 5$. The colours indicate for which scores they are in the top $U = 50$ for the different conditions. OA, O, and S denote Organic Acid, Other, and Sugar respectively.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Extracting information from Gene Coexpression Networks of R*hizobium leguminosarum*
Publication Status	Published
Publication Details	Pardo-Diaz J, Beguerisse-Díaz M, Poole PS, Deane CM, Reinert G. Extracting Information from Gene Coexpression Networks of *Rhizobium leguminosarum*. J Comput Biol. 2022 Jul;29(7):752-768. doi: 10.1089/cmb.2021.0600. Epub 2022 May 19. PMID: 35588362.

Student Confirmation

Student Name:	Javier Pardo Díaz (Javier Pardo-Diaz)		
Contribution to the Paper	Conceptualisation (equal), data curation (lead), formal analysis (lead), investigation (lead), methodology (lead), software, validation (lead), visualization (lead), writing—original draft (lead), and writing—review and editing (equal). The Poisson approximation in Section 3.4 and SI D was developped by Professor Gesine Reinert.		
Signature		Date	9th September 2022

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Professor Gesine Reinert		
Supervisor comments This is an accurate description.		
Signature 	Date	September 20, 2022

This completed form should be included in the thesis, at the end of the relevant chapter.

5

Analysis of gene expression in the bacteroid

5.1 Overview

R. leguminosarum experiences large changes when it infects the plant. NifA is the central, essential activator of nitrogen fixation across most studied symbiotic organisms (Dixon and Kahn, 2004). It is primarily involved in the final stage of bacteroid differentiation (Hirsch and Smith, 1987). Key targets of NifA include the nitrogenase (*nif*) genes found across nitrogen-fixing species. FnrN is a another key regulator of nitrogen fixation in several species, but not essential in Rlv3841 bacteroids, according to (Wheatley *et al.*, 2020). Nevertheless, to our knowledge, there is still not a clear explanation for all the metabolic changes taking place in the bacteroid.

In this chapter, we aim to study those genes that are highly upregulated in the bacteroid, compared to free-living conditions (bacteria grown in liquid

cultures in the laboratory). For this purpose, we use RNA-Seq data from both conditions in bacteroid and free-living, to identify the genes which are highly upregulated in the bacteroid.

We also identify those transcriptional regulators that are essential in the bacteroid but not in previous life-cycle stages (Wheatley *et al.*, 2020). We expect the selected regulators to be responsible for the transcriptional changes occurring in the bacteroids, which lead to the metabolic changes mentioned in Section 1.4.2. To study the function of these genes in the bacteroid in more detail, for this work, we construct knock-out mutants and use them to infect *P. sativum* (pea) plants. Nine of the mutants are integration mutants, in which the gene of interest has been interrupted by the insertion of a transposon sequence. The remaining mutants are markless mutants already available in the lab.

We evaluate the mutants' ability to fix nitrogen *in planta*. We also extract the RNA from the bacteroids and compare the level of expression of all the genes in the mutants with their expression levels in the wild type in order to identifying which genes they are regulating.

The link between this chapter and Chapters 2 – 4 is relatively week since the selection of the knocked out transcriptional regulators and the trascriptomic analysis of the mutants do not involve the use of gene coexpression networks. However, we apply the methods developed in Chapter 4 to the gene coexpression networks constructed following the methodology described in Chapter 3 to characterise the genes highly coexpressed with the knocked out transcriptional regulators. This additional information results useful for improving our knowledge of bacteroid metabolism.

5.2 Material and methods

5.2.1 Bacterial strains

Bacterial strains used in this work are listed in 5.1. All strains were stocked in 15% glycerol and stored at –80°C.

Strain	Species	Description	Source
Rlv3841	*R. leguminosarum*	*Rhizobium leguminosarum* bv. *viciae*, derivative of strain 300, Str^R	Johnston and Beringer (1975)
OPS1944	*R. leguminosarum*	pK19 mutant of pRL120694 in Rlv3841 background	This work
OPS1977	*R. leguminosarum*	pK19 mutant of pRL120291 in Rlv3841 background	This work
OPS1948	*R. leguminosarum*	pK19 mutant of pRL120397 in Rlv3841 background	This work
OPS1949	*R. leguminosarum*	pK19 mutant of pRL100133 in Rlv3841 background	This work
OPS1950	*R. leguminosarum*	pK19 mutant of pRL100212 in Rlv3841 background	This work
OPS1978	*R. leguminosarum*	pK19 mutant of RL0727 in Rlv3841 background	This work
OPS2033	*R. leguminosarum*	pK19 mutant of RL0652 in Rlv3841 background	This work
OPS1602	*R. leguminosarum*	pK19 mutant of pRL90058 in 3841gusA background	Brandon Ford, Poole Lab
OPS1604	*R. leguminosarum*	pK19 mutant of RL0149 in 3841gusA background	Brandon Ford, Poole Lab
OPS1737	*R. leguminosarum*	Rlv3841 with a markerless deletion of pRL100196	Paul J Rutten, Poole Lab
OPS2676	*R. leguminosarum*	Rlv3841 with a markerless deletion of pRL80079	Carolin Schulte, Poole Lab
LMB351	*R. leguminosarum*	Integration into Rlv3841 disrupting pRL120695	Graham Hood, Poole Lab
LMB400	*R. leguminosarum*	pK19 mutant of RL2924 in Rlv3841 background	Graham Hood, Poole Lab
CR538	*R. leguminosarum*	Tn5 mutant of RL3426 in Rlv300	Alan Downie

Table 5.1: Bacterial strains used in this chapter

5.2.2 Expression data and gene coexpression networks

In this chapter, we use two different Rlv3841 RNA-Seq datasets:

- Bacteroid RNA-Seq: A RNA-Seq dataset obtained in this work from bacteroids from three-week-old pea plants, following the methods described in Sections 5.2.4.2 and 5.2.4.4. It contains 17 samples: three corresponding to wild type (WT) bacteroids and 14 corresponding to mutants (one sample per mutant).

- Free-living RNA-Seq: A RNA-Seq dataset obtained from WT bacteria grown in 20mM succinate 10mM NH_4^+. It contains three samples. Dr Raphael Ledermann (Prof Poole Lab) carried out the RNA extraction from the bacterial growth cultures. Unpublished

We detail how the RNA was extracted in Section 5.2.4.4. Both the datasets were processed following the pipeline described in Section 5.2.5.

We use a microarray dataset (three replicates) that compares the expression in four-week-old bacteroids (sample) against 20mM succinate 10mM NH_4^+ (control, free-living conditions). This dataset was published in Karunakaran *et al.* (2009). For the microarray dataset, as well as for the RNA-Seq dataset, we calculate the fold-change expression values (FC) of the genes as the quotient of their expression in the bacteroid divided by their expression in free-living conditions.

In this chapter, we work with the complete network presented in Chapter 4, constructed using the methodology described in Chapter 3.

5.2.3 Identification of transcriptional regulators

We select 14 Rlv3841 genes that fulfil the following requirements:

- They are essential or defective in the bacteroid, accordingly to Wheatley *et al.* (2020).

- They are neutral or advantaged in free-living conditions and in the rhizosphere, accordingly to Wheatley *et al.* (2020).

- They encode for transcriptional regulators, accordingly to the annotation in Young *et al.* (2006).

- The encoded protein presents at least one DNA-binding motif, assessed with ScanProsite (De Castro *et al.*, 2006).

The selected genes are shown in Table 5.2.

Gene ID	**Gene Name**	**Gene Description**	**FL**	**R**	**B**
pRL80079		putative DeoR family transcriptional regulator (repressor)	NE	NE	ES
pRL90058	*rmrR*	putative TetR family transcriptional regulator of RND family efflux transporter (regulator of rmrAB. Homologous with region in R. etli)	NE	NE	ES
pRL100133		putative IclR family transcriptional regulator	NE	NE	DE
pRL100196	*nifA*	transcriptional regulator NifA	NE	NE	DE
pRL100212		putative GntR family transcriptional regulator			DE
pRL120291		putative MarR family transcriptional regulator	NE	NE	DE
pRL120397		putative GntR family transcriptional regulator	NE	NE	ES
RL120694		putative LacI family transcriptional regulator (repressor)	NE	NE	DE
pRL120695		putative TetR family transcriptional regulator of RND family efflux transporter	NE	NE	DE
RL0149		putative XRE family transcriptional regulator	NE	NE	DE
RL0652		putative LacI family transcriptional regulator (repressor)	NE	NE	DE
RL0727	*rem*	putative two-component sensor/regulator; transcriptional regulator	NE	NE	DE
RL2924		putative MarR family transcriptional regulator	NE	NE	ES
RL3426	*dctD*	putative two-component sensor/regulator; transcriptional regulator C4-dicarboxylate transport(sigma-54)	NE	NE	DE

Table 5.2: Selected transcriptional regulators. The last three columns show whether the genes are essential (ES, in red), defective (DE, in purple) or neutral (NE, in black) in free-living conditions, the rhizosphere and the bacteroid, according to Wheatley *et al.* (2020).

5.2.4 Experimental methods

5.2.4.1 Mutant Construction with pK19mob

We use the pK19mob mutagenesis system (Schäfer *et al.*, 1994) to generate nine mutants with specific disruptions of the chosen genes (see Table 5.2). For this purpose, we (1) obtain a pK19mob mutagenic plasmid containing a fragment of the gene of interest, (2) transform *E. coli* with the plasmid, and (3) conjugate the plasmid from *E. coli* into Rlv3841. In Rlv3841, the plasmid is integrated by homologous recombination into the genome, disrupting the gene of interest. This system is well-established for use in multiple *Rhizobium* species (e.g. Sudtachat *et al.* (2009); Karunakaran *et al.* (2010); Becker *et al.* (1993)) and the insertion mutations generated are structurally similar to those generated in INSeq studies (see 1.4.4).

Generation of the pK19mob mutagenic plasmid

We purify the pK19mob vector (containing a kanamycin resistance cassette and lacZα) and digest it with HindIII (Thermo Scientific), according to the manufacturer's instructions. We also amplify an internal fragment (300 – 500 base pairs) of the target gene from the WT strain genomic DNA with PCR using Phusion® polymerase (New England Biolabs) and primers containing extensions in their 5' end: TGATTACGCCAAGCT (forward primer) and GCAGGCATGCAAGCT (reverse primer). Afterwards, the amplified gene fragment is cloned into the digested pK19mob vector using InFusion HD cloning, according to the manufacturer's instructions. The inserted fragment disrupts the lacZα reading frame in the vector (i.e. the protein is no longer produced).

Transformation of *E. coli* with pK19mob mutagenic plasmid

We thaw on ice 50 µL of chemically competent *E.coli* DH5α cells. We add 5 µL of plasmid to the cells and incubate them on ice for 30 min. We then heat-shock at 42°C for 45 seconds, followed by 2 min on ice. Afterwards, we add 400 µL SOC medium (Hanahan, 1983) and incubate at 37°C and 200 rpm for 1 hour. To select the cells containing the correct construct (i.e. the plasmid containing the gene fragment instead), we plate all cells on LB·Km_{20}·$Xgal_{40}$ agar: non-transformed

bacteria cannot grow in media with kanamycin; bacteria transformed with the empty plasmid are blue (LacZα complements the inactive β-galactosidase in the cell, which transforms Xgal into a blue product); bacteria transformed with the right construct are white (they lack LacZα). We sequence the plasmids using the pK19A and pK19B primers to verify the insert.

Conjugation to transfer the plasmid pK19mob mutagenic plasmid from *E.coli* to *R.leguminosarum*

We grow the recipient strains of Rlv3841 (WT) on TY·Str_{500} agar (Beringer, 1974) plates at 28°C for 72 hours prior to conjugation, and the *E. coli* plasmid donor strains and *E. coli* pRK2013 helper strain on LB·Km_{20} agar plates at 37°C overnight before the day of conjugation. We transfer the three strains to a TY agar plate without antibiotics, spread them in a 2 – 3 cm diameter circle in the centre of the plate, and grow the conjugation plates overnight at 28°C. We plate the bacteria on TY·Str_{500}·Neo_{240} agar plates and incubate at 28°C for 72 hours to select for rhizobial conjugants. We map the mutagenesis using mapping colony PCR and sequence the PCR product for verification.

5.2.4.2 Plant infection, nodule picking, and regrowning checks

We surface sterilise pea (*Pisum sativum* cv *Avola*) seeds by immersion in 70% ethanol for 45 seconds and then 2% bleach for 5 min, followed by 10 washes with sterile distilled water. We bury the seeds in sterilised 1 L plastic beakers (two seeds per beaker) filled with medium vermiculite and 400 mL of N-free rooting solution (1 mM $CaCl_2{\cdot}2H_2O$, 100 µM KCl, 800 µM $MgSO_4{\cdot}7H_2O$, 10 µM Fe·EDTA, 35 µM H_3BO_3, 9 µM $MnCl_2.4H_2O$, 0.8 µM $ZnCl_2$, 0.5 µM $Na_2MoO_4{\cdot}2H_2O$, 0.3 µM $CuSO_4{\cdot}5H_2O$, 3.67 µM KH_2PO_4, and 3.52 µM Na_2HPO_4).

We resuspend the cultures of the strains to be inoculated, grown on TY·Str_{500}·Neo_{240} agar slopes at 28°C for 72 hours, in 4 mL of sterile distilled water. In sterile conditions, we inoculate the beakers with 2 mL of the resuspended culture (only one strain per beaker) and seal the beakers with cling film afterwards. We carry out the inoculation assays in six different batches (3 of the batches are for RNA-Seq,

the other 3 are for Acetylene Reduction Assays, ARA); Table 5.3 indicates the distribution. Each batch includes two beakers in which we add sterile water instead of bacterial culture (water controls) and three beakers inoculated with the WT strain.

Batch	**Strains***	**Reps**	H_2O	**Experiment**
1	pRL120694, pRL120291, pRL120397, pRL120397, pRL100212, WT	3	2	ARA
2	RL0652, pRL100196, pRL120695, RL0149, RL2924, WT	3	2	ARA
3	RL3426, pRL80079, pRL90058, RL0727, WT	3	2	ARA
4	pRL120694, pRL120291, pRL120397, pRL120397, RL0727, pRL100212, WT	3	2	RNA-Seq
5	RL0652, pRL100196, pRL120695, RL0149, RL2924, WT	3	2	RNA-Seq
6	RL3426, pRL80079, pRL90058, WT	3	2	RNA-Seq

Table 5.3: Batches of plants prepared for follow-up experiments, either acetylene reduction assays (ARA) or RNA-Seq. In the "Strains*" column we name the different mutants using the name of the gene that has been knocked out. The "Reps" column indicates how many beakers are prepared per strain. The "H_2O" column indicates the number of control plants in the batch (i.e. plants that have not been infected).

After seven days the cling film is cut open and we discard the smaller pea seedling. After 21 days, we harvest the plants. Then, we pick five nodules from each plant and surface-sterilise them with 2% bleach for five min, followed by six washes in sterile distilled water. In sterile 1.5 mL tubes, we ground the nodules with sterile plastic pestles. Using the pestles, we streak TY plates and incubate them at 28°C for three days. Afterwards, we select five single colonies from each plate, restreak them on TY·Str_{500}·Neo_{240}, and incubate over three days at 28°C to check that bacteria have not lost the resistance to neomycin. We map three colonies from each nodule to check through PCR that the bacteria still have the mutation and also that there has not been cross-contamination.

5.2.4.3 Acetylene reduction assays

We harvest the plants from batches 1, 2, and 3 in Table 5.3 to carry out acetylene reduction assays (ARA). We place the plants in 500 mL Duran bottles with some wet tissue paper that we close with a rubber septum lid. We draw 15.25 mL of air from each bottle using a syringe and then inject 12.2 mL of acetylene. After one hour, we draw three samples of 1 mL from each bottle and inject them into the Perkin Elmer Clarus® 480 Gas Chromatograph. We measure the ratio between the areas under the acetylene and ethylene peaks and calculate the moles of acetylene reduced. We use the acetylene reduction as a proxy for nitrogenase activity (Shah *et al.*, 1975).

5.2.4.4 RNA extraction and sequencing

We harvest the plants from batches 4, 5, and 6 in Table 5.3 to extract the RNA from the bacteroids in the nodules. Firstly, we pick the nodules in the roots of the plants (see Figure 5.3) and freeze them using liquid nitrogen; we store the frozen nodules in 1.5 mL tubes at –80°C (each tube containing the nodules from a plant). We grind the frozen nodules with 5 mL of filter sterilised (0.22 µm filter) isolation buffer (sucrose 0.73 mM, K_2HPO_4 8 mM, KH_2PO_4 2 mM, and $MgCl_2$ 5 mM in distilled water) (Karunakaran *et al.*, 2009). We centrifuge the mix in a 15 mL Falcon tube (1,000 rpm for 10 min at 4°C) to remove plant material. Then, we centrifuge the supernatant in 1.5 mL tubes (6,000 rpm for 5 min at 4°C) to separate the bacteroids. We resuspend the pellets in 500 µL of 10 mM Tris-Cl pH 8.0 and add the bacteroid suspension into pre-cooled autoclaved Ribolyser tubes filled with 300 mg of 0.1 mm silica beads and 100 mg of 0.1 mm glass beads containing 500 µL of RLT Plus buffer (Qiagen RNeasy Plus Mini Kit) and 10 µL of β-mercaptoethanol. We use a ribolyser (FastPrep FP120, Thermo FisherScientific) to lyse the bacteroids (speed 6.5 for 30 seconds with three repeats with 30 seconds intervals between the repeats) and then incubate them on ice for 3 min. We centrifuge at 13,000 rpm for 3 min and transfer the supernatant to a gDNA eliminator column (Qiagen RNeasy Plus Mini Kit). We continue the RNA purification using the Qiagen RNeasy Plus Mini Kit, according to the manufacturer's instructions, and elute the RNA in

50 µL of warm RNAse-free water. We remove the genomic DNA in the samples using the invitrogene TURBO DNAse kit, following the manufacturer's instructions. Lastly, we use the Zymo Research RNA Clean & Concentrator Kit, according to the manufacturer's instructions, and eluting in 15 µL of warm RNAse-free water. Our 17 samples were sent to Novogene for transcriptomic sequencing.

5.2.5 RNA-Seq analysis

As a preliminary check, we use Bowtie 2 (Langmead and Salzberg, 2012) to obtain the number of reads mapping to the Rlv3841 genome, to the Rlv3841 rRNA, and to the pea genome. Then, we map the data for each library against the Rlv3841 genome using EDGE-pro (Magoc *et al.*, 2013). EDGE-pro uses Bowtie 2 to map the reads to the genome and calculates the frequencies per nucleotide. EDGE-pro calculates the number of reads and RPKM value for each gene feature in the genome including mRNA, rRNA, and tRNA. Afterwards, we construct a matrix containing the TPM values for all the genes according to the latest genome annotation for all the samples. The annotation files required for the reads mapping were created by former Poole's lab member Dr Vinoy K. Ramachandran.

We use DESeq2 (Love *et al.*, 2014) to calculate the differential expression between the bacteroid and free-living conditions, using as input values the number of reads for each gene in the three WT bacteroid samples and three free-living samples. We assess the correlation between the FC values obtained using RNA-Seq and those obtained using the microarray dataset (Karunakaran *et al.*, 2009). For the RNA-Seq dataset, we compute the log2FC values and select those genes with a log2FC value higher than 2.5 and with an expression in the bacteroid higher than the median expression. We consider these genes to be (*genes upregulated expressed in the bacteroid*).

We study the function of each mutated transcriptional regulator we calculate the FC values for all the genes by dividing the TPM values in the mutant by the TPM values in the WT. For each mutant, we use the WT that was in the same batch of plants (see Table 5.3). We consider a gene to be upregulated in the mutant if its FC value is higher than three and its expression in the mutant

bacteroid is higher than the median expression. Likewise, we say that a gene is downregulated in the mutant if its FC value is lower than 1/3 and its expression in the WT bacteroid is higher than the median expression.

5.3 Results and discussion

5.3.1 Identification of genes which are differentially expressed in the bacteroid

We compare the gene expression in the bacteroid and in free-living bacteria using the FC values obtained from the RNA-Seq data produced for this work and microarray data from Karunakaran *et al.* (2009) For the RNA-Seq, we observe more extreme FC values, ranging from $2^{-11.74}$ to $2^{15.19}$, whereas for the microarray data the FC values are between $2^{-3.83}$ and $2^{7.22}$. The extreme FC values observed for the RNA-Seq should be analysed carefully since they may correspond to those genes which have very low expression values for at least one of the studied conditions.

A total of 6,656 genes have values for both datasets. The Pearson correlation of the log2FC values of these pairs of genes across the two datasets is 0.54. We observe that the set of highly overexpressed genes (high FC values) is similar for both methods (e.g. 50% of the 100 highest overexpressed genes in the RNA-Seq dataset are among the 100 highest overexpressed genes in the microarray dataset; 42% of the 200 highest overexpressed genes in the RNA-Seq dataset are among the 200 highest overexpressed genes in the microarray dataset). We observe that the genes that are not included in the overlap of the top 200 highest expressed genes in the bacteroid have a lower expression in the bacteroid (according to the RNA-Seq data) than those present in the overlap (see Figure 5.1).

For the microarray dataset, there are 112 genes with a log2FC value greater than 2.5 and an expression value in the bacteroid higher than the median expression. For the RNA-Seq dataset, we observe 468 genes with an expression in the bacteroid higher than the median expression and a log2FC value higher than 2.5. Out of this

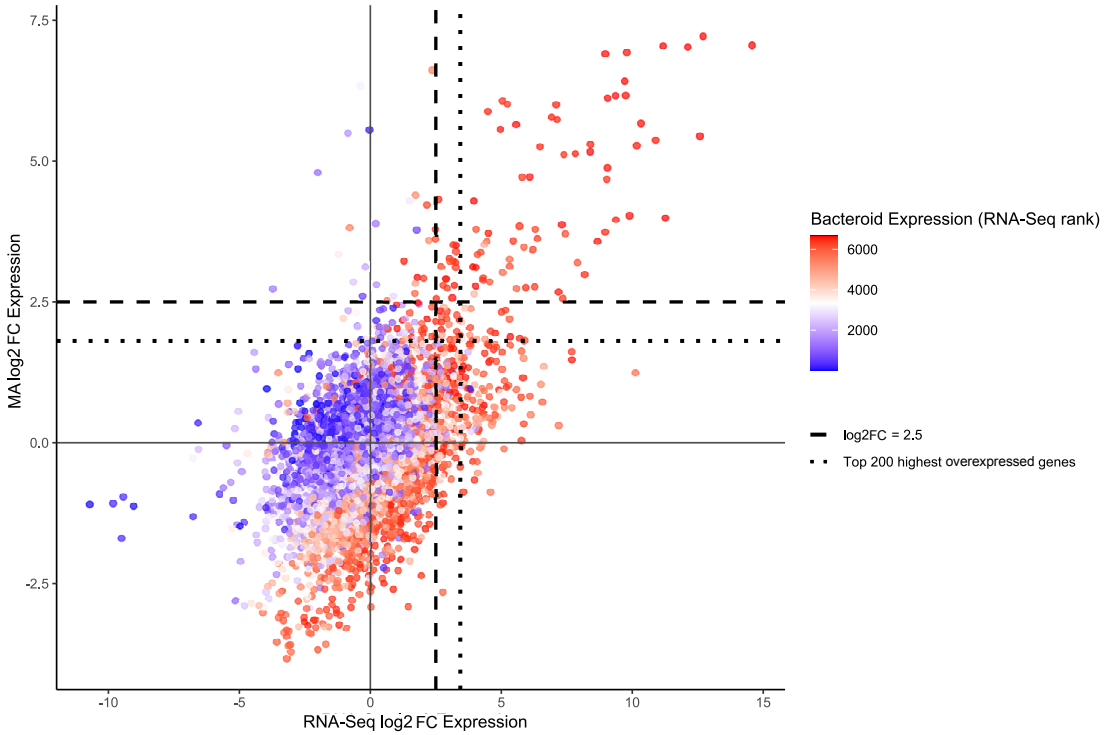


Figure 5.1: Scatter plot that represents the FC expression values in the bacteroid compared to free-living conditions of 6,656 genes obtained using the RNA-Seq and microarray datasets. The x-axis indicates the log2FC values according to the RNA-Seq dataset. The y-axis indicates the log2FC values according to the microarray dataset. Each point represents a different gene. The colour of the points indicates the expression in the bacteroid as a rank from dark blue (the least expressed gene) to red (the most expressed gene). Dashed lines indicate the value log2FC = 2.5. The dotted lines indicate the log2FC values for the 200th-highest overexpressed genes.

set, 383 genes are included in the microarray dataset, and 382 appear in our complete gene coexpression network. Figure 5.2 represents the induced subgraph of these 382 genes. We observe that 347 of them are in the same component, suggesting a high correlation in the expression of the genes (i.e. highly expressed in the bacteroid, low expression in other conditions). The genes which are not included in the largest connected component show lower values in the bacteroid than the average value.

In the subgraph, we observe a cluster of genes with high expression in the bacteroid (blue nodes in Figure 5.2) as well as high differential expression values (larger nodes in the figure); all the genes with FC values higher than 200, except RL2270, pRL100122, and pRL100144 are in this cluster. Most of these genes are involved in nitrogen fixation, respiration, storage compound biosynthesis and

metabolism (e.g. *fixABCX*, *fixNOPQ1*, *nifH*, *nifA*, *nifB*, *acsA3*, *fdxB1*, *nifN*, *nifE*, *phaC2*, *phaE*, *nifK*, *nifD*). Other genes with high expression and FC values are pRL100099 pRL100100, pRL100101, pRL100103, pRL100104, pRL100119, pRL100120, pRL100201, pRL100202, all of them under a NifA-dependent promoter and already identified in Karunakaran *et al.* (2009) (see *nifA* mutant results); some of these genes are pseudogenes and are not included in our gene coexpression network.

We present the 50 genes with the highest FC values according to the RNA-Seq data in Table E.1 in Appendix E. There are four genes with high expression and FC values that do not appear in the network in Figure 5.2 because they did not appear in the previous genome annotation. These genes are RL_RS39930 (hypothetical protein, downstream pRL100101), RL_RS28420 (hypothetical protein, located between fixP1 and fixG1), RL_RS28350 (putative nitrogen fixation protein NifT), and RL_RS38450 (4Fe-4S binding protein); the latter two genes are adjacent in the genome and located between pRL100193 and pRL100195 ($nifB$).

There are multiple factors that might explain the discrepancies between the RNA-Seq and the microarray results, all of them reported in different studies, for example, some probes might not hybridise perfectly or might be cross-reactions and RNA-Seq is superior to microarrays in detecting low abundance transcripts (Zhao *et al.*, 2014). Other reasons might be related to differences in the experimental set-up (i.e. how the plants are infected and harvested and how RNA is extracted). RNA-Seq is more powerful than microarrays as we do not need to know the sequence, there is no saturation limit, it allows us to detect low-expressed genes, and there are no hybridisation problems. Nevertheless, Figure 5.1 suggests that we can use microarrays as a good proxy for expression studies, especially when attending to the highest expressed genes.

5.3.2 Acetylene Reduction Assays and phenotypic analysis of mutants

Following the protocol described in Section 5.2.4.3, we test the ability of the mutants in Table 5.2 to fix nitrogen during symbiosis with pea plants. We discard

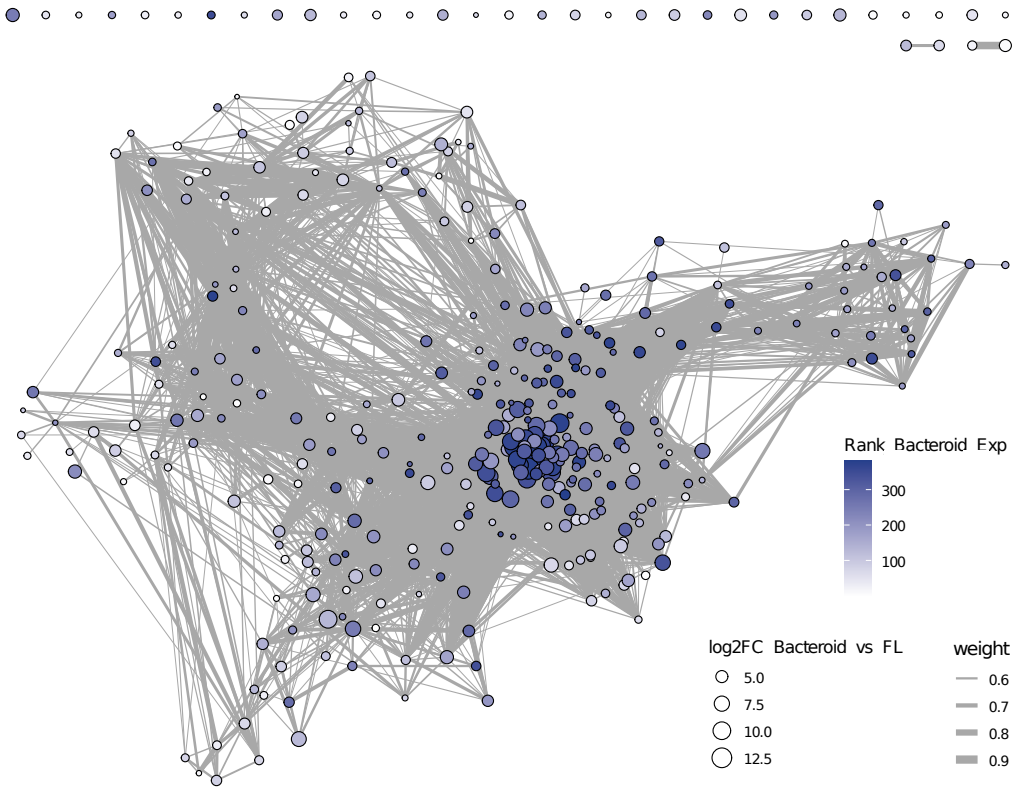


Figure 5.2: Gene coexpression network subgraph. It includes all the genes that are overexpressed in the bacteroid, compared to FL conditions in the RNA-Seq data (log2FC > 2.5 and with an expression in the bacteroid higher than the median expression). There are 382 genes. The size of the nodes indicates the log2FC values (the bigger, the more upregulated in the bacteroid). The colour indicates the expression in the bacteroid as a rank from dark blue (highest) to light blue (lowest). The width of the edges indicates the correlation in the expression of the genes (the thicker, the higher the coexpression).

the plants infected with RL3421 (*dctD*) mutants (see Section 5.3.4). We observe that all the strains are able to fix nitrogen at the same rate as the WT, except the pRL100196 (*nifA*) mutant. All the plants have a phenotype similar to the control plants infected with the WT (i.e. green healthy plants with red nodules) except those plants infected with the $nifA$ mutant, which are smaller, yellower, and with small white nodules (see Figure 5.3).

5.3.3 Transcriptomic analysis of the mutants

We analyse the RNA-Seq data with Bowtie 2 to count the number of reads mapped to the Rlv3841 genome. We observe that for all our samples, except the sample obtained from the RL3426 mutant (see Section 5.3.4), we obtain more than 7 million

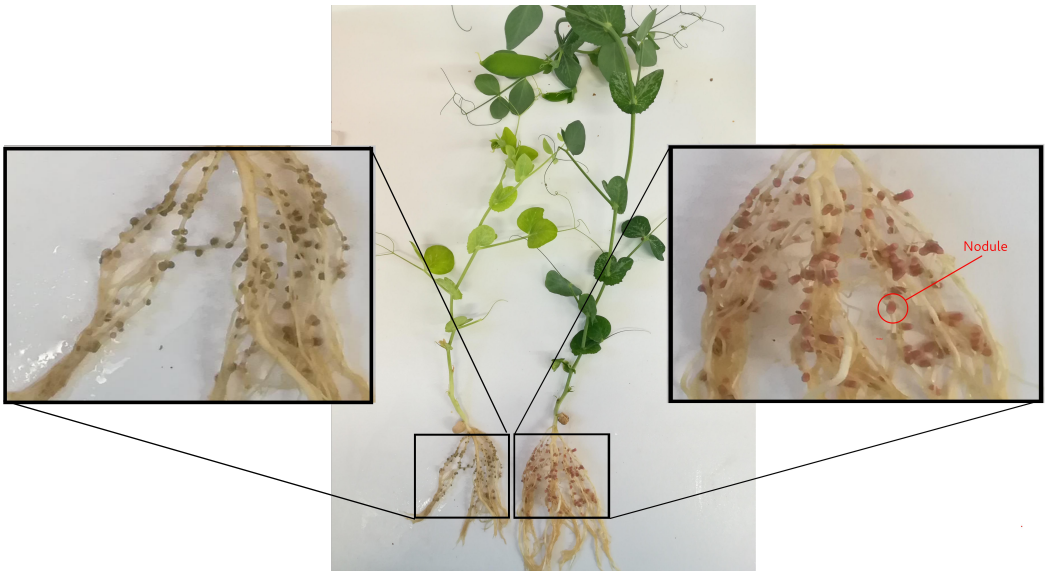


Figure 5.3: Pea plants 21 days after inoculation. The plant on the left was inoculated with a knock-out mutant of *nifA*. This gene is essential in the bacteroid and for nitrogen fixation. The nodules in this plant do not fix nitrogen, which affects plant growth. The plant on the right was inoculated with the WT strain, which fixes nitrogen. This second plant looks much healthier.

reads mapping to the Rlv3841 genome, excluding the reads mapping to rRNA. Table 5.4 indicates the results of this analysis.

Strain	% Reads Rlv3841	% Reads rRNA Rlv3841	% Reads Pea	#Reads	#Reads Rlv3841
pRL80079	73.24	0	17.55	12894762	9444124
pRL90058	73.4	0	17.97	12212516	8963987
pRL100133	76.78	0	16.84	10741743	8247510
pRL100196	95.18	0	2.89	19223797	18297210
pRL100212	85.57	0.1	9.63	12624147	10802483
pRL120291	79.3	0	14.48	13471313	10682751
pRL120397	75.79	0	15.6	10299280	7805824
pRL120694	82.78	0	11.72	14360927	11887975
pRL120695	77.01	0	16.96	11348506	8739484
RL0149	78.93	0	15.52	8102790	6395532
RL0652	82.75	0	11.97	11998657	9928889
RL0727	81.02	0	13.04	11835922	9589464
RL2924	79.24	0	14.45	12027869	9530883
RL3426	45.27	0	23.97	11569347	5237443

Strain	% Reads Rlv3841	% Reads rRNA Rlv3841	% Reads Pea	#Reads	#Reads Rlv3841
WT (1)	85.21	0	10.58	13667213	11645832
WT (2)	78.45	0	15.69	10811965	8481987
WT (3)	87.58	0	7.8	12426376	10883020

Table 5.4: Results of the analysis of the RNA-Seq samples using Bowtie 2. For each sample, we obtain the number of reads, the number of reads mapped to Rlv3841 genome, the number of reads mapped to Rlv3841 rRNA, and the number of reads mapped to the pea genome. We remove from the analysis the sample obtained from the RL3426 mutant (see Section 5.3.4).

In the next sections, we present the transcriptomic analysis for the transcriptional regulators in Table 5.2, excluding RL3426 and pRL120695 (see Section 5.3.4). For each regulator, we identify the genes they are coexpressed with, according to the gene coexpression network; Table E.2 in Appendix E includes the genes coexpressed with each of the regulators. Afterwards, we present which genes are highly differentially expressed in the mutant bacteroids compared to the WT, focusing on those genes that are upregulated in the bacteroid compared with the free-living conditions (i.e. those represented in 5.2); Table E.3 in Appendix E includes the genes which are differentially expressed at each of the mutants. We combine all this information to hypothesise the role of the studied transcriptional regulators.

5.3.3.1 pRL80079

The gene pRL80079 codes for a putative DeoR family transcriptional regulator (repressor); these regulators typically act as repressors of genes encoding metabolic enzymes, preventing transcription of catabolic genes when their substrates are not present in the cell (Hunter *et al.*, 2012). When analysing the genes which are highly coexpressed with pRL80079 (Table E.2), we observe two genes with a score in the top 50 for all three methods: RL3481 (putative MaoC family protein; MaoC proteins are involved in polyhydroxyalkanoate (PHA) biosynthetic pathways (Wang *et al.*, 2013)) and RL4651 (putative solute binding-protein of ABC peptide

transporter). We also identify two genes with a score in the top five for two of the methods: RL2880 (putative transcriptional regulator with GGDEF sensory box domain with diguanylate cyclase activity) and RL4627 (*mtgA*; biosynthetic peptidoglycan transglycosylase). We observe that pRL80079 is slightly upregulated in the bacteroid compared to free-living conditions (log2FC = 1.60).

In the mutant, we observe a cluster of adjacent genes (pRL80080 – pRL80095) that are overexpressed compared to the WT, with FC values higher than 5 for all of them and higher than 100 for some of them (see Table E.3). According to Vanderlinde *et al.* (2014), in Rlv3841, pRL80083-pRL80080 and pRL80084 – pRL80088 form a gene cluster that is required for the transport and metabolism of L-homoserine and that is upregulated by L-homoserine compared to mannitol. Homoserine is a major component in pea root exudates (Van Egeraat, 1975). In Vanderlinde *et al.* (2014), they confirmed that the upregulation of the cluster in Vincent media (Vincent *et al.*, 1970) with root exudates and that the expression in the bacteroid is similar to the expression in non-induced free-living conditions. In the same study, the authors carried out a luminescence assay to show that in a pRL80079 mutant, the expression of pRL80083 – pRL80080 and pRL80084 – pRL80088 increases compared to the WT when grown in Vincent's media with mannitol, suggesting that pRL80079 is a negative regulator that represses expression of the homoserine gene cluster operons in the absence of L-homoserine.

The results we obtain are in line with those in the literature in that we observe an increase in the expression of the pRL80083 – pRL80080 and pRL80084 – pRL80088 clusters in the pRL80079 mutant. In addition, we observe that pRL80090 – pRL80095 are also overexpressed in the mutant. These genes are predicted to code for a putative fucose metabolism operon protein (pRL80090) whose loss reduces the fitness of the bacteria in the rhizosphere (Wheatley *et al.*, 2020) and components of an ABC carbohydrate uptake transporter (pRL90092 – pRL90095). The region pRL80059 – pRL80095 has lower G/C content than the average G/C content for pRL8 (Vanderlinde *et al.*, 2014); therefore, it is likely that these genes are

related to the other ones and controlled by the same transcriptional regulator (namely, pRL80079).

The pRL80079 mutant is predicted to be growth defective in the bacteroid according to Wheatley *et al.* (2020). As we do not observe an effect on nitrogen fixation, the INSeq output might be a result of a decrease in the competitiveness of the mutant. In the mutant, there is an upregulation of the homoserine metabolism cluster, altering the bacteroid metabolism and potentially making adapting to the host more difficult. In the mutant, we also see an upregulation of RL0451 (cold shock protein), which supports the idea that the bacteria are metabolically compromised, resulting in elevated stress levels. We also observe a decrease in the expression of nitrogen limitation-related transporters (RL3737 – RL3741), which may have an effect on competitiveness.

We observe that two of the genes upregulated in the bacteroid compared to free-living conditions (those included in Figure 5.2) are downregulated in the pRL80079 mutant: RL1106 (*pspA*; putative phage shock protein A regulator) and RL2912 (conserved hypothetical protein). To our knowledge, these genes are not involved in nitrogen fixation or bacteroid development.

5.3.3.2 pRL90058

The gene pRL90058 (*rmrR*) codes for a putative TetR family transcriptional regulator that positively controls *rmrAB* (pRL90059 and pRL90060) (Ramos *et al.*, 2005), which code for the membrane fusion protein and the inner membrane protein components of an efflux transporter, respectively. The expression of pRL90058 is higher in the bacteroid than in free-living conditions (log2FC = 2.60); *rmrAB* are also overexpressed in the bacteroid compared to free-living conditions (log2FC = 3.8 and 1.17). When analysing the genes that are coexpressed with pRL90058 using our gene coexpression network, we observe that all the genes with a score in the top 50 for all three methods (pRL110324, RL3867, and RL4099) and those with a score in the top 10 for two of the methods (pRL110321, RL4099, and RL4535) do not have a functional annotation (i.e. they are annotated as “hypothetical protein”

or similar descriptions). The only exception to this trend is RL3273, which is annotated as a putative von Willebrand factor; nevertheless, Whittaker and Hynes (2002) suggests that these factors are not required for prokaryotic existence and are acquired through horizontal gene transfer from eukaryotes.

In the pRL90058 mutant, we observe a decrease in the expression of *rmrAB* (FC = 0.51 and 0.20), supporting the idea that *rmrR* is a positive activator for *rmrAB*. González-Pasayo and Martínez-Romero (2000) propose that, in *R. etli*, *rmrAB* encodes an inducible export system that prevents the accumulation of toxic plant compounds within the bacteria cell. Our pRL90065 mutant, with reduced expression of *rmrAB*, might not be able to export some toxic plant compounds, affecting the competitiveness of the bacteria and therefore explaining the INSeq phenotype assigned to pRL90058 in (Wheatley *et al.*, 2020) (growth defective). These results agree with those presented in Ford (2021), where it is shown that *rmrR* mutant forms nodules and fixes nitrogen at the same level as the WT strain but is less competitive than the WT strain. Another efflux pump was shown to be involved in the symbiosis between *Bradyrhizobium japonicum* and soybean (Lindemann *et al.*, 2010), reinforcing the idea that the transporter RmrAB plays an important role in root colonisation. None of the genes upregulated in the WT bacteroid compared to WT free-living conditions are differentially expressed in the pRL90058 mutant. In the mutant, a secondary effect of the dysregulation is an increase in the stress level of the bacteria, reflected by the overexpression of the cold shock protein as well as a decrease in the expression of nitrogen limitation-related transporters (RL3737 – RL3741).

5.3.3.3 pRL100133

The gene pRL100133 codes for a putative IclR family transcriptional regulator that is downregulated in the bacteroid compared to free-living conditions according to the RNA-Seq data (log2FC = –1.96). According to Prof Poole's microarray database, pRL100133 is overexpressed in succinate/GABA media compared to nitrogen-limiting succinate/ammonia media (FC = 5.92) and in succinate vs TY

media (FC = 7.3). When analysing the genes pRL100133 is coexpressed with, using our gene coexpression network, we identify several genes (see Table E.2). We highlight RL2509 (*citZ*, a putative citrate synthase I), which has the highest score for all three methods and is upregulated in nitrogen-limited conditions (FC = 8.26), succinate against glucose (FC = 3.88), and acetoacetate against glucose (FC = 4.58) and downregulated in 7-day and 28-day bacteroids compared to free-living conditions (FC = 0.192 and FC = 0.201). RL3654 (*pssD*) codes for a putative polysaccharide biosynthesis protein essential in the bacteroid and nodule bacteria (Wheatley *et al.*, 2020). The scores obtained for RL3654 are in the top four for all three methods; RL3654 is downregulated in 15-day bacteroids compared to free-living conditions (FC = 0.38). pRL110198 (hypothetical protein) scores are among the top eight values for all three methods and it is as well downregulated in 15-day bacteroids compared to free-living conditions (FC = 0.38). Other coexpressed genes, with scores in the top 50 for all three methods or in the top 10 for at least two methods, are pRL100072 (putative solute binding-protein of ABC polar amino acid transporter upregulated in protocatechuate against pyruvate ammonia (FC = 4.55)), pRL100255 (conserved hypothetical protein), pRL110488 (putative solute binding-protein of ABC carbohydrate uptake transporter upregulated in protocatechuate against pyruvate ammonia (FC = 8.15), succinate/GABA against succinate/ammonia (FC = 5.30), glucose/glutamate vs glucose/aspartate (FC = 6.57), and nitrogen-limited conditions (FC = 3.66) and whose homologue in *S. meliloti* is induced by galactosamine), pRL120637 (putative transmembrane protein), RL1012 (*cbpA*, putative curved DNA-binding protein), and RL3251 (conserved hypothetical protein).

In the pRL100133 mutant, compared to the WT strain, we observe an upregulated cluster of adjacent genes (pRL100127 – pRL100136). pRL100134 – pRL100136 code for orthologues of *A. tumefaciens* AatKLM. pRL100134 (*aatK*) is upregulated in pea rhizosphere one, three, and seven days post inoculation (FC = 3.04, 4.87, and 3.28, respectively). pRL100135 and pRL100136 (*aatLM*) are upregulated in succinate/GABA compared to succinate/ammonia (FC = 10.00 and 8.59). AatKLM

in *A. tumefaciens* are involved in the degradation of gamma-butyrolactone to succinic acid (Chai *et al.*, 2007). The transcription of the *attKLM* operon is repressed by AttJ, an IclR-type transcription factor encoded by an adjacent, divergent gene (Zhang *et al.*, 2002). Our results suggest that pRL100133 encodes the repressor for *aatKLM* in Rlv3841. To check this hypothesis, we compute the protein identity between pRL100133 and AttJ in *A. tumefaciens* using Blast (Altschul, 1993). We observe that both proteins have 59.7% of identity; this result backs our hypothesis.

In Rlv3841, as in *A. tumefaciens*, the expression of plasmid transfer genes (*tra* genes) requires the accumulation of N-(3-oxooctanoyl)-L-homoserine lactone (3OC8HSL) (Lithgow *et al.*, 2000). In *A. tumefaciens*, AttM inactivates 3OC8HSL by hydrolysis of its homoserine lactone ring. In the pRL100133 mutant, AttM is overexpressed, causing the quick degradation of lactones. N-acylhomoserine lactones play an important role in cell communication (quorum sensing). These molecules, produced by a synthase gene (e.g. *cinI* or *rhiI* in Rlv3841), accumulate in the environment depending on bacterial cell density and are sensed by the bacteria community. Once a threshold concentration is reached, this triggers the expression of certain bacterial genes involved in all sorts or processes, such as DNA transfer, virulence, and biofilm production. The dysregulation in the mutant might have an effect on plasmid transfer, affecting competition during early infection. These results suggest that 3OC8HSL signalling might play an important role during plant infection and that the timing for switching lactone degradation off affects competitiveness (i.e. high-rate degradation of lactones might hinder the adaptation to the nodule environment). To confirm this hypothesis, competition assays are required.

In the pRL100133 mutant, the cluster pRL100127 – pRL100131 is also upregulated. These genes code for the components of a branched-chain amino acid ABC transporter. pRL100130 and pRL10031 are slightly upregulated in succinate/GABA vs succinate/ammonia (FC values 6.39 and 5.43). Future research will focus on the role of this transporter.

We observe that genes that are upregulated in the mutant are not upregulated in the WT bacteroid (where the repressor pRL100133 is downregulated) compared to

free-living conditions. This suggests that even a low expression of the regulator (6.15 TPM in the bacteroid) still has an effect on the regulation of the expression of its target genes or that there is an additional regulator controlling the expression of these genes.

Lastly, out of the genes that are highly upregulated in the WT bacteroid compared to the free-living conditions, RL3066 (putative transmembrane protein) is the only one that is differentially expressed in the pRL100133 mutant, with an FC value is 0.09. To our knowledge, this gene has never been described to be involved in nitrogen fixation or bacteroid development.

5.3.3.4 pRL100196

NifA, encoded by pRL100196, is the central and essential activator of nitrogen fixation across most studied symbiotic and non-symbiotic diazotrophs (Dixon and Kahn, 2004). NifA is primarily involved in the final stage of bacteroid differentiation (Lazazzera *et al.*, 1993). One of the main targets of NifA is the $nifHDK$ operon that codes for the nitrogenase components, together with other nif genes involved in the biosynthesis of the Fe-S clusters. Karunakaran *et al.* (2009) found six copies which are similar or identical to the $nifH$ upstream region (promoter) in the Rlv3841 genome; this promoter is recognised by NifA, activating their expression. They checked that mutants for several of those genes, leaving apart nif genes, were able to fix nitrogen.

In our mutant, we observe 48 downregulated genes, compared to the WT (genes with FC $< 1/3$ and an expression in the expression WT bacteroid higher than the median expression). All of the genes under the control of promoters which are similar to the $nifH$ promoter identified in Karunakaran *et al.* (2009) are downregulated in our mutant, confirming that those genes are regulated by NifA. One of these operons is pRL10094 – pRL10098; Karunakaran *et al.* (2009) reported that this operon is not essential for nodule development and nitrogen fixation. Most of the downregulated genes we observe in the mutant are well-known genes involved in nitrogen fixation, such as nif and fix genes. Out of the set of 468

genes upregulated in the WT bacteroid compared to free-living conditions (see Figure 5.2), 31 genes are highly downregulated (FC < 0.25) in the $nifA$ mutant, compared to the WT bacteroid. This result suggests that NifA might be involved in the activation of such a set of genes.

We observe 3,096 upregulated genes in the mutant compared to the WT bacteroid. This effect may be a result of secondary effects of the unavailability of fixing nitrogen and developing functional nodules. It might as well be due to artefacts from the sequencing: $nifA$ is very abundant in the WT (1809.3 TPM), as well as all the NifA-controlled genes that are highly downregulated in the mutant; therefore, the TPM values the genes that are expressed in the mutant are likely to be higher than for the WT strain even if their expression does not change. In order to correct this effect, we shall, as future work, repeat the RNA sequencing by adding a known concentration of a set of synthetic RNA spike-ins to RNA samples before library preparation, adapting the protocol described in Blackburn *et al.* (2019). These sequences can be used to measure technical biases and to act as internal quantitative and qualitative controls throughout the sequencing workflow.

Among the downregulated genes, we highlight the cluster of genes pRL110020 – pRL110025, all of them in the same predicted operon and with FC values between 87.4 and 334.32. pRL110020 – pRL110022 code for the components of a nitrate transporter NitT. This result is interesting since there should not be nitrate in the nodules and suggests that these genes might have another function. pRL110023 code for a putative FAD-dependent oxidoreductase. pRL110024 and pRL110025 (*caiD*) code for a putative acyl-CoA transferase and a putative carnitinyl-CoA dehydratase, respectively. In the mutant, we also observe the downregulation of chaperonins, such as groLch1; these proteins favour correct protein folding. In *B. japonicum*, one copy of the chaperonin groESL is NifA regulated (Fischer *et al.*, 1993) and in (Fischer *et al.*, 1999) they claim that GorESL chaperones (also involved in protein folding) are essential for nitrogen fixation. Our results suggest that in Rlv3841 chaperonins might also be regulated by NifA.

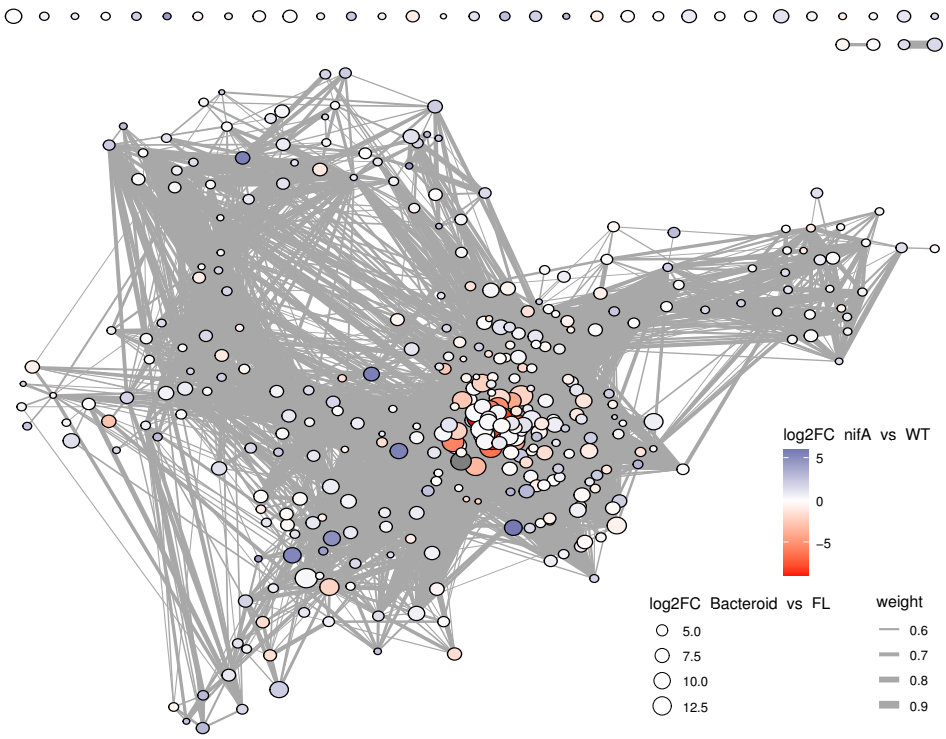


Figure 5.4: Gene coexpression network subgraph. It includes all the genes that are overexpressed in the bacteroid, compared to free-living conditions in the RNA-Seq data (log2FC > 2.5 and with an expression in the bacteroid higher than the median expression). There are 382 genes. The size of the nodes indicates differential expression in the WT bacteroid compared to the free-living conditions (the bigger, the more upregulated in the bacteroid). The colour indicates the differential expression in the $nifA$ mutant bacteroid compared to the WT bacteroid from dark blue (genes highly upregulated in the mutant) to red (genes highly downregulated in the mutant). The width of the edges indicates the correlation in the expression of the genes (the thicker, the higher the coexpression).

Figure 5.4 is an induced subgraph of the genes that are upregulated in the bacteroid compared to the free-living conditions (log2FC > 2.5 and expression in the bacteroid higher than the median expression) using the same layout as in Figure 5.2 for easier comparison. The size of the nodes indicates the differential expression of the genes in the bacteroid compared to the free-living conditions. The colour of the nodes indicates the differential expression of the genes in the $nifA$ mutant bacteroid compared to the WT bacteroid. We observe that all the genes that are highly downregulated in the mutant (red nodes) are in the central cluster that includes the genes which are most highly upregulated in the bacteroid, compared to the free-living conditions.

5.3.3.5 pRL100212

The gene pRL100212 encodes a putative GntR transcriptional regulator that is upregulated in the bacteroid (log2FC = 2.76). This gene does not appear as differentially expressed for any condition in the microarray datasets analysed in this work and it is not included in our gene coexpression network.

In the pRL100212 mutant, we do not observe a big specific effect in upregulation (i.e. high FC values). In contrast, a large number of genes which are described as important for competition in Mendoza-Suárez *et al.* (2021) are moderately upregulated (FC values lower than 7): genes involved in the degradation of rhamnose (pRL110407 – pRL110416), *nod* genes (pRL100175, pRL100183, pRL100185, and pRL10087), proline transporters (pRL100080 and pRL100081), *glnII* (RL3549), and *phnG* (RL0177). We propose that none of those genes is directly regulated by pRL100212 (in fact, the rhamnose catabolism genes have their own regulator RhaR and the nod genes are regulated by NodD) and that the upregulation effects we observe in the mutant might be a secondary effect of a major dysregulation process taking place at some point during the initial stages of root colonisation. This dysregulation affects the competitiveness of the bacteria, explaining the INSeq value (growth defective) assigned to pRL100212 in the bacteroid in Wheatley *et al.* (2020). In order to identify the direct targets of pRL100212, future research will focus on analysing the transcriptome of the mutant strain during plant colonisation.

A cluster of genes (RL4454 – RL4457) is highly downregulated in the mutant (FC values between 0.1 and 0.16). RL4454 (putative solute binding-protein of an ABC polyamine/opine/phosphonate uptake transporter) and RL4455 (*speB*2; putative agmatinase) are both included in the set of genes which are highly upregulated in the bacteroid, as well as RL2912 (conserved hypothetical exported protein). This result suggests that pRL100212 might be involved in the activation of these genes in the bacteroid.

5.3.3.6 pRL120291

pRL120291 is a putative MarR family transcriptional regulator that is downregulated in the bacteroid (log2FC = –1.36). MarR family DNA binding proteins are typically repressors, and repression is alleviated by exposure to a ligand (Alekshun and Levy, 1997). Regarding the genes coexpressed with pRL120291, according to the gene coexpression network, we highlight pRL110465, pRL120738, RL2882, and RL2891. pRL110465 is a hypothetical protein that is essential in free-living conditions, rhizosphere and root (Wheatley *et al.*, 2020) and is upregulated in nitrogen-limited conditions (FC = 4.26). pRL120728 is a transmembrane protein that has the highest score for two of the network analysis methods we employ and is upregulated in succinate against TY media (FC = 6.90). The scores for RL2885 and RL2891 (putative acetyltransferase and putative cationic efflux system protein) are among the seven highest values for two of the three methods and they are not differentially expressed in any of the analysed microarrays. Other coexpressed genes are pRL120196 (putative fructose-biphosphate aldolase), pRL120238 (*ilvD*; putative dihydroxy-acid dehydratase), RL0233 (*optF*; putative aminotransferase), RL0586 (putative transmembrane protein), RL0658 (conserved hypothetical protein), RL2481 (putative transmembrane GGDEF sensory box protein), and RL2535 (putative rhomboid family transmembrane protein).

In the pRL120291 mutant, we observe an upregulation of a cluster of genes (pRL120287-pRL120290) with FC values between 40.45 and 2781.81. This cluster is downregulated in the WT bacteroid, compared to free-living conditions. pRL120287 is the permease component of an ABC carbohydrate uptake transporter induced by galactose, arabinose, and fucose in *S. meliloti* (Mauchline *et al.*, 2006). The other components of the transporter (pRL120285 and pRL120285) are slightly upregulated in the mutant (FC = 2.27 and 5.04). pRL120288 is a gluconolactonase involved in glucose and glucose-1P degradation and sorbitol biosynthesis, according to BioCyc (Karp *et al.*, 2019). pRL120289 is a choline dehydrogenase. pRL120290 is a hydrolase (Karp *et al.*, 2019). Our results suggest that pRL120291 is a repressor of the mentioned genes. We propose that pRL120288 catalyses the carboxylic ester

bond in a lactonase, releasing a complex choline molecule that is used as a substrate by pRL120289. One of the side chains in the choline might be a sugar imported by the transporter (e.g. arabinate). The upregulation of this metabolic pathway in the bacteroid might have an undesired effect on the adaptation to the new environment. Future research is needed to assess the competitiveness of the mutant; assays using the mutant and the WT strain in growth media containing a lactonase with the mentioned structure will assist in confirming the molecular function of pRL120291.

None of the genes which are highly upregulated in the bacteroid compared to the free-living conditions is downregulated in the pRL120291 mutant.

5.3.3.7 pRL120397

pRL120397 is a putative GntR family transcriptional regulator that is slightly downregulated in the bacteroid (log2FC = -1.55). pRL120397 is downregulated in the rhizosphere, according to the microarray data available. Attending to the gene coexpression network, we observe that pRL120397 is coexpressed with several genes that are essential or defective in all environments analysed in Wheatley *et al.* (2020). We highlight the cluster of genes pRL110628 – pRL110630 (*cobHIJ*) and RL2829 (*cobO*); all of them are involved in precorrin metabolism. Other coexpressed genes which are essential or defective according to the INSeq dataset are RL1638 (*plsX*; putative fatty acid/phospholipid synthesis cluster protein), RL1783 (*rpsQ*; ribosomal protein), RL3309 (*murD*; putative UDP-N-acetylmuramoylalanine-D-glutamate ligase likely to be related to cell wall biosynthesis), RL3422 (*greA*; transcription elongation factor) and some genes without functional annotation (RL1763 and RL4734) These coexpression results suggest that pRL120397 is a constitutive gene in free-living conditions.

In the mutant, we observe an overexpression of its neighbouring genes (pRL120380 – pRL120396) when compared to the WT. pRL120381 – pRL120385 and pRL120390 – pRL120393 code for the components of two carbohydrate uptake transporters, respectively. According to BioCyc (Karp *et al.*, 2019), one of these transporters might be involved in sn-glycerol-3-phosphate uptake. pRL120386 and pRL120387 (*dhaLp*12

and *dhaKp*12) code for a dihydroxyacetone kinase. pRL120389, pRL120394, and pRL120395 code for a GFO/IDH/MocA oxidoreductase, a semialdehyde dehydrogenase (Karp *et al.*, 2019), and a four-carbon acid sugar kinase (Karp *et al.*, 2019), respectively. pRL120396 codes for a putative ribulose bisphosphate carboxylase subunit; orthologous genes in photosynthetic rhizobia, as *Bradyrhizobium sp.*, are required for efficient symbiotic nitrogen fixation (Gourion *et al.*, 2011). However, Rlv3841 is not photosynthetic, so there is no Calvin Cycle where ribulose bisphosphate carboxylase could play a role; thus, to our knowledge, this gene is not essential for nodule formation and nitrogen fixation in Rlv3841. RL0451 (*cspA*2; putative cold shock protein) and RL4565 (*glnK*; putative nitrogen regulatory protein PII) are also upregulated in the mutant.

The available annotation suggests that the upregulated genes are involved in a catalytic pathway, starting with the importing of glycerol-3-phosphate by the transporter (pRL120390 – pRL120393). Then, this molecule is transformed into dihydroxyacetone and then phosphorylated (pRL120386 – pRL120387) and degraded in the glycolysis pathway. The activation of this catabolic process might have an effect on nitrogen metabolism, explaining the upregulation of *glnK*. We propose that pRL120397 is a repressor of the upregulated cluster and that it is only activated in the presence of its substrate (as derived from the coexpression analysis, the repressor is constitutively expressed). The removal from the repressor has an effect on the fitness and competitiveness of the bacteria. To confirm these hypotheses, in future work, we shall analyse the transcriptome of the WT strain in growth media with glycerol-3-phosphate as a carbon source, as well as carry out a competition assay.

Lastly, we observe that RL3891 (conserved hypothetical protein), RL4454 (putative solute binding-protein of ABC transporter polyamine/opine/phosphonate uptake transporter), and RL4455 (*speB*2; putative agmatinase) are downregulated in the mutant compared to the WT bacteroid. All these three genes are included in the set of genes which are highly overexpressed in the bacteroid, compared to free-living conditions.

5.3.3.8 pRL120694

pRL120694 is a putative LacI family transcriptional regulator (repressor). Like most LacI family regulators, it is likely that it inhibits the expression of genes coding for proteins involved in the metabolism of sugars when the sugar is not available, ensuring that the bacterium does not produce the enzymes. Once the sugar becomes available, the repressor stops binding the DNA to bind to the sugar, allowing the transcription of the genes (Jacob and Monod, 1961). According to the microarray collection, pRL120694 is slightly upregulated in 7-day-old bacteroid, compared to the free-living conditions (FC = 2.57). Using the coexpression network, we observe that pRL120694 is highly coexpressed with RL3887 (it retrieves the highest score for all the methods), pRL120362, and pRL100298 (scores among the highest seven values for all the methods). RL3887 (*sitD*; putative permease component of ABC manganese/zinc/iron chelate uptake transporter) is upregulated in 7-day-old bacteroid too (FC = 3.13) and its homologue in *S. meliloti* is induced by manganase limitation. pRL120362 (*katG*; putative catalase-peroxidase) is upregulated in 7-day and 21-day-old bacteroid. pRL100298 is a conserved hypothetical protein that is not differentially expressed in any of the analysed microarrays. Other genes highly which are coexpressed with pRL120694 are RL0425 (*ptsN*; nitrogen regulatory IIA protein defective in the bacteroid and nodule bacteria), RL1302 (*osmC*2; putative OsmC-like organic hydroperoxide resistance protein), RL2508 (*gltA*2, putative cytrate synthase II), RL2604 and RL3441 (conserved hypothetical proteins), and RL3869 (*rLub*; putative ribosomal large subunit pseudouridine synthase B).

When attending to the differential expression values between the pRL120694 mutant and the WT strain, there is a highly overexpressed cluster (pRL120685 – pRL120691) with FC values from 28.67 (pRL120690) to 269.31 (pRL120688). All these genes show a low expression in WT bacteroids, according to the RNA-Seq data. pRL120688, which is upregulated in root juice (FC = 3.78), codes for a putative polygalacturonase that might be involved in the degradation of the pectin in plant cell walls. It is reported in the literature that some phytopathogens excrete polygalacturonases into the plant host as a mean to weaken the pectin network

and acquire nutrients (De Lorenzo and Ferrari, 2002). Ljunggren and Fåhraeus (1961) conclude that polygalacturonase plays an important part in the root infection process by weakening the cell wall of the root hair and facilitating bacterial invasion. pRL120685, pRL120689, pRL120690, and pRL120691 are components of ABC carboydrate uptake transporters that might be responsible for the intake of the sugars obtained from the plant; pRL120690 and pRL120691 are upregulated in root juice. pRL120687 is an unsaturated glucoronyl hydrolase that catalyses the hydrolytic release of unsaturated glucuronic acids from oligosaccharides that might be involved in the digestion of sugars.

Our results suggest that pRL120694 is a repressor of the genes in the cluster and that the mutation triggers their expression, which affects the competitiveness of the mutated strain. A constitutive expression of these genes may lower the fitness due to unnecessary degradation of the plant cell wall. In addition, a high rate of degradation of the plant cell wall might affect the formation of the infection thread and provoke the plant defence reactions, as the plant senses Rlv3841 as a pathogen. Future research should evaluate the performance of our pRL120694 mutant in competition assays.

None of the upregulated genes in the bacteroid, compared to the free-living conditions, are differentially expressed in the pRL120694 mutant.

5.3.3.9 RL0149

RL0149 is a putative XRE family transcriptional regulator that is not differentially expressed in the bacteroid but upregulated in TY media against succinate NH_4^+. According to the scores derived from the gene coexpression network, we find that pRL120448 (putative AraC family transcriptional regulator) has the highest or second highest score for two of the three methods employed. pRL100455 (putative conserved hypothetical ATPase), pRL120131 (putative solute binding-protein of ABC transporter PepT), RL0101 (*gabD*2; putative succinate-semialdehyde dehydrogenase), RL0523 (putative purine deaminase, RL1690 (*lon*; putative ATP-dependent protease), RL1998 (putative permease component of ABC carbohydrate

uptake transporter slightly upregulated in phenylalanine/pyruvate media (FC = 2.3)), RL2398 (*uvrA*; putative component of DNA repair by nucleotide excision complex UvrABC system), RL3212 (putative alcohol dehydrogenase), RL3498 (conserved hypothetical protein), RL4180 (putative efflux transporter), and RL4560 (putative cNMP-binding protein) have a score among the top highest 50 scores for each method or a score among the top ten highest scores for two methods.

RL0149 is highly upregulated in mutants where the repressor PraR (RL0390) is knocked out (Frederix, 2010). PraR mutants show a hyper attachment phenotype thought to be due to the overproduction of exopolysaccharides (EPS) and cell-surface attachment factors which are no longer repressed by PraR, with this leading to greater biofilm formation, root attachment and competitiveness against WT (Frederix *et al.*, 2014). Ford (2021) assessed the competitiveness of an RL0149 mutant and observed that its competitive nodulation ability was significantly impaired compared to the WT. Those results confirmed that RL0149 is involved in the induction of the EPS and cell-surface attachment factors that PraR represses and that the disruption of RL0149 leads to defective attachment and reduced competitiveness for nodulation.

In our RL0149 mutant, we observe a slight upregulation (FC = 3.18) of RL0991 (*matB*; malonyl-CoA synthetase involved in the transformation of malonate into acetyl-CoA). To our knowledge, there is no obvious explanation for why an increase in the production of acetyl-CoA from malonate has an effect on competition. pRL100427 (putative permeate component of ABC transporter upregulated in root attachment assays and whose loss increases the fitness of the bacteria in the rhizosphere and root (Wheatley *et al.*, 2020)) is also upregulated in the mutant (FC = 3.35).

We also observe that pRL100169 (*rhiA*) and pRL100170 (*rhiB*) are downregulated in the mutant compared to the WT strain (FC = 0.19 and 0.13). These two genes are highly induced in the rhizosphere and are repressed by flavonoids that induce *nod* genes, but their biochemical role, to our knowledge, has not been established (Cubo *et al.*, 1992; Economou *et al.*, 1989). We suggest that the

deletion of RL0149 has a drastic effect on the rhizosphere and that the differentially expressed genes that we observe in the bacterioid might be a result of a side effect. Future research will focus on the analysis of the transcriptome of the RL0149 in the rhizosphere.

None of the genes which highly overexpressed expressed in the bacteroid compared to the free-living conditions are differentially expressed in the RL0149 mutant.

5.3.3.10 RL0652

RL0652 is a putative LacI family transcriptional repressor that is slightly downregulated in the bacteroid (log2FC = –0.77). According to the microarray collection, RL0652 is upregulated in protocatechuate media compared to pyruvate (FC = 4.4). From the coexpression network analysis, we observe that, unlike for other genes, there is considerable agreement between the scores retrieved for the three methods used in the network analysis. This agreement suggests that there is a high correlation between the selected genes that are identified by all three methods. We highlight those genes whose scores are among the 10 highest scores for the three methods: RL0586 (putative membrane protein), RL0634 (putative short-chain dehydrogenase), RL0990 (*matA*; putative malonyl-CoA decarboxylase), RL1008 (putative transmembrane GGDEF/EAL sensory box protein), RL1339 (conserved hypothetical protein), and RL1363 (putative oxidoreductase). We also observe some coexpressed genes that are essential or defective to all analysed conditions in Wheatley *et al.* (2020): pRL100046 (putative transposase-related protein), RL0404 (*mviN*; putative transmembrane virulence factor homologue), RL0960 (conserved hypothetical protein), RL0973 (*dxs*; putative 1-deoxy-d-xylulose-5-phosphate synthase), and RL1014 (*pdxH*; putative pyridoxamine 5'-phosphate oxidase), and genes that are essential in the bacteroid (Wheatley *et al.*, 2020): RL0690 (putative chemotaxis protein methyltransferase), RL1032 (*rnhA*1; putative ribonuclease HI), and RL1503 (putative RNA binding protein).

In the RL0652 mutant, we observe an overexpression of RL0646-RL0650. These genes code for metabolic enzymes: putative aldolase (RL0646), putative GFO/IDH/MocA

family oxidoreductase (RL0647; downregulated in protochatechuate media), putative D-lyxose/D-mannose family sugar isomerase (RL0648), putative PfkB family sugar kinase (RL0649), and putative sorbitol dehydrogenase (RL0650). RL0648 is a conserved hypothetical protein. Our results suggest that RL0652 is likely to be regulating the expression of this operon by repressing it. This operon might be involved in sorbitol catabolism. However, according to Oresnik *et al.* (1998), the genes which are required for growth in sorbitol are in pRL11. In this study, they also showed that mutants which are unable to catabolise sorbitol were not affected in their competitive ability, suggesting that sorbitol catabolism genes might not be necessary during root colonisation. In the case that our mutant has constitutively induced the sorbitol catabolism genes due to the lack of the repressor, that might results in a loss of competitiveness during the early stages of infection compared to other strains, thus explaining the INSeq values in Wheatley *et al.* (2020). In the mutant we also observe a decrease in the expression of pRL100169 (*rhiA*), backing our hypothesis about the mutation having an effect during colonisation. To confirm these results, in future work, we shall study the transcriptome of the mutant in the rhizosphere, as well as assess the sorbitol metabolism rate in the mutant.

None of the genes which are highly upregulated in the bacteroid compared to the free-living conditions is differentially expressed in the RL0652 mutant.

5.3.3.11 RL0727

RL0727 encodes Rem, which is part of a putative two-component sensor/regulator that positively regulates the expression of *flaA*, *flaB*, *flaC*, *flaD*, *motA*, *motB*, *che*1, and *mcpD* (Tambalo *et al.*, 2010), all of them involved in cell motility. The expression of rem is positively regulated by the dimer VisNR; in Rlv3841, VisNR and Rem are expressed throughout the exponential phase but expression drops sharply during the stationary phase, although it is still sufficient for continued motility (McCarter, 2006). rem downregulated in the bacteroid (log2FC = –2.76), and upregulated in nitrogen-limiting conditions (FC = 3.60), according to our microarray collection. In the WT strain, *rem* is coexpressed with several genes involved

in cell motility too: RL0212 (*cpaA*; putative pilus assembly protein), RL0703 (*motA*; putative chemotaxis motility protein, which retrieves a score among the top five highest scores for all the methods), RL0708 ($flgC$; putative flagellar basal-body protein FlgC), RL0731 ($flaF1$; putative flagellar synthesis related protein), RL0732 ($fltB$; putative flagellar biosynthesis repressor protein that binds to and promotes degradation of flagellar mRNA), RL0733 ($flgD$; putative flagellar hook formation protein), and RL4386 ($cheW2$; putative chemotaxis protein). Other genes which are coexpressed with RL0727 are pRL1110001 ($repAp11$; putative replication protein RepA in pRL11), RL0425 ($ptsN$; putative nitrogen regulatory IIA protein which is one of the major regulators in carbon and nitrogen metabolism (Sánchez-Cañizares *et al.*, 2020)), RL1379 ($rosR$; putative transcriptional regulator involved in nodulation competitiveness, which is essential for growth in the rhizosphere, root attachment and in the bacteroid (Wheatley *et al.*, 2020)), RL1729 ($divK$; response regulator receiver protein defective or essential in free-living conditions, rhizosphere, and root (Wheatley *et al.*, 2020)), RL2508 ($gltA2$; putative citrate synthase II), and RL3150 (conserved hypothetical protein that retrieves the highest and second highest scores for two of the network analysis methods).

In our *rem* mutant, we observe that there is a moderate upregulation (FC values between 3.6 and 6.3) of some genes involved in cell motility: RL0719 ($flaB$; putative flagellin B, whose expression is supposed to be activated by Rem), RL0728 ($flgE1$; putative flagellar hook protein), RL0729 ($flgK1$; putative flagellar hook-associated protein), RL0730 ($flgL1$; putative flagellar hook-associated protein), RL0731 ($flaF1$; putative flagellar hook-associated protein), RL0732 ($fltB$; putative flagellar biosynthesis repressor protein that is coexpressed with Rem), and RL0733 ($flgD$; putative flagellar hook formation protein that is coexpressed with *rem*). These results do not fully agree with the published literature since Rem is known to regulate the activation of motility genes and we expected its deletion to result in a downregulation of these genes. Nevertheless, they might indicate that Rem, besides acting normally as an activator, when it has a low expression (as in the bacteroid, where *rem* is downregulated) has a mild repressor activity. Another hypothesis

that could justify our results is that the pK19 insertion in *rem* might trigger the expression of the adjacent cluster of genes (RL0728 – RL0732); yet, this hypothesis does not explain the upregulation of $flaB$. In future research, we will create a *rem* markless mutant to assess whether the expression values obtained in this study are just a result of the pK19 insertion. Afterwards, this markless mutant shall be complemented with rem under promoters of different strengths to evaluate how the expression of the motility genes changes depending on the expression of *rem*.

In our *rem* mutant, we also observe a moderate upregulation (FC = 3.12) of RL1416 (putative myo-inositol degradation/rhizopine catabolism protein). We do not observe a differential expression for any of the genes which are highly upregulated in the bacteroid compared to the free-living conditions.

5.3.3.12 RL2924

RL2924 is a putative MarR transcriptional regulator. According to the microarray collection, RL2924 is upregulated in the rhizosphere, 21 days post-inoculation (FC = 6.02). Attending the gene coexpression network, we observe that RL2924 is coexpressed with RL2925 ($mutT$; putative mutator protein), which has the highest score for two of the three methods, and the sixth highest score for the third, RL4302 (putative oxidoreductase) that has the highest score for one of the methods, and the second highest for another one, and RL4378 (putative LacI family transcriptional repressor) with second and third highest scores for two of the methods. We also highlight RL2646 (putative transmembrane protein), RL2900 ($dhaK1$; putative dihydroxyacetone kinase), RL2906 ($dhaKa2$; putative PTS-dependent dihydroxyacetone kinase), RL3002 ($dapB2$; putative dihydrodipicolinate reductase), RL3784 (putative efflux transporter, and pRL120277 ($ligI$; putative 2-pyrone-4,6-dicarboxylic acid hydrolase).

In the RL2924 mutant, we observe an overexpression of RL2925 – RL2927. These genes are upregulated in the bacteroid. RL2925 ($mutT$), which is coexpressed with RL2924, shows a mild upregulation (FC = 5.69) compared to the other two genes (FC = 49.08 and 288.23). Martínez-Salazar *et al.* (2009) claim that MutT

might be related to stress in *R. etli*. We propose that in low concentrations, RL2924 acts as a mild repressor for its own operon, which also controls the expression of *mutT*. Therefore, the complete removal of RL2924 causes a low overexpression of *mutT*. This hypothesis is compatible with the information derived from the gene coexpression network (i.e. RL2924 and *mutT* are highly coexpressed). RL2926 is a putative adenylyl cyclase and RL2927 codes for OhrB, a putative organic hydroperoxide resistance protein. Our results suggest that RL2924 acts as a repressor for RL2926 and *ohrB*, both of them in the same operon.

We do not observe a differential expression for any of the genes which are highly upregulated in the bacteroid compared to free-living conditions. However, we observe that in the mutant, the rhizosphere genes *rhiA* and *rhiB* are downregulated, suggesting that the high overexpression of *ohrB* might have an effect on the rhizosphere, probably related to high oxidative stress that may also affect the early stages of the colonisation process, and thus having an effect in competitiveness.

5.3.4 Discarded results

We discard the plants infected with the RL3426 mutant from the acetylene reduction assay and the RNA-Seq analysis since, after analysing the RNA-Seq results we observed that the mutant was in a different background. We found out that the WT parent strain was not Rlv3841 but another *R. leguminosarum* strain. After analysing the RNA-Seq results for the sample obtained from the plant infected with the pRL120695 mutant we observed that the sample was contaminated with RNA from the pRL120291 mutant (we observed the pK19 sequence fused with the pRL120291 coding sequence). As we performed a bacteroid reisolation test after harvesting the plants, the contamination is likely to have taken place during the preparation of the RNA samples for sequencing.

5.4 Conclusions

Bacteroids have a drastically different transcriptome compared to free-living bacteria. Here, we analysed the transcriptome of WT bacteroids and confirmed these differences using high-throughput sequencing approaches. This dataset, which is more accurate than the currently available microarray dataset, will help to address future research questions. We used this dataset to identify those genes which are highly upregulated in the bacteroid, compared to the free-living conditions.

Many of the essential symbiosis genes are induced by low oxygen via either FnrN or NifA. However, many more genes outside these regulons are differentially expressed in bacteroids, highlighting additional roles for other transcription factors in bacteroids. In this work, we addressed several of those transcriptional regulators. We identified 14 transcriptional regulators that are essential in the bacteroid, according to (Wheatley *et al.*, 2020), including NifA. It remains unknown to which degree these transcription factors contribute to symbiosis since all the evaluated mutants, except the $nifA$ mutant, are able to fix nitrogen at the same rate as the WT strain. We were not able to identify which transcriptional regulators (if any) control the expression of the genes which are upregulated in the WT bacteroid since the only mutant that showed a differential expression in a considerable number of upregulated genes, most of them already known, was the $nifA$ knock-out mutant (Rutten and Poole, 2019; Ledermann *et al.*, 2021). The reason the knock-out study produced weak results may be that the original InSeq competition assay was arguably selecting as genes essential in the development of the bacteroid genes which are important in the root colonisation process. Future research will focus on developing new transcriptional regulator mutants (including those that are not essential in the bacteroid, according to (Wheatley *et al.*, 2020), as $fnrN$ for example) to evaluate their transcriptome. Using the gene coexpression network, jointly with the already available transcriptomic data, will help to identify which genes shall be mutated.

We observed that most of the mutated transcriptional regulators regulate the expression of the genes adjacent to them. These observations and our gene coexpression network allowed us to hypothesise for functional annotations. Table

5.5 summarises the main hypothesis drawn from this analysis. We suggest that that most of them might play an important role in root colonisation. Future research will assess the competitiveness of these strains in competitiveness assays as well as in field conditions.

Gene	Hypothesis	Follow-up experiment
pRL80080 – pRL80095	Their overexpression negatively affects competitiveness	Competition assay using the pRL80079 mutant
pRL90059 & pRL90060	Export some toxic plant compounds and their loss negatively affects the competitiveness	Competition assay of the pRL90059 mutant (done in Ford (2021))
pRL100136	Degrades 3OC8HSL and its overexpression negatively affects the competitiveness	Competition assay using the pRL100133 mutant
pRL100212	Important regulatory effect in the rhizosphere	Transcriptomic analysis using the mutant in the rhizosphere and competition assay
pRL120288 – pRL120289	Catabolise lactonases and their overexpression negatively affects the competitiveness	Growth curves in media containing lactonases and competition assay using the pRL120291 mutant
pRL120386 – pRL120387 & pRL120390 – pRL120393	Import and catabolise glycerol-3-phosphate. Its overexpression negatively affects the competitiveness	Growth curve in media containing glycerol-3-phosphate as carbon source and competition assay using the pRL120397 mutant
pRL120386 – pRL120387 & pRL120390 – pRL120393	Import and catabolise glycerol-3-phosphate. Its overexpression negatively affects the competitiveness	Growth curve in media containing glycerol-3-phosphate as carbon source and competition assay using the pRL120397 mutant
pRL120685 – pRL120691	Involved in plant colonisation and plant cell wall degradation. Their overexpression negatively affects the competitiveness	Competition assay using the pRL120694 mutant
RL0149	Important regulatory effect in the rhizosphere	Transcriptomic analysis of the mutant in the rhizosphere and competition assay

Gene	Hypothesis	Follow-up experiment
RL0652	Important regulatory effect in the rhizosphere	Transcriptomic analysis of the mutant in the rhizosphere and competition assay
RL0646 – RL0650	Catabolise sorbitol	Groth curve in media containing sorbitol using the RL0652 mutant
RL0727	Mild repressor of RL0719 and RL0728 – RL0732 when it has a low expression	Trascriptomic analysis of a RL0727 markless mutant, and mutants with RL0727 under the control of promoter with different strengths
RL2924	Mild repressor of RL2925 – RL2927 when it has a low expression	Transcriptomic analysis of RL2924 under the control of promoter with different strengths

Table 5.5: Main hypothesis drawn from the transcriptomic and coexpression analysis of the mutants. All the presented hypothesis will need a follow-up experiment to confirm the results.

6

Analysis of promoter regions of genes which are highly expressed in free-living conditions

6.1 Introduction and objectives

In Chapter 5 we have analysed the role of the transcriptional regulators that are essential or defective in the bacteroid (Wheatley *et al.*, 2020). In this chapter, we focus on those genes that are expressed in free-living conditions. Among these genes, there are genes whose expression remains constant during plant infection, whereas others increase or decrease their expression. We are interested in investigating whether there is a molecular signal that may regulate these expression patterns. To our knowledge, there is not a specific signal that regulates differential gene expression for these groups of genes.

In bacteria, gene expression is regulated by transcription factors (Balleza *et al.*, 2008), which bind to specific DNA regions and activate or repress gene expression. In addition, the DNA sequence upstream of the transcription starting site (TSS) has an important role in transcription. The -10 region (Pribnow box) is a region enriched in thymine (T) and adenine (A). This region is recognized and bound by the RNA polymerase during the initiation of transcription and where the first DNA base pairs separate (Feklistov and Darst, 2011). The AT-richness is important to allow this separation since adenine and thymine are easier to break apart (Yakovchuk *et al.*, 2006), therefore, G/C content in the region surrounding the TSS might have an effect on the changes in expression of the genes expressed in free-living conditions when infecting the plant.

We focus our analysis on three groups of genes: genes which are expressed in the free-living conditions whose expression increases in the bacteroid, genes which are expressed in the free-living conditions whose expression decreases in the bacteroid, and genes which are expressed in the free-living conditions whose expression remains constant in the bacteroid. As telling apart low expressed genes from not expressed genes is not trivial, we identify which genes are highly expressed in free-living conditions by applying a strict threshold to the free-living expression data. Then, out of this initial list of genes, we select genes based on their expression change in the bacteroid using RNA-Seq data from WT nodules (data presented in Chapter 5). Afterwards, we increase these groups using the gene coexpression network methods

described in Chapter 4. We evaluate and compare the G/C content of the genes in the different groups in the regions surrounding the transcription starting sites. We use the sets of genes obtained just from the expression data, the sets of genes resulting from the network analysis, and the intersection and overlap from those sets.

6.2 Methods

6.2.1 Identification of groups of genes which are highly expressed in free-living conditions

We first select those genes that are highly expressed in free-living conditions. For this purpose, we use the Free-living RNA-Seq presented in Chapter 5; the RNA extraction experiments for obtaining this dataset were carried out by Dr Raphael Ledermann. We select those genes that are among the 10% highest expressed genes for all three samples included in the dataset. Then, we analyse the differential expression of the selected genes between free-living conditions and the bacteroid, using the RNA-Seq datasets presented in Chapter 5 and DESeq2 (Love *et al.*, 2014) to obtain the log2FC values and the p-value associated to them. From these initial set of genes, we select the genes that have a log2FC value greater than 2 and a p-value less than 0.05 (genes that increase their expression in the bacteroid), those that have a log2FC value less than –2 and a p-value less than 0.05 (genes that decrease their expression in the bacteroid), and those that have a log2FC value greater than –0.5 and less than 0.5 and a p-value greater than 0.05 (genes whose expression remains constant in the bacteroid). We refer to these three sets of genes as those obtained *without the network*.

The conditions to obtain the sets of genes without the network are quite stringent, in particular when selecting the genes that are expressed in free-living conditions. To overcome this situation, we use the sets of genes obtained without the network as seed lists to identify the genes that are highly coexpressed with them. For this purpose, we use the complete network and the methodology presented in Chapter

4. We refer to the resulting sets of genes as those obtained *with the network*. In our further analyses, we evaluate separately the sets of genes obtained with and without the network, as well as their union and intersection sets.

6.2.2 Identification of Transcription Starting Sites

For each of the genes included in any of the sets presented in Section 6.2.1, we identify its TSS using a TSS library available in Prof Poole's lab created by former member Dr Vinoy K. Ramachandran. This library, elaborated using RNA-Seq data, includes the specific nucleotides where transcription starts, the direction of transcription (forward or reverse), and the first gene downstream of the TSS. For those genes included in an operon, we use the TSS upstream of the first coding sequence in the operon. We analyse the region starting 40 nucleotides upstream and ending 40 nucleotides downstream of the TSS. In each of the 12 sets of regions (e.g. regions surrounding the TSS of genes that increase their expression in the bacteroid obtained using the network), we delete those repeated sequences (if two selected genes are in the same operon, they will have the same TSS assigned).

6.2.3 Evaluation of regions surrounding the Transcription Starting Sites

We calculate the G/C content in a six-nucleotide-long window starting at each nucleotide each position of the selected sequence regions. In other words, the value for the position i is the fraction of guanine or cytosine nucleotides in the region from i to $i + 6$; for each region we obtain 81 values. For each of the different methods (e.g. sets obtained without the network), we perform 81 Dunn Tests (one per window) to assess whether the G/C-content values for the three groups of genes (i.e. genes whose expression increases, decreases, or remains constant in the bacteroid) are significantly different. We obtain 243 p-values since there are three comparisons per window. We then correct to account for the number of tests using a Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995) with a

false discovery rate of 20%. Following this procedure, for each p-value, we obtain a critical value resulting from diving the rank of the p-value across the 243 p-values by the number of p-values and multiplying by the false discovery rate. We consider significant all the p-values lower than the largest p-value below its corresponding critical value. We perform the analysis of the different methods separately.

6.3 Results and discussion

6.3.1 Identification and analysis of groups of genes which highly expressed in free-living conditions

Attending the RNA-Seq data (i.e. without using the network), we identify 121 genes whose expression decreases in the bacteroid, 64 genes whose expression increases in the bacteroid and 88 genes whose expression remains constant in the bacteroid; all of them are highly expressed in free-living conditions. After using these sets of genes as seed lists (i.e. using the network), we retrieve 25 genes whose expression decreases in the bacteroid, 49 genes whose expression increases in the bacteroid and 26 genes whose expression remains constant in the bacteroid. The list of genes included in each group is available in Table E.4 in Appendix E.

In Figure 6.1, we present the expression of the genes in each set in the free-living conditions (panels on the left) and their FC values, which resulted from dividing the expression values in the bacteroid by the expression values in the bacteroid (panels on the right). We observe that for the set of genes obtained without using the network, the distribution of the expression values in the free-living conditions is larger than for the set obtained without using the network. Thus, using our coexpression network allows us to identify genes that, while showing the same expression pattern as the genes in the employed seed lists, have lower absolute expression values. In the set of genes whose expression increases in the bacteroid obtained using the network there are 6 genes (RL0050, RL1355, RL3093, RL3378,

RL3752, pRL80013) whose expression decreases in the bacteroid. We remove these genes from the next steps of our analysis.

Figure 6.2 illustrates the induced subgraphs from our gene coexpression network including the genes in each studied group of genes. In the figure, the genes retrieved without using the network are represented in red, those retrieved using the coexpression network are represented in blue, and those retrieved by both methods are represented in black. We observe that the genes obtained using the network (in blue) increase the connectivity of the graphs.

6.3.2 Evaluation of regions surrounding the Transcription Starting Sites

We obtain the list of regions associated with the genes included in the different analysed groups; Table 6.1 indicates the number of regions analysed for each group after removing repeated regions. We perform the Dunn tests and the Benjamini–Hochberg procedure adjustments. As per these adjustments, for the genes obtained without using the network, we consider significant all the p-values lower than 0.0014. The equivalent values for the other sets are 0.0033 (with the network), 0.0031 (intersection), and 0.0074 (union).

	Exp. decreases	**Exp. increases**	**Exp.remains ct.**
Without Network	103	51	84
With Network	21	33	24
Union	112	71	100
Intersection	10	13	5

Table 6.1: Number of regions surrounding TSS analysed per each group of genes. The rows indicate the method employed to retrieve the genes. The columns indicate the expression pattern of the genes: their expression decreases in the bacteroid ("Exp. decreases"), their expression increases in the bacteroid ("Exp. increases") or their expression remains constant in the bacteroid ("Exp. remains ct."), compared to free-living conditions.

Figure 6.3 illustrates the variation of the G/C content across the different

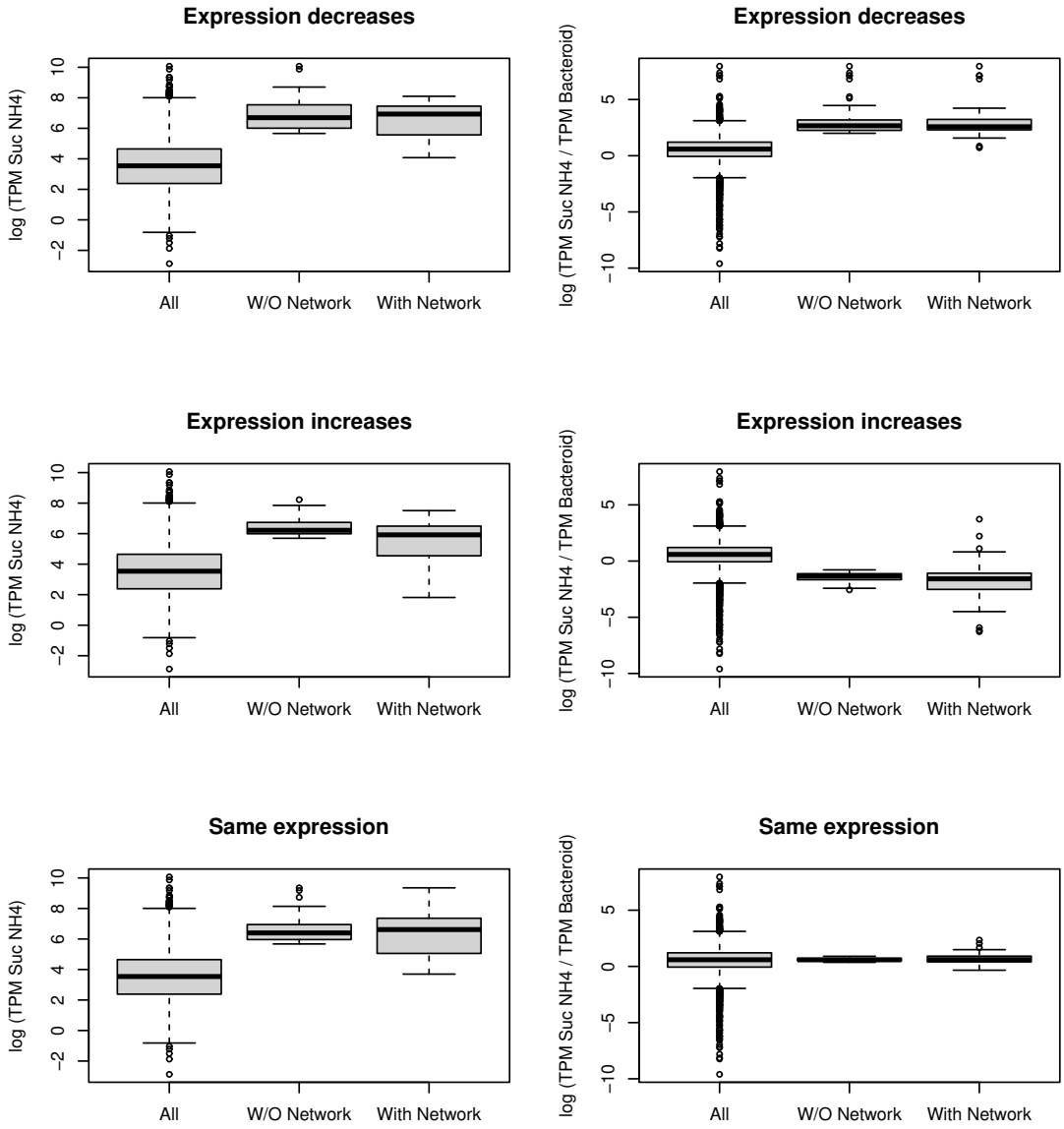


Figure 6.1: Boxplots representing the gene expression values, expressed as the logarithm of the TPM values, in free-living conditions (panels on the left) and the logFC values resulting from dividing the expression values in free-living conditions by the expression values in the bacteroid (panels on the left) of the studied groups of genes. Each panel includes the values for all the genes in the dataset ("All"), the sets of genes obtained without using the network ("W/O Network"), and the set of genes retrieved using the coexpression network ("With Network"). The notches display a confidence interval around the median which is based on the median +/- 1.58*IQR/$\sqrt{n}$, where IQR indicate the middle 50% of values when ordered from lowest to highest and n the number of genes in the set.

.

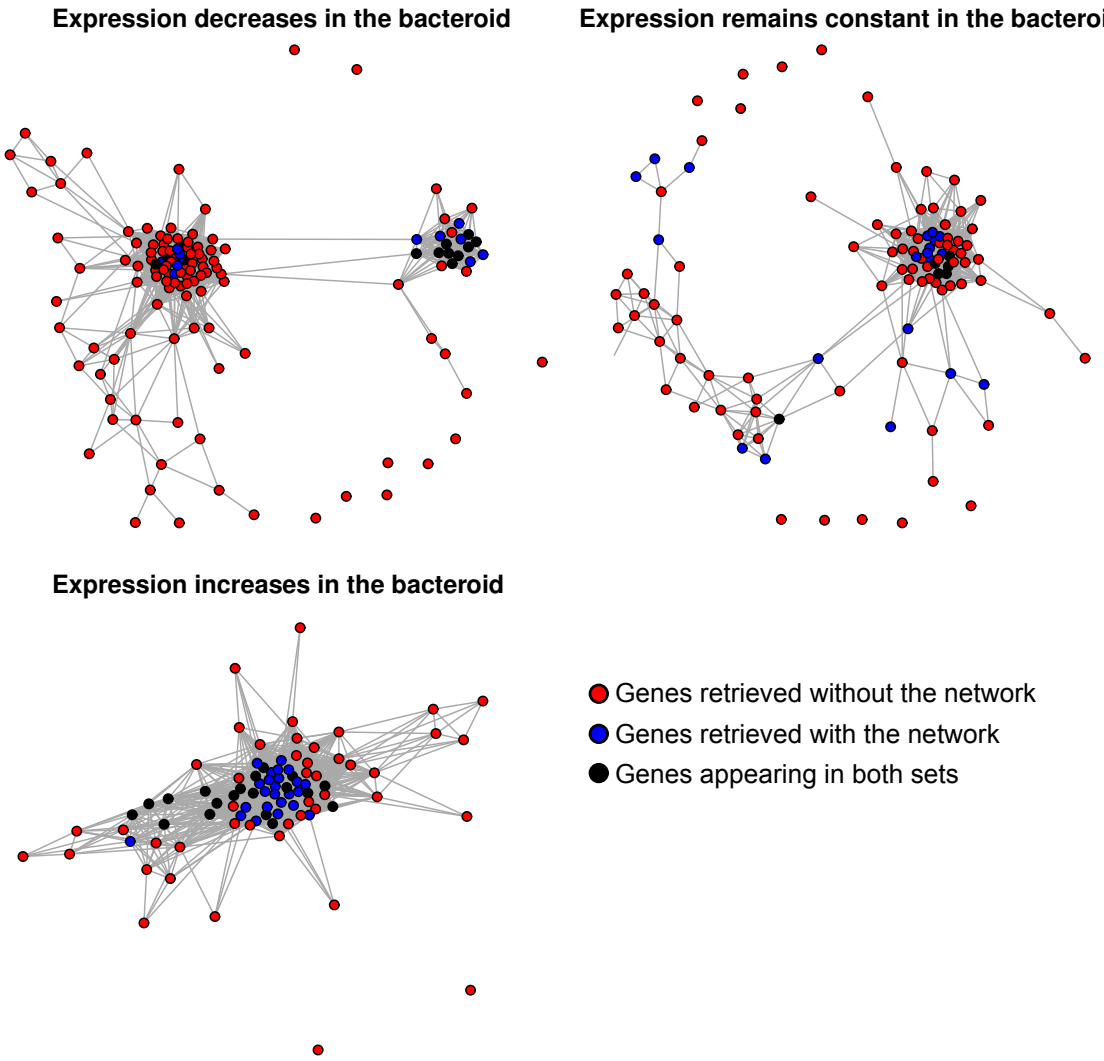


Figure 6.2: Induced subgraphs of the genes included in the different groups of genes studied. The genes retrieved without using the network are represented in red, those retrieved using the coexpression network are represented in blue, and those retrieved by both methods are represented in black.

analysed windows for the different methods and expression patterns. We observe that, for all cases, there is a decrease of G/C content in the area around the position –10, corresponding to the Pribnow box. In the figure, the black dots at the bottom indicate significant G/C content comparisons between the different expression patterns. For all the methods, we observe some windows for which there are significant differences in their G/C content across the groups of genes with different expression patterns. For example, the G/C content of the windows starting at 13, 12, and 10 nucleotides upstream of the TSS of genes whose expression

decreases in the bacteroid and obtained without using the network is significantly different from the values obtained for genes whose expression remains constant in the bacteroid and obtained employing the same method. We observe the highest number of significant comparisons (10) when using the union of the sets of genes obtained using with and without using the network, suggesting that coexpression information might ease the evaluation of differences between different groups. We do not observe any comparison that is significant for all four methods. We note that for each method, due to the false discovery rate we have used, up to 20% of the comparisons identified as significant might be false positives.

6.3.3 Future directions

We have seen that there are some statistically significant differences in the G/C content of the genes which are highly expressed in free-living conditions and whose expression increases, decreases or remains constant in the bacteroid, compared to free-living conditions. Nevertheless, these differences depend on the methods employed to generate the different groups of genes. Future research will focus on how better redefine the groups of genes, attending to the coexpression information contained in the network, and studying more in detail the relationship between the variation of G/C content in the regions surrounding the TSS and the expression pattern of the genes. Future research shall also include a more detailed analysis of the different groups of genes, including, for example, sequence alignment approaches.

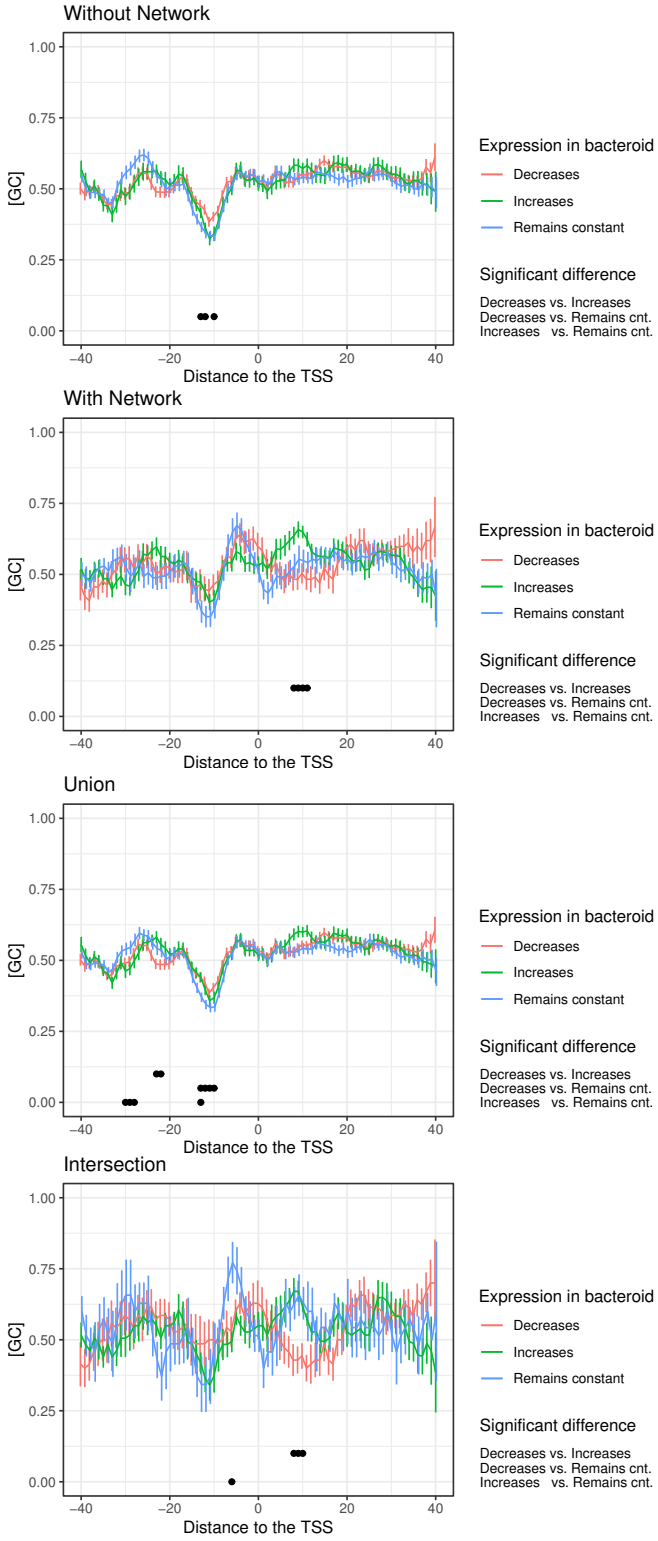


Figure 6.3: Variation of G/C content in the regions surrounding the TSS of studied genes. Each panel includes information on the sets of genes retrieved employing different methods. In the panels, y-axes indicate the proportion of G or C in six-nucleotide-long windows starting in the position indicated in the x-axes. Each line represents the mean G/C content of the genes with different expression patterns: their expression decreases (red), increases (green) or remains constant (blue) in the bacteroid, compared to free-living conditions. The error bars represent the standard deviation. The black dots at the bottom indicate significant G/C content comparisons between the different expression patterns.

7

Conclusions and future directions

Gene coexpression networks ease the study and visualisation of gene expression data and can be applied to study Rlv3841. In this thesis, I have proposed new methods to construct (Chapters 2 and 3) and analyse (Chapter 4) gene coexpression networks and applied them to Rlv3841 gene expression data, resulting in the first gene coexpression network for this organism. I have also carried out a detailed analysis of the transcriptome of the bacteroid by analysing gene expression data from several transcriptional regulator mutants (Chapter 5).

7.1 Construction of gene coexpression networks

In this thesis, I have proposed the concept of signed distance correlation as a powerful metric to measure gene coexpression. I have shown that coexpression networks constructed using this measurement capture more biological information and are more robust than those obtained using other well-established methods, such as the Pearson correlation coefficient or mutual information. I also showed

that this approach is suitable for microarray, RNA-Seq, and scRNA-Seq gene expression data. Since the publication of the distance correlation method presented in this thesis (Pardo-Diaz *et al.*, 2021), other authors have used distance and signed distance correlation to construct biological networks. Hou *et al.* (2022) used distance correlation to construct gene coexpression networks and showed that it provides better performance in detecting complex relationships than Pearson's correlation, Spearman's correlation, and maximal information coefficient. They also combined distance correlation with the module detection function implemented in WGCNA (Langfelder and Horvath, 2008), resulting in an enhancement of the result of enrichment analysis and an improvement of the module stability. Čuperlović-Culf *et al.* (2022) used the signed distance correlation metric to generate metabolic networks and concluded that signed distance correlation values indicate both strength and overall directionality of correlations. These studies reinforce the suitability of (signed) distance correlation for studying biological data.

Moreover, I have presented a principled method for threshold selection in the construction of unweighted and weighted gene coexpression networks based on the similarity of the networks obtained from overlapping gene expression datasets using COGENT (Bozhilova *et al.*, 2020). This approach identifies the threshold without requiring previous biological information, resulting in a convenient tool to analyse gene expression data from organisms which lack detailed functional annotation. I showed that weighted and thresholded networks capture more biological information than unweighted networks while still sparse.

A limitation of my pipeline to construct gene coexpression networks is that it disregards negative correlations in the gene expression . While this information might not be of interest for finding groups of genes with similar expression patterns and associating them with biological processes as carried out in this dissertation, this question remains open and shall be dealt with in future research. A scenario in which negative edges might provide insightful information is when studying transcriptional repression. These analyses would require using only a smaller range of conditions to construct the coexpression network (e.g. only expression data from

the bacteroid). An approach that may be useful for this purpose is evaluating which of the pairs of genes which are coexpressed when using a broad range of conditions (as the datasets employed in this thesis to construct the networks) have a negative correlation in their expression when reducing the set of samples.

In Chapter 2, I selected the empirical probability distribution as the entropy estimator to compute the mutual information. However, there are other approaches to estimate the entropy which shall be considered in future research (e.g. Miller-Madow asymptotic bias corrected empirical estimator or Schurmann-Grassberger estimate of the entropy of a Dirichlet probability distribution (Meyer *et al.*, 2008). In addition, the choice of density estimation can be delicate, as the accuracy of the mutual information calculation depends on the quality of the density estimation. In particular, the estimated density should accurately reflect the true underlying relationship between the two variables in order to accurately compute the mutual information. Future research shall carefully select the methodology to compute mutual information from gene expression data to ensure the best possible results; and then compare the resulting gene expression networks with those derived from our signed distance correlation pipeline.

7.2 Extracting information from gene coexpression networks

I have introduced and combined three score-based methods, namely community detection co-occurrence, personalised Page Rank, and edge weights, to extract biological information from gene coexpression networks. These methods include two flavours: the first one retrieves those genes which are coexpressed with a predefined set of genes; the second one retrieves those genes which are coexpressed with genes that are highly expressed in a particular condition or environment. One application of this approach is the creation of an induced subgraph that includes only the retrieved genes (e.g. a subgraph of the genes that are coexpressed with the ribosomal genes or a subgraph of the genes which are coexpressed with the

genes expressed in the bacteroid). Thus, this exploratory tool might be used as a starting point to address different research questions, such as studying a particular gene or process in detail.

I have confirmed that the three scores are able to capture information from the coexpression networks. I have shown that combining the rankings of the genes for the different methods by picking the genes that are highly ranked in one of them and also those that are somewhat less highly ranked, but fairly highly ranked across all the methods, captures considerable signals. I have also proposed an approach for combining the different scores to obtain a prioritised list of genes but further research shall focus on how other combination strategies (e.g. introducing a selection parameter that allows selecting those genes with a score among the highest scores for at least two methods) might affect the final result. Other open research questions from this project relate to how the community detection co-occurrence score compares to tracing the community membership of the genes through multiple partitions obtained from using different resolution values (Lewis *et al.*, 2010), and the possibility of including overlapping communities in the analysis (Riccio-Rengifo *et al.*, 2021).

Future research will also consider other methods to extract information from gene coexpression networks, assess how they compare to those proposed, and how implement them in my pipeline. For example, NetMix2 (Chitra *et al.*, 2022) is a principled algorithm for identifying groups of densely connected nodes with high scores from a wide range of subnetworks obtained following different approaches. These approaches include methods in which the nodes are ranked based on random walks or diffusion processes. The idea underpinning this method resembles the approach I have presented, which also combines algorithms based on the identification of highly connected groups of nodes (community detection) and propagation techniques (PageRank). Network analysis methods based on Diffusion State Distances (DSD) (Cao *et al.*, 2013) have also been shown to be able to quantify topological similarity in biological networks and predict protein function (Lazarsfeld *et al.*, 2019). DSD is a metric that measures the distance from each pair of nodes in

the network based on a random walk and correcting for highly visited nodes; the distances are then clustered or used to retrieve nodes with a low DSD value among those studied. Future research shall focus on how this method compares to the combination of PageRank and the use of edge weights, as proposed in this work.

7.3 Transcriptomic analysis of *R. leguminosarum*

I have performed a detailed analysis of the Rlv3841 transcriptome and have constructed and analysed two gene coexpression networks for this organism. These two networks, obtained from two different datasets, were similar but not identical. These results open the door to further studies on how altering the expression data may affect the resulting network. The network-comparison metrics implemented in COGENT might be useful to evaluate the structural similarity of two networks obtained using different datasets. The methods used throughout this thesis to obtain biological information from the networks shall be applied to networks resulting from different datasets in order to compare the information they contain.

The networks I have presented as a tool to study Rlv3841 (see Appendix A) have already been used by different members of Professor Poole's lab (e.g. Ford (2021)), especially in guiding preliminary studies. Beyond extracting some rough biological insights from these coexpression networks, as future research, it would be possible to attempt to survey the gene coexpression networks themselves. This analysis would entail looking at the overall structure of the network and identify different groups of nodes highly connected (i.e. communities) and relate them to a particular biological process or function. In other words, if a high proportion of the genes in one of those communities have a well-defined function, this approach would allow us to hypothesise the biological function of the uncharacterised genes in the community.

The networks presented in this thesis have some limitations. Firstly, one cannot generate a network for a specific condition but only generate an induced subgraph including the genes which are highly expressed in said condition or the genes which are coexpressed with the highly expressed genes by using the approaches described

in Chapter 4. Secondly, the networks have a fixed number of nodes, determined by the genes present in the dataset used to generate the network. Therefore, one cannot study genes that are not present in the original network. To overcome this issue, one might construct a new network from a new dataset or use other genes already present in the network as a proxy (e.g. use as a proxy a gene which is in the same operon that the gene which is not included in the network).

In this work, I have analysed the role of transcriptional regulators that are essential in the bacteroid and shortlisted a selection of genes for further investigation. However, I have not been able to identify the signals that are responsible for the upregulation of most of the genes which are upregulated in the bacteroid. The results in Chapter 5 suggest that some of those transcriptional regulators might be involved in other processes, such as root attachment. Future research will focus on studying the transcriptome of the mutants during root colonisation and compare the results with those already obtained. That information will be combined with data from competition assays. In Chapter 6, I have presented a preliminary study to disentangle the molecular signals that may affect the expression during plant infection of the genes which are expressed in free-living conditions.

Ultimately, it would be useful to have a multilayer network (Kivelä *et al.*, 2014) in which the different layers are the different stages during root colonisation and nodule formation (i.e. rhizosphere, root attachment, nodule bacteria, and bacteroid). To construct these networks, I suggest two possible approaches. The first one involves using the already available complete coexpression network and assigning scores to the nodes based on their expression in the different stages. For this purpose, RNA-Seq data, as well as the methods to extract biological information from the networks presented in Chapter 4, can be applied. The second approach relies on generating different coexpression networks from different sets of RNA-Seq data. For example, to construct the layer for the bacteroid, one could use only expression data from the bacteroid. This approach would lead to a set of layers containing different sets of genes (e.g. as *fix* genes are only expressed in the bacteroid, they would not appear in the other layers). However, having a network derived from a smaller set of samples

(and with fewer nodes) might be useful for studying negative edges. The resulting multilayer networks could be thought of as a temporal network in which edges connecting different genes rewire as a result of the changes in Rlv3841 transcriptome during plant colonisation, and could be analysed using the tools developed for this purpose (Lubba *et al.*, 2019; Fulcher *et al.*, 2013; Fulcher and Jones, 2017).

7.4 Closing remarks

In this thesis, I have presented and evaluated novel methods to construct and analyse gene coexpression networks. I have illustrated the benefits of using signed distance correlation over other commonly-used methods (Chapter 2) and the advantages of weighted networks over unweighted networks (Chapter 3). I have shown that combining different strategies to extract information from the networks might increase the signal captured 4. These methods can be used and adapted by researchers to carry out their own investigations.

I have applied the methods developed in this thesis to construct and analyse the first gene coexpression networks for Rlv3841. These networks are a tool that will allow continuing studying this important bacteria. Lastly, I have carried out an exhaustive transcriptomic analysis of the bacteroid (Chapter 5) and an assessment of the transcription starting sites of sets of genes with defined gene expression patterns (Chapter 6). These results constitute the starting point for future work to identify signals that trigger the expression of bacteroid-specific genes.

Bibliography

Alekshun, M. N. and Levy, S. B. (1997). Regulation of chromosomally mediated multiple antibiotic resistance: the mar regulon. *Antimicrobial agents and chemotherapy*, **41**(10), 2067–2075.

Altschul, S. F. (1993). A protein alignment scoring system sensitive at all evolutionary distances. *Journal of molecular evolution*, **36**(3), 290–300.

Amaratunga, D. and Cabrera, J. (2001). Analysis of data from viral DNA microchips. *Journal of the American Statistical Association*, **96**(456), 1161–1170.

Balleza, E., Lopez-Bojorquez, L. N., Martínez-Antonio, A., Resendis-Antonio, O., Lozada-Chávez, I., Balderas-Martínez, Y. I., Encarnación, S., and Collado-Vides, J. (2008). Regulation by transcription factors in bacteria: beyond description. *FEMS microbiology reviews*, **33**(1), 133–151.

Barnes, M., Freudenberg, J., Thompson, S., Aronow, B., and Pavlidis, P. (2005). Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms. *Nucleic Acids Research*, **33**(18), 5914–5923.

Becker, A., Kleickmann, A., Keller, M., Arnold, W., and Pühler, A. (1993). Identification and analysis of the rhizobium meliloti exoamonp genes involved in exopolysaccharide biosynthesis and mapping of promoters located on the exohklamonp fragment. *Molecular and General Genetics MGG*, **241**(3), 367–379.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, **57**(1), 289–300.

Beringer, J. E. (1974). R factor transfer in rhizobium leguminosarum. *Microbiology*, **84**(1), 188–198.

Blackburn, J., Wong, T., Madala, B. S., Barker, C., Hardwick, S. A., Reis, A. L., Deveson, I. W., and Mercer, T. R. (2019). Use of synthetic dna spike-in controls (sequins) for human genome sequencing. *Nature Protocols*, **14**(7), 2119–2151.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, **2008**(10), P10008.

Bolstad, B. M., Irizarry, R. A., Åstrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**(2), 185–193.

Bozhilova, L. V., Pardo-Diaz, J., Reinert, G., and Deane, C. M. (2020). COGENT: evaluating the consistency of gene co-expression networks. *Bioinformatics*. btaa787.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine.

Butte, A. (2002). The use and analysis of microarray data. *Nature reviews drug discovery*, **1**(12), 951–960.

Butte, A. J. and Kohane, I. S. (1999). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In *Biocomputing 2000*, pages 418–429. World Scientific.

Cain, A. K., Barquist, L., Goodman, A. L., Paulsen, I. T., Parkhill, J., and Van Opijnen, T. (2020). A decade of advances in transposon-insertion sequencing. *Nature Reviews Genetics*.

Cao, M., Zhang, H., Park, J., Daniels, N. M., Crovella, M. E., Cowen, L. J., and Hescott, B. (2013). Going the distance for protein function prediction: a new distance metric for protein interaction networks. *PloS one*, **8**(10), e76339.

Carter, S. L., Brechbühler, C. M., Griffin, M., and Bond, A. T. (2004). Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics*, **20**(14), 2242–2250.

Chai, Y., Tsai, C. S., Cho, H., and Winans, S. C. (2007). Reconstitution of the biochemical activities of the attj repressor and the attk, attl, and attm catabolic enzymes of agrobacterium tumefaciens. *Journal of bacteriology*, **189**(9), 3674–3679.

Chao, M. C., Abel, S., Davis, B. M., and Waldor, M. K. (2016). The design and analysis of transposon insertion sequencing experiments. *Nature Reviews Microbiology*, **14**(2), 119–128.

Chitra, U., Park, T. Y., and Raphael, B. J. (2022). Netmix2: Unifying network propagation and altered subnetworks. In *International Conference on Research in Computational Molecular Biology*, pages 193–208. Springer.

Cubo, M., Economou, A., Murphy, G., Johnston, A., and Downie, J. (1992). Molecular characterization and regulation of the rhizosphere-expressed genes rhiabcr that can influence nodulation by rhizobium leguminosarum biovar viciae. *Journal of Bacteriology*, **174**(12), 4026–4035.

Čuperlović-Culf, M., Yilmaz, A., Stewart, D., Surendra, A., Akyol, S., Vishweswaraiah, S., Shao, X., Alecu, I., Nguyen-Tran, T., McGuinness, B., Passmore, P., Kehoe, P. G., Maddens, M. E., Green, B. D., Graham, S. F., and Bennett, S. A. (2022). Signed distance correlation (sidco): A network analysis application of distance correlation for identifying metabolic networks disrupted in dementia with lewy bodies. *medRxiv*.

De Castro, E., Sigrist, C. J., Gattiker, A., Bulliard, V., Langendijk-Genevaux, P. S., Gasteiger, E., Bairoch, A., and Hulo, N. (2006). Scanprosite: detection of prosite signature matches and prorule-associated functional and structural residues in proteins. *Nucleic acids research*, **34**(suppl_2), W362–W365.

De Lorenzo, G. and Ferrari, S. (2002). Polygalacturonase-inhibiting proteins in defense against phytopathogenic fungi. *Current opinion in plant biology*, **5**(4), 295–299.

Denarie, J., Debelle, F., and Prome, J.-C. (1996). Rhizobium lipo-chitooligosaccharide nodulation factors: signaling molecules mediating recognition and morphogenesis. *Annual Review of Biochemistry*, **65**(1), 503–535.

Dixon, R. and Kahn, D. (2004). Genetic regulation of biological nitrogen fixation. *Nature Reviews Microbiology*, **2**(8), 621–631.

Downie, J. A. (2010). The roles of extracellular proteins, polysaccharides and signals in the interactions of rhizobia with legume roots. *FEMS Microbiology Reviews*, **34**(2), 150–170.

Economou, A., Hawkins, F., Downie, J., and Johnston, A. (1989). Transcription of rhia, a gene on a rhizobium leguminosarum bv. viciae sym plasmid, requires rhir and is repressed by flavanoids that induce nod genes. *Molecular microbiology*, **3**(1), 87–93.

Feklistov, A. and Darst, S. A. (2011). Structural basis for promoter- 10 element recognition by the bacterial rna polymerase σ subunit. *Cell*, **147**(6), 1257–1269.

Fischer, H., Babst, M., Kaspar, T., Acuna, G., Arigoni, F., and Hennecke, H. (1993). One member of a gro-esl-like chaperonin multigene family in bradyrhizobium japonicum is co-regulated with symbiotic nitrogen fixation genes. *The EMBO journal*, **12**(7), 2901–2912.

Fischer, H.-M., Schneider, K., Babst, M., and Hennecke, H. (1999). Groel chaperonins are required for the formation of a functional nitrogenase in bradyrhizobium japonicum. *Archives of microbiology*, **171**(4), 279–289.

Ford, B. L. (2021). *Investigating the basis of competitive nodulation and bacteroid development in Rhizobium leguminosarum*. Ph.D. thesis, University of Oxford.

Frederix, M. (2010). *A novel mechanism of coupling quorum sensing systems in Rhizobium leguminosarum bv. viciae 3841*. Ph.D. thesis, University of East Anglia.

Frederix, M., Edwards, A., Swiderska, A., Stanger, A., Karunakaran, R., Williams, A., Abbruscato, P., Sanchez-Contreras, M., Poole, P. S., and Downie, J. A. (2014). Mutation of prar in r hizobium leguminosarum enhances root biofilms, improving nodulation competitiveness by increased expression of attachment proteins. *Molecular microbiology*, **93**(3), 464–478.

Fulcher, B. D. and Jones, N. S. (2017). hctsa: A computational framework for automated time-series phenotyping using massive feature extraction. *Cell systems*, **5**(5), 527–531.

Fulcher, B. D., Little, M. A., and Jones, N. S. (2013). Highly comparative time-series analysis: the empirical structure of time series and their methods. *Journal of the Royal Society Interface*, **10**(83), 20130048.

Fuller, T., Langfelder, P., Presson, A., and Horvath, S. (2011). Review of weighted gene coexpression network analysis. In *Handbook of Statistical Bioinformatics*, pages 369–388. Springer.

George, G., Singh, S., Lokappa, S. B., and Varkey, J. (2019). Gene co-expression network analysis for identifying genetic markers in Parkinson's disease-a three-way comparative approach. *Genomics*, **111**(4), 819–830.

Ghazalpour, A., Doss, S., Zhang, B., Wang, S., Plaisier, C., Castellanos, R., Brozell, A., Schadt, E. E., Drake, T. A., Lusis, A. J., and Hovarth, S. (2006). Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genetics*, **2**(8), e130.

Gleich, D. F. (2015). Pagerank beyond the web. *Siam Review*, **57**(3), 321–363.

González-Pasayo, R. and Martínez-Romero, E. (2000). Multiresistance genes of rhizobium etli cfn42. *Molecular plant-microbe interactions*, **13**(5), 572–577.

Gourion, B., Delmotte, N., Bonaldi, K., Nouwen, N., Vorholt, J. A., and Giraud, E. (2011). Bacterial rubisco is required for efficient bradyrhizobium/aeschynomene symbiosis. *PloS one*, **6**(7), e21900.

Guo, X., Xiao, H., Guo, S., Dong, L., and Chen, J. (2017). Identification of breast cancer mechanism based on weighted gene coexpression network analysis. *Cancer gene therapy*, **24**(8), 333–341.

Gutiérrez, R. A. (2012). Systems biology for enhanced plant nitrogen nutrition. *Science*, **336**(6089), 1673–1675.

Gygi, S. P., Rochon, Y., Franza, B. R., and Aebersold, R. (1999). Correlation between protein and mRNA abundance in yeast. *Molecular and Cellular Biology*, **19**(3), 1720–1730.

Hanahan, D. (1983). Studies on transformation of escherichia coli with plasmids. *Journal of molecular biology*, **166**(4), 557–580.

Herrero, J., Díaz-Uriarte, R., and Dopazo, J. (2003). Gene expression data preprocessing. *Bioinformatics*, **19**(5), 655–656.

Hirsch, A. M. and Smith, C. A. (1987). Effects of rhizobium meliloti nif and fix mutants on alfalfa root nodule development. *Journal of bacteriology*, **169**(3), 1137–1146.

Hou, J., Ye, X., Feng, W., Zhang, Q., Han, Y., Liu, Y., Li, Y., and Wei, Y. (2022). Distance correlation application to gene co-expression network analysis. *BMC bioinformatics*, **23**(1), 1–24.

Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D., Kidd, M. J., King, A. M., Meyer, M. R., Slade, D., Lum, P. Y., Stepaniants, S. B., Shoemaker, D. D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M., and Friend, S. H. (2000). Functional discovery via a compendium of expression profiles. *Cell*, **102**(1), 109–126.

Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T. K., Bateman, A., Bernard, T., Binns, D., Bork, P., Burge, S., *et al.* (2012). Interpro in 2011: new developments in the family and domain prediction database. *Nucleic acids research*, **40**(D1), D306–D312.

Ideker, T., Thorsson, V., Siegel, A. F., and Hood, L. E. (2000). Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *Journal of Computational Biology*, **7**(6), 805–817.

Imlay, J. A. (2006). Iron-sulphur clusters and the problem with oxygen. *Molecular microbiology*, **59**(4), 1073–1082.

Jacob, F. and Monod, J. (1961). On the regulation of gene activity. In *Cold Spring Harbor symposia on quantitative biology*, volume 26, pages 193–211. Cold Spring Harbor Laboratory Press.

Johnston, A. W. and Beringer, J. (1975). Identification of the rhizobium strains in pea root nodules using genetic markers. *Microbiology*, **87**(2), 343–350.

Karp, P. D., Billington, R., Caspi, R., Fulcher, C. A., Latendresse, M., Kothari, A., Keseler, I. M., Krummenacker, M., Midford, P. E., Ong, Q., *et al.* (2019). The biocyc collection of microbial genomes and metabolic pathways. *Briefings in bioinformatics*, **20**(4), 1085–1093.

Karunakaran, R., Ramachandran, V., Seaman, J., East, A., Mouhsine, B., Mauchline, T., Prell, J., Skeffington, A., and Poole, P. (2009). Transcriptomic analysis of Rhizobium leguminosarum biovar viciae in symbiosis with host plants Pisum sativum and Vicia cracca. *Journal of Bacteriology*, **191**(12), 4002–4014.

Karunakaran, R., Haag, A. F., East, A. K., Ramachandran, V. K., Prell, J., James, E. K., Scocchi, M., Ferguson, G. P., and Poole, P. S. (2010). Baca is essential for bacteroid development in nodules of galegoid, but not phaseoloid, legumes. *Journal of bacteriology*, **192**(11), 2920–2928.

Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., and Porter, M. A. (2014). Multilayer networks. *Journal of complex networks*, **2**(3), 203–271.

Knights, H. E., Jorrin, B., Haskett, T. L., and Poole, P. S. (2021). Deciphering bacterial mechanisms of root colonization. *Environmental Microbiology Reports*, **13**(4), 428–444.

Kothapalli, R., Yoder, S. J., Mane, S., and Loughran, T. P. (2002). Microarray results: how accurate are they? *BMC Bioinformatics*, **3**(1), 22.

Langfelder, P. and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**(1), 559.

Langfelder, P. and Horvath, S. (2014). Tutorials for the WGCNA Package. `https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/`.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature methods*, **9**(4), 357–359.

Lazarsfeld, J., Rodríguez, J., Erden, M., Liu, Y., and Cowen, L. J. (2019). Majority vote cascading: a semi-supervised framework for improving protein function prediction. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 51–60.

Lazazzera, B. A., Bates, D. M., and Kiley, P. J. (1993). The activity of the escherichia coli transcription factor fnr is regulated by a change in oligomeric state. *Genes & development*, **7**(10).

Ledermann, R., Schulte, C. C., and Poole, P. S. (2021). How rhizobia adapt to the nodule environment. *Journal of bacteriology*, **203**(12), e00539–20.

Lee, H. K., Hsu, A. K., Sajdak, J., Qin, J., and Pavlidis, P. (2004). Coexpression analysis of human genes across many microarray data sets. *Genome Research*, **14**(6), 1085–1094.

Lewis, A. C., Jones, N. S., Porter, M. A., and Deane, C. M. (2010). The function of communities in protein interaction networks at multiple scales. *BMC systems biology*, **4**(1), 1–14.

Lim, W. K., Wang, K., Lefebvre, C., and Califano, A. (2007). Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. *Bioinformatics*, **23**(13), i282–i288.

Lindemann, A., Koch, M., Pessi, G., Müller, A. J., Balsiger, S., Hennecke, H., and Fischer, H.-M. (2010). Host-specific symbiotic requirement of bdeab, a regr-controlled rnd-type efflux system in bradyrhizobium japonicum. *FEMS microbiology letters*, **312**(2), 184–191.

Lithgow, J. K., Wilkinson, A., Hardman, A., Rodelas, B., Wisniewski-Dyé, F., Williams, P., and Downie, J. A. (2000). The regulatory locus cinri in rhizobium leguminosarum controls a network of quorum-sensing loci. *Molecular microbiology*, **37**(1), 81–97.

Ljunggren, I. and Fåhraeus, G. (1961). The role of polygalacturonase in root-hair invasion by nodule bacteria. *Microbiology*, **26**(3), 521–528.

Lodwig, E. and Poole, P. (2003). Metabolism of rhizobium bacteroids. *Critical reviews in plant sciences*, **22**(1), 37–78.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, **15**(12), 1–21.

Lubba, C. H., Sethi, S. S., Knaute, P., Schultz, S. R., Fulcher, B. D., and Jones, N. S. (2019). catch22: Canonical time-series characteristics. *Data Mining and Knowledge Discovery*, **33**(6), 1821–1852.

Lupwayi, N. Z., Clayton, G. W., and Rice, W. A. (2006). Rhizobial inoculants for legume crops. *Journal of Crop Improvement*, **15**(2), 289–321.

Magoc, T., Wood, D., and Salzberg', S. L. (2013). Edge-pro: estimated degree of gene expression in prokaryotic genomes. *Evolutionary Bioinformatics*, **9**, EBO–S11250.

Magwene, P. M. and Kim, J. (2004). Estimating genomic coexpression networks using first-order conditional independence. *Genome Biology*, **5**(12), R100.

Makrodimitris, S., Reinders, M. J., and van Ham, R. C. (2020). Metric learning on expression data for gene function prediction. *Bioinformatics*, **36**(4), 1182–1190.

Mandon, K., Michel-Reydellet, N., Encarnación, S., Kaminski, P. A., Leija, A., Cevallos, M. A., Elmerich, C., and Mora, J. (1998). Poly-β-hydroxybutyrate turnover inazorhizobium caulinodans is required for growth and affectsnifa expression. *Journal of bacteriology*, **180**(19), 5070–5076.

Martínez-Salazar, J. M., Salazar, E., Encarnación, S., Ramírez-Romero, M. A., and Rivera, J. (2009). Role of the extracytoplasmic function sigma factor rpoe4 in oxidative and osmotic stress responses in rhizobium etli. *Journal of bacteriology*, **191**(13), 4122–4132.

Mauchline, T., Fowler, J., East, A., Sartor, A., Zaheer, R., Hosie, A. H., Poole, P. S., and Finan, T. (2006). Mapping the sinorhizobium meliloti 1021 solute-binding protein-dependent transportome. *Proceedings of the National Academy of Sciences*, **103**(47), 17933–17938.

McCarter, L. L. (2006). Regulation of flagella. *Current opinion in microbiology*, **9**(2), 180–186.

Mendoza-Suárez, M., Andersen, S. U., Poole, P. S., and Sánchez-Cañizares, C. (2021). Competition, nodule occupancy, and persistence of inoculant strains: Key factors in the rhizobium-legume symbioses. *Frontiers in Plant Science*, page 1684.

Mendoza-Suárez, M. A., Geddes, B. A., Sánchez-Cañizares, C., Ramírez-González, R. H., Kirchhelle, C., Jorrin, B., and Poole, P. S. (2020). Optimizing rhizobium-legume symbioses by simultaneous measurement of rhizobial competitiveness and n2 fixation in nodules. *Proceedings of the National Academy of Sciences*, **117**(18), 9822–9831.

Meyer, P. E., Lafitte, F., and Bontempi, G. (2008). minet: Ar/bioconductor package for inferring large transcriptional networks using mutual information. *BMC bioinformatics*, **9**(1), 1–10.

Min, Y. and Kramer, E. M. (2020). Transcriptome profiling and weighted gene co-expression network analysis of early floral development in aquilegia coerulea. *Scientific reports*, **10**(1), 1–15.

Newman, M. E. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, **69**(2), 026113.

Nookaew, I., Papini, M., Pornputtapong, N., Scalcinati, G., Fagerberg, L., Uhlén, M., and Nielsen, J. (2012). A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. *Nucleic Acids Research*, **40**(20), 10084–10097.

Oldroyd, G. E. (2013). Speak, friend, and enter: signalling systems that promote beneficial symbiotic associations in plants. *Nature Reviews Microbiology*, **11**(4), 252.

Oresnik, I. J., Pacarynuk, L. A., O'Brien, S. A., Yost, C. K., and Hynes, M. F. (1998). Plasmid-encoded catabolic genes in rhizobium leguminosarum bv. trifolii: evidence for a plant-inducible rhamnose locus involved in competition for nodulation. *Molecular plant-microbe interactions*, **11**(12), 1175–1185.

Pardo-Diaz, J., Bozhilova, L. V., Beguerisse-Díaz, M., Poole, P. S., Deane, C. M., and Reinert, G. (2021). Robust gene coexpression networks using signed distance correlation. *Bioinformatics*. btab041.

Pini, F., East, A. K., Appia-Ayme, C., Tomek, J., Karunakaran, R., Mendoza-Suarez, M., Edwards, A., Terpolilli, J., Rowoth, J., Downie, J. A., and Poole, P. S. (2017). Lux bacterial biosensors for in vivo spatiotemporal mapping of root secretion. *Plant Physiology*, **174**(3), 1289–1306.

Poole, P. S. and Ledermann, R. (2022). Maintaining osmotic balance in legume nodules. *Journal of Experimental Botany*, **73**(1), 8–10.

Presson, A. P., Sobel, E. M., Papp, J. C., Suarez, C. J., Whistler, T., Rajeevan, M. S., Vernon, S. D., and Horvath, S. (2008). Integrated weighted gene co-expression network analysis with an application to chronic fatigue syndrome. *BMC systems biology*, **2**(1), 95.

Prieto, C., Rivas, M., Sánchez, J., López-Fidalgo, J., and De Las Rivas, J. (2006). Algorithm to find gene expression profiles of deregulation and identify families of disease-altered genes. *Bioinformatics*, **22**(9), 1103–1110.

Prieto, C., Risueño, A., Fontanillo, C., and De Las Rivas, J. (2008). Human gene coexpression landscape: confident network derived from tissue transcriptomic profiles. *PloS One*, **3**(12), e3911.

Ramos, J. L., Martínez-Bueno, M., Molina-Henares, A. J., Terán, W., Watanabe, K., Zhang, X., Gallegos, M. T., Brennan, R., and Tobes, R. (2005). The tetr family of transcriptional repressors. *Microbiology and molecular biology reviews*, **69**(2), 326–356.

Recourt, K., Van Brussel, A., Driessen, A., and Lugtenberg, B. (1989). Accumulation of a nod gene inducer, the flavonoid naringenin, in the cytoplasmic membrane of rhizobium leguminosarum biovar viciae is caused by the ph-dependent hydrophobicity of naringenin. *Journal of Bacteriology*, **171**(8), 4370–4377.

Reichardt, J. and Bornholdt, S. (2006). Statistical mechanics of community detection. *Physical Review E*, **74**(1), 016110.

Riccio-Rengifo, C., Finke, J., and Rocha, C. (2021). Identifying stress responsive genes using overlapping communities in co-expression networks. *BMC bioinformatics*, **22**(1), 1–17.

Riley, M. (1993). Functions of the gene products of escherichia coli. *Microbiology and Molecular Biology Reviews*, **57**(4), 862–952.

Rutten, P. J. and Poole, P. S. (2019). Oxygen regulatory mechanisms of nitrogen fixation in rhizobia. *Advances in Microbial Physiology*, **75**, 325–389.

Salem, M., Kenney, P. B., Rexroad III, C. E., and Yao, J. (2017). Microarray gene expression analysis in atrophying rainbow trout muscle: a unique non-mammalian muscle degradation model. *Physiological Genomics*.

Sánchez-Cañizares, C., Jorrín, B., Poole, P. S., and Tkacz, A. (2017). Understanding the holobiont: the interdependence of plants and their microbiome. *Current Opinion in Microbiology*, **38**, 188–196.

Sánchez-Cañizares, C., Prell, J., Pini, F., Rutten, P., Kraxner, K., Wynands, B., Karunakaran, R., and Poole, P. S. (2020). Global control of bacterial nitrogen and carbon metabolism by a ptsntr-regulated switch. *Proceedings of the National Academy of Sciences*, **117**(19), 10234–10245.

Schäfer, A., Tauch, A., Jäger, W., Kalinowski, J., Thierbach, G., and Pühler, A. (1994). Small mobilizable multi-purpose cloning vectors derived from the escherichia coli plasmids pk18 and pk19: selection of defined deletions in the chromosome of corynebacterium glutamicum. *Gene*, **145**(1), 69–73.

Schulte, C. C., Ramachandran, V. K., Papachristodoulou, A., and Poole, P. S. (2022). Genome-scale metabolic modelling of lifestyle changes in rhizobium leguminosarum. *Msystems*, **7**(1), e00975–21.

Shah, V. K., Davis, L., and Brill, W. J. (1975). Nitrogenase: Vi. acetylene reduction assay: dependence of nitrogen fixation estimates on component ratio and acetylene concentration. *Biochimica et Biophysica Acta (BBA)-Enzymology*, **384**(2), 353–359.

Smith, C., Hill, A. K., and Torrente-Murciano, L. (2020). Current and future role of haber–bosch ammonia in a carbon-free energy landscape. *Energy & Environmental Science*, **13**(2), 331–344.

Sprent, J. I. and James, E. K. (2007). Legume evolution: where do nodules and mycorrhizas fit in? *Plant physiology*, **144**(2), 575–581.

Stertz, L., Henriette, R., and Walss-Bass, C. (2018). T192. Weighted Gene Coexpression Network Analysis of IPSC generated from patients with schizophrenia. *Biological Psychiatry*, **83**(9), S202–S203.

Stuart, J. M., Segal, E., Koller, D., and Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**(5643), 249–255.

Sudtachat, N., Ito, N., Itakura, M., Masuda, S., Eda, S., Mitsui, H., Kawaharada, Y., and Minamisawa, K. (2009). Aerobic vanillate degradation and c1 compound metabolism in bradyrhizobium japonicum. *Applied and environmental microbiology*, **75**(15), 5012–5017.

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., Jensen, L. J., and von Mering, C. (2019). STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, **47**(D1), D607–D613.

Tambalo, D. D., Del Bel, K. L., Bustard, D. E., Greenwood, P. R., Steedman, A. E., and Hynes, M. F. (2010). Regulation of flagellar, motility and chemotaxis genes in rhizobium leguminosarum by the visn/r-rem cascade. *Microbiology*, **156**(6), 1673–1685.

Terpolilli, J. J., Masakapalli, S. K., Karunakaran, R., Webb, I. U., Green, R., Watmough, N. J., Kruger, N. J., Ratcliffe, R. G., and Poole, P. S. (2016). Lipogenesis and redox balance in nitrogen-fixing pea bacteroids. *Journal of bacteriology*, **198**(20), 2864–2875.

Thalamuthu, A., Mukhopadhyay, I., Zheng, X., and Tseng, G. C. (2006). Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics*, **22**(19), 2405–2412.

Traag, V. A., Waltman, L., and van Eck, N. J. (2019). From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, **9**(1), 1–12.

Van Egeraat, A. (1975). The possible role of homoserine in the development of rhizobium leguminosarum in the rhizosphere of pea seedlings. *Plant and Soil*, pages 381–386.

van Noort, V., Snel, B., and Huynen, M. A. (2003). Predicting gene function by conserved co-expression. *TRENDS in Genetics*, **19**(5), 238–242.

Vanderlinde, E. M., Hynes, M. F., and Yost, C. K. (2014). Homoserine catabolism by r hizobium leguminosarum bv. viciae 3841 requires a plasmid-borne gene cluster that also affects competitiveness for nodulation. *Environmental Microbiology*, **16**(1), 205–217.

Venkatasubbarao, S. (2004). Microarrays–status and prospects. *TRENDS in Biotechnology*, **22**(12), 630–637.

Vincent, J. M. *et al.* (1970). A manual for the practical study of the root-nodule bacteria. *A manual for the practical study of the root-nodule bacteria.*

Walker, T., Dumadag, S., Lee, C. J., Lee, S. H., Bender, J. M., Abbott, J. C., and She, R. C. (2016). Clinical impact of laboratory implementation of verigene BC-GN microarray-based assay for detection of gram-negative bacteria in positive blood cultures. *Journal of Clinical Microbiology*, **54**(7), 1789–1796.

Wang, H., Zhang, K., Zhu, J., Song, W., Zhao, L., and Zhang, X. (2013). Structure reveals regulatory mechanisms of a maoc-like hydratase from phytophthora capsici involved in biosynthesis of polyhydroxyalkanoates (phas). *PLoS One*, **8**(11), e80024.

Wang, Z., Gerstein, M., and Snyder, M. (2009). Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, **10**(1), 57–63.

Watson, J. D. and Crick, F. H. (1953). Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, **171**(4356), 737–738.

Weirauch, M. T. (2011). Gene coexpression networks for the analysis of DNA microarray data. *Applied Statistics for Network Biology: Methods in Systems Biology*, **1**, 215–250.

Wheatley, R. M., Ramachandran, V. K., Geddes, B. A., Perry, B. J., Yost, C. K., and Poole, P. S. (2017). Role of O2 in the growth of Rhizobium leguminosarum bv. viciae 3841 on glucose and succinate. *Journal of Bacteriology*, **199**(1), e00572–16.

Wheatley, R. M., Ford, B. L., Li, L., Aroney, S. T. N., Knights, H. E., Ledermann, R., East, A. K., Ramachandran, V. K., and Poole, P. S. (2020). Lifestyle adaptations of rhizobium from rhizosphere to symbiosis. *Proceedings of the National Academy of Sciences*, **117**(38), 23823–23834.

Whittaker, C. A. and Hynes, R. O. (2002). Distribution and evolution of von willebrand/integrin a domains: widely dispersed domains with roles in cell adhesion and elsewhere. *Molecular biology of the cell*, **13**(10), 3369–3387.

Wikipedia (2022). Microarray analysis techniques — Wikipedia, the free encyclopedia. `http://en.wikipedia.org/w/index.php?title=Microarray%20analysis%20techniques&oldid=1103069851`. [Online; accessed 06-October-2022].

Wolfe, C. J., Kohane, I. S., and Butte, A. J. (2005). Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. *BMC Bioinformatics*, **6**(1), 227.

Yakovchuk, P., Protozanova, E., and Frank-Kamenetskii, M. D. (2006). Base-stacking and base-pairing contributions into thermal stability of the dna double helix. *Nucleic acids research*, **34**(2), 564–574.

Yang, Z., Algesheimer, R., and Tessone, C. J. (2016). A comparative analysis of community detection algorithms on artificial networks. *Scientific reports*, **6**, 30750.

Young, J. P. W., Crossman, L. C., Johnston, A. W., Thomson, N. R., Ghazoui, Z. F., Hull, K. H., Wexler, M., Curson, A. R., Todd, J. D., Poole, P. S., Mauchline, T. H., East, A. K., Quail, M. V., Churcher, C., Arrowsmith, C., Cherevach, I., Chillingworth, T., Clarke, K., Cronin, A., Davis, P., Fraser, A., Hance, Z., Hauser, H., Jagels, K., Moule, S., Mungall, K., Norbertczak, H., Rabbinowitsch, E., Sanders, M., Simmonds, M., Whitehead, S., and Parkhill, J. (2006). The genome of Rhizobium leguminosarum has recognizable core and accessory components. *Genome Biology*, **7**(4), R34.

Zhang, B. and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, **4**(1).

Zhang, H.-B., Wang, L.-H., and Zhang, L.-H. (2002). Genetic control of quorum-sensing signal turnover in agrobacterium tumefaciens. *Proceedings of the National Academy of Sciences*, **99**(7), 4638–4643.

Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., and Liu, X. (2014). Comparison of rna-seq and microarray in transcriptome profiling of activated t cells. *PloS one*, **9**(1), e78644.

Appendices

A

Online application to access *R. leguminosarum* gene coexpression networks

I have developed an online application using `shiny` to access the *R. leguminosarum* gene coexpression networks described in this thesis. This application is available at `https://javierpardodiaz.shinyapps.io/rl3841_gcn/`. The application allows the user to select a group of genes and plots the induced subgraph. The user can increase the number of neighbours k of the selected genes shown in the image. If they select more than one gene, there are two possible options:

- Union: All the genes at a distance equal to or less than k from any of the selected genes are included.
- Intersection: The genes at a distance equal to or less than k from all the selected genes are included.

They can also retrieve the gene(s) with the highest k co-occurrence community detection score ("Communities Score" option) and those with the k highest personalised Pagerank score ("PageRank" option) (see 4. The option of plotting the genes in the same community as the selected gene(s) at a given k resolution is also implemented ("Communities" option).

The tab "Table" shows the correlation matrix containing the correlation values for each pair of included genes. The tab "Gene info" shows the basic information (name, description and Riley code) of the included genes as well as the average correlation of that gene with the rest of included genes. The Riley codes (Riley, 1993) are 3-digits codes that relate to functional annotation of the genes. For example, all the codes starting with 1 make reference to cell processes. The application allows colouring the nodes (genes) in the network according to their Riley code. The user can also choose which dataset they want to work with: *standard dataset* or *complete dataset*, presented in chapters 2 and 4, respectively.

B

Supplementary Information: Robust gene coexpression networks using signed distance correlation

Supplementary Material: Robust gene coexpression networks using signed distance correlation

Javier Pardo-Diaz [1,3,*], Lyuba V. Bozhilova [1], Mariano Beguerisse-Díaz [2], Philip S. Poole [3], Charlotte M. Deane [1] and Gesine Reinert [1,†]
[1]Department of Statistics, University of Oxford, Oxford, UK
[2]Mathematical Institute, University of Oxford, Oxford, UK
[3]Department of Plant Sciences, University of Oxford, Oxford, UK
[*]jdiaz@stats.ox.ac.uk
[†]reinert@stats.ox.ac.uk

1 *Rhizobium leguminosarum* expression data

Table S1 provides the details of the gene expression data employed to generate the gene coexpression networks for *R. leguminosarum*. We used 54 different microarrays obtained under 18 conditions (first column). The second column indicates the growth condition of the bacteria; for example, "7 days pea, 1 day PI" stands for inoculation of seven days old pea plant sampled one day post-inoculation. The last column indicates the type of sample, which can be "Free-living" (bacteria grown in liquid culture), "Rhizosphere" (bacteria isolated from the fraction of soil in contact with the plant), or "Bacteroid" (bacteria that have infected the plant and differentiated). This data has been published previously [8, 6, 3] but never analysed jointly.

Microarrays (channels)	Growth condition	Type
kk21 (Cy3), kk41 (Cy5), kk45 (Cy3)	Succinate NH_4	Free-living
Jay10 (Cy5), Jay11 (Cy5), Jay12 (Cy3)	Glucose Glutamate	Free-living
Jay10 (Cy3), Jay11 (Cy3), Jay12 (Cy5)	Glucose Aspartate	Free-living
kk46 (Cy3), kk48 (Cy3), kk50 (Cy3)	Pyruvate NH_4	Free-living
Vinoy43 (Cy5), Vinoy44 (Cy5), Vinoy45 (Cy3)	Pyruvate NH_4 Hespertin	Free-living
Ade7 (Cy3), Ade9 (Cy3), Ade18 (Cy5)	Pyruvate NH_4 IAA	Free-living
Ade8 (Cy5), Ade11 (Cy3), Ade15 (Cy5)	Pyruvate NH_4 Kinetin	Free-living
kk39 (Cy5), kk43 (Cy5), kk51 (Cy3)	Inositol NH_4	Free-living
Vinoy26 (Cy5), Vinoy31 (Cy5), Vinoy34 (Cy3)	7 days pea, 1 day PI	Rhizosphere
Vinoy27 (Cy5), Vinoy32 (Cy5), Vinoy35 (Cy3)	7 days pea, 3 day PI	Rhizosphere
Vinoy28 (Cy5), Vinoy33 (Cy5), Vinoy36 (Cy3)	7 days pea, 7 day PI	Rhizosphere
Vinoy29 (Cy5), Vinoy37 (Cy5), Vinoy39 (Cy3)	14 days pea, 1 day PI	Rhizosphere
Vinoy30 (Cy5), Vinoy38 (Cy5), Vinoy40 (Cy3)	21 days pea, 1 day PI	Rhizosphere
kk67 (Cy5), kk71 (Cy3), kk72 (Cy5)	1 week pea, 7 days PI	Bacteroid
kk65 (Cy5), kk68 (Cy3), kk69 (Cy3)	1 week pea, 15 days PI	Bacteroid
kk73 (Cy5), kk74 (Cy3), kk75 (Cy5)	1 week pea, 21 days PI	Bacteroid
kk64 (Cy5), Vinoy22 (Cy3), Vinoy24 (Cy3)	1 week pea, 28 days PI	Bacteroid
kk70 (Cy5), kk60 (Cy3), kk61 (Cy5)	Vetch seed, 28 days PI	Bacteroid

Table 1: *Rhizobium leguminosarum* gene expression data description

2 Spearman correlation networks for *R. leguminosarum*

We construct a Spearman correlation matrix R from the pre-processed gene expression matrix M^* by computing the Spearman correlation between every pair of gene expression vectors. We threshold R to construct two networks $NR(d_S)$ and $NR(d_P)$ with edge density d_S and d_P, respectively. Similarly to the networks obtained using Pearson correlation $NP(d_S)$ and $NP(d_P)$, the ones based on Spearman correlation have a larger and less densely connected largest connected component than $NS(d_S)$ and $NS(d_P)$. The summaries of these two networks can be found in Table S2. We evaluate the Spearman networks using STRING as described in the paper. The results obtained are higher than the obtained by the Pearson correlation networks but lower than those using signed distance correlation. Table S3 and Figure S1 show the results of the STRING evaluation for the Spearman networks.

Network	Number of Edges	Number of vertices in LCC	Edge density LCC * 100	Global clustering coeff. LCC
$NR(d_S)$	313,348	6,664	1.41	0.515
$NR(d_P)$	406,977	6,831	1.34	0.515

Table 2: Summaries of Spearman networks for *R. leguminosarum* LCC Denotes largest connected component.

Network	All of STRING information (C)	Only coexpression information ($C^\dagger$)	All information except coexpression ($C^\ddagger$)
$NR(d_S)$	28.52	8.67	25.51
$NR(d_P)$	25.77	7.38	23.15

Table 3: Evaluation of the biological content of the Spearman networks with STRING

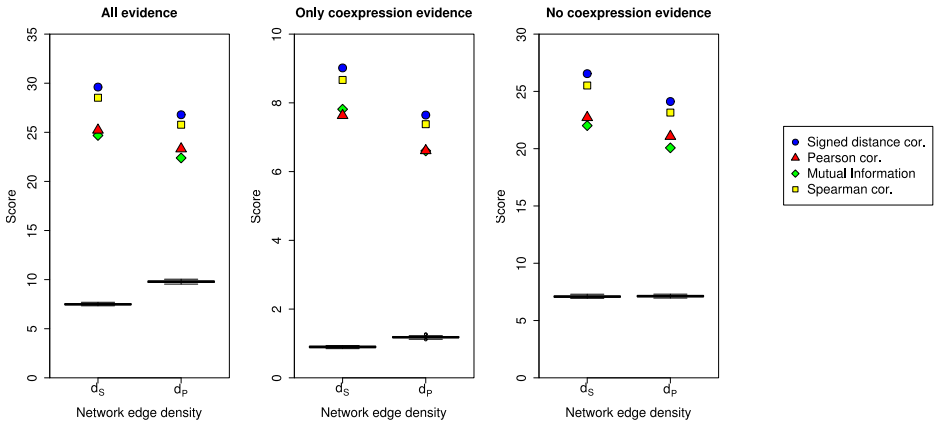


Figure 1: Scores obtained for the *R. leguminosarum* gene coexpression networks using STRING. All panels show the score for the different networks in the y-axis, and the network density on the x-axis. The scores are the result of adding up the confidence scores with all evidence (C), only coexpression evidence ($C^\dagger$) and everything excluding coexpression ($C^\ddagger$) from STRING associated with the edges in the networks, each computed using different of information. The black box plots correspond to the scores obtained by 30 random networks. Blue circles, red triangles and yellow squares represent signed distance correlation, Pearson correlation and Spearman correlation, respectively.

3 Gene coexpression network analysis for the study of Yeast RNA-Seq data

We use our signed correlation pipeline to generate a gene coexpression network from a dataset obtained using RNA-Seq of yeast (*Saccharomyces cerevisiae*) expressing pathways designed to increase ATP or GTP consumption. We obtain all the raw-counts for experiment E-MTAB-5174 in Expression Atlas [5] and remove the genes with zero expression variance. The final dataset which we feed into our pipeline includes the expression of 6,930 genes across 209 samples.

Following the pipeline described in the methodology section, we pre-process the data and obtain the distance correlation matrix D, the Pearson correlation matrix P, the signed correlation matrix S, and the matrix $|P|$ of absolute values of the Pearson correlation. Table S4 and Figure S2 show the summaries and distribution of these matrices.

Correlation matrix	Min	1st q	Median	3rd q	Max	Mean
D	0.00	0.26	0.40	0.57	1.00	0.42
$\|P\|$	0.00	0.13	0.30	0.51	1.00	0.34
P	-0.96	-0.29	0.00	0.31	1.00	0.01
S	-0.97	-0.39	-0.05	0.40	1.00	0.01

Table 4: Summaries for the correlation matrices of the Yeast dataset

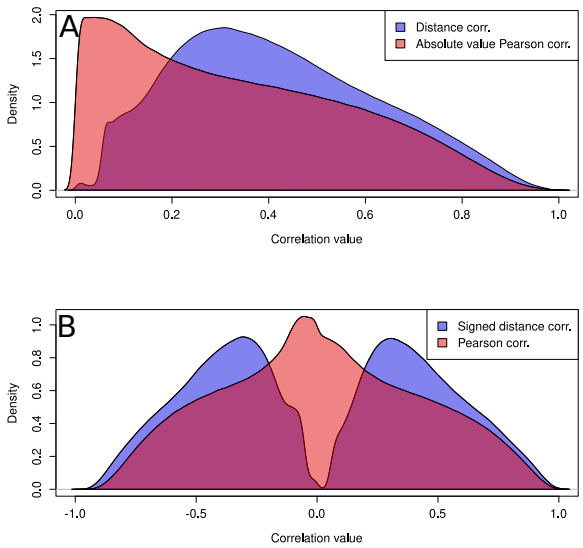


Figure 2: Density plots of the distribution of the values of the correlation matrices from the Yeast dataset. Panel A shows the distribution of values in the distance correlation matrix (blue) and of the absolute value of the values in the Pearson correlation matrix (red). Panel B shows the distribution of values in the signed distance correlation matrix (blue) and of the values in the Pearson correlation matrix (red).

We estimate the optimal threshold values θ^*, $\theta^\star$, and $\theta^\diamond$ to construct the unweighted gene coexpression networks $A_S(\theta^*)$, $A_P(\theta^\star)$, and and $A_I(\theta^\diamond)$ using COGENT [2]. We follow our pipeline to calculate the self-consistency of the networks. We adjust the similarity score by subtracting the network density. Fig S3 shows the variation of the score function for the correlation matrices across different edge densities. For the three matrices, there is a edge density value for which the score function reaches its maximum. This value is $d_S = 0.0131$ for the signed distance correlation (score of 0.746, which is achieved for $\theta^*= 0.84$), $d_P = 0.0146$ for the Pearson correlation (score of 0.733, which is achieved for $\theta^\star = 0.81$), and $d_I = 0.024$ for Mutual Information (score of 0.381, which is achieved for $\theta^\diamond = 1.04$. The results show that the use of signed distance correlation offers more stable networks. As in the study of *R. leguminosarum*, the scores obtained using signed mutual information are lower than those retrieved using the correlation methods. However, the difference between the Pearson correlation and signed distance correlation methods is lower than the observed in the study of *R. leguminosarum*. We analyse the networks $NS(d_S)$ (edge density d_S), $NS(d_P)$ (edge density d_P), and $NS(d_I)$ (edge density d_I) retrieved from S, the networks $NP(d_S)$ (edge density d_S), $NP(d_P)$ (edge density d_P) and $NP(d_I)$ (edge density d_I) retrieved from P, the networks $NI(d_S)$ (edge density d_S),$NI(d_P)$ (edge density d_P), and $NI(d_I)$ (edge density d_I) retrieved from I, and the networks $NR(d_S)$ (edge density d_S), $NR(d_P)$ (edge density d_P), and $NR(d_I)$ (edge density d_I) obtained using Spearman correlation. The summaries of the 12 networks are detailed in Table S5.

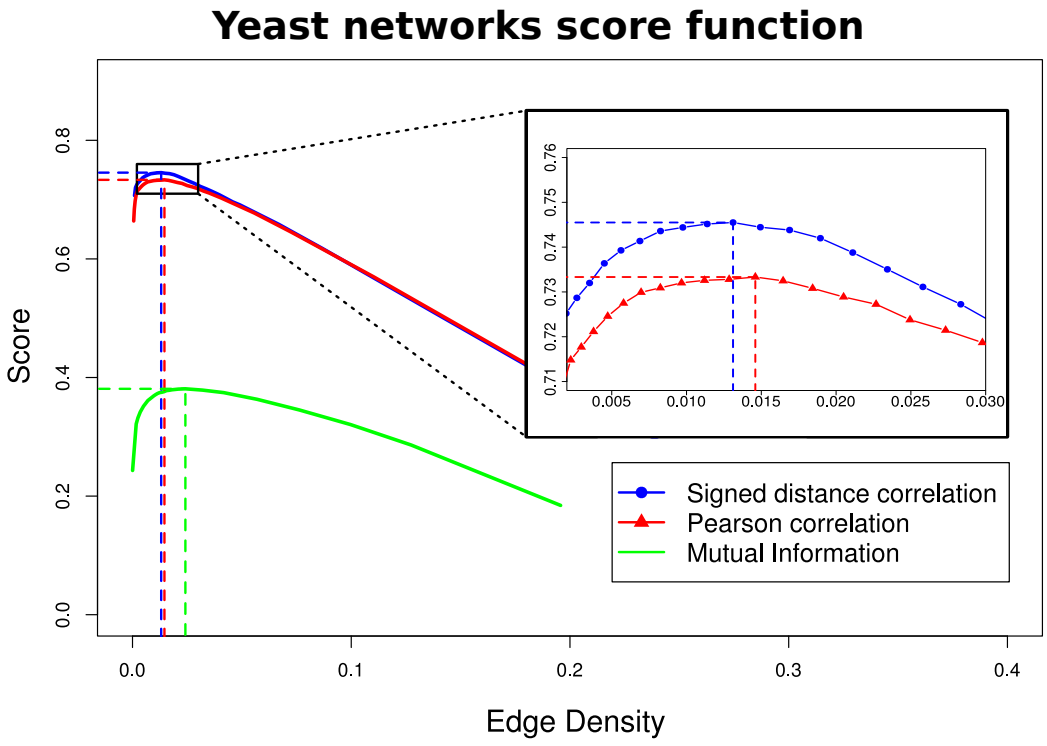


Figure 3: Score function values for different edge densities using the Yeast dataset. The blue line with circles shows the scores obtained using signed distance correlation and the red line with triangles those obtained using Pearson correlation. The green line those obtained using mutual information. The dotted lines indicate the position of the highest score point for each line. This value is 0.746 for the signed distance correlation network (giving edge density 0.0131 which is achieved with $\theta^* = 0.84$), 0.733 for the Pearson correlation network (giving edge density 0.0146 which is achieved with $\theta^\star = 0.81$) and 0.381 for the mutual information network (giving edge density 0.024 which is achieved with $\theta^\diamond = 1.04$.

We evaluate all the networks using STRING [9]. We use the three different sets of confidence scores:

Network	Number of Edges	Number of vertices in LCC	Edge density LCC * 100	Global clustering coeff. LCC
$NS(d_S)$	314,771	3,987	3.91	0.651
$NS(d_P)$	350,549	4,154	4.02	0.607
$NS(d_I)$	580,360	4,954	4.73	0.715
$NP(d_S)$	314,771	4,497	3.07	0.431
$NP(d_P)$	350,549	4,615	3.25	0.516
$NP(d_I)$	580,360	5,334	4.08	0.644
$NI(d_S)$	314,771	4,001	3.93	0.411
$NI(d_P)$	350,549	4,085	4.20	0.46
$NI(d_I)$	580,360	4,424	5.93	0.44
$NR(d_S)$	314,771	5,060	2.39	0.578
$NR(d_P)$	350,549	5,117	2.44	0.578
$NR(d_I)$	580,361	5,432	3.92	0.594

Table 5: Summaries of yeast networks. LCC denotes largest connected component.

obtained using all the evidence in STRING (C), obtained using only coexpression information ($C^{\dagger}$), and obtained using all the evidence except the coexpression information ($C^{\ddagger}$). Table S6 and Figure S4 present the results. For all the studied cases, the results obtained using the signed distance correlation networks are the highest. Unlike with the *R. leguminosarum* dataset, the Pearson correlation networks perform better than the Spearman correlation networks and the mutual information networks. In fact, the results for the signed distance correlation and Pearson correlation are almost identical (as in the case of the self-consistency study). These results reinforces our hypothesis that a high similarity in the self-consistency of the networks may imply a similarity in the amount of biological information that the networks are able to capture.

We generate 90 random networks of which 30 have edge density d_S, 30 have edge density d_P and other 30 have edge density d_I. We evaluate these networks with the STRING information and find that the all the networks in Table S5 outperform all of them. The highest difference to random is obtained for signed distance correlation networks when using only coexpression information ($C^{\dagger}$) to evaluate the networks; for $NS(d_S)$ the score is 17.69 times higher than the mean score obtained by random networks with matching densities.

Network	All of STRING information (C)	Only coexpression information ($C^{\dagger}$)	All information except coexpression ($C^{\ddagger}$)
$NS(d_S)$	**117.36**	**85.03**	**67.14**
$NP(d_S)$	116.31	84.85	66.3
$NI(d_S)$	68.31	45.70	41.67
$NR(d_S)$	74.84	48.11	50.48
RE d_S	14.02±0.14	4.81±0.08	11.89±0.12
$NS(d_P)$	**111.29**	**79.69**	**64.59**
$NP(d_P)$	110.14	79.37	63.64
$NI(d_P)$	66.36	44.15	40.72
$NR(d_P)$	72.45	46.05	49.19
RE d_P	14.02±0.14	4.79±0.07	11.91±0.13
$NS(d_I)$	**86.65**	**58.62**	**53.34**
$NP(d_I)$	85.56	57.94	52.60
$NI(d_I)$	58.30	37.37	36.93
$NR(d_I)$	62.03	37.67	43.10
RE d_I	13.98±0.08	4.78±0.04	11.87±0.08

Table 6: Evaluation of the biological content of the networks with STRING. RE indicates the expected (mean) result based on random networks with the indicated edge density and its standard deviation.

We conclude that the Yeast gene coexpression networks obtained using signed distance correlation are more stable and recover more biological information than those based on Pearson correlation, mutual information, or Spearman correlation.

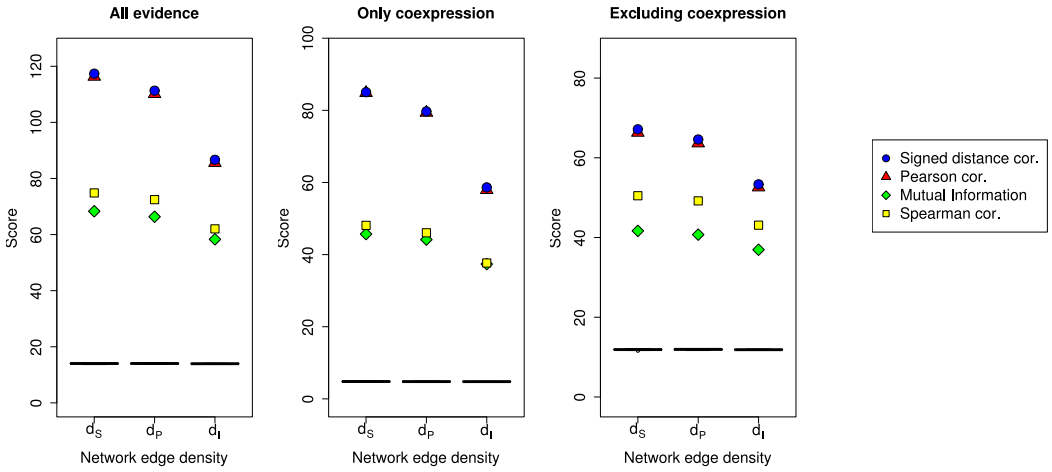


Figure 4: Scores obtained for the yeast gene coexpression networks using STRING. All panels show the score for the different networks in the y-axis, and the network density on the x-axis. The scores are the result of adding up the confidence scores from STRING associated with the edges in the networks. Each plot corresponds to a different set of values: using all evidence C, only coexpression evidence $C^{\dagger}$ and excluding coexpression evidence $C^{\ddagger}$. The black box plots correspond to the scores obtained by 30 random networks. Blue circles, red triangles, green diamonds, and yellow squares represent signed distance correlation, Pearson correlation, mutual information, and Spearman correlation, respectively.

4 Gene coexpression network analysis for the study of human liver single-cell RNA-Seq data

We use our signed correlation pipeline to generate a gene coexpression network from a dataset obtained using single-cell RNA-Seq of human liver cells [4]. Most of the methods to study single-cell RNA-Seq data aim to cluster cells according to their levels of gene expression [11, 10]. However, in this section we aim to group genes based on their expression across different samples (or cells, in this context) by constructing a gene coexpression network. Both approaches are complementary and we hope to combine them in the future. The original dataset measures the expression of 15,353 genes in 1,622 cells.

This dataset is considerably different from the previous two, both in the data itself – the measurements correspond to different cells instead of to different samples – and the organism – while both *R. leguminosarum* and *S. cerevisiae* are unicellular organisms, humans are not. Hence, a considerable proportion of genes are not expressed in the studied cells (for example, specific genes in neurons will not be expressed in cells from liver). For this reason, we employ a different pre-processing strategy in this case. In the first place, we quantile-normalise the data [1] to make the measurements in the different cells comparable. Afterwards, as in [7], we identify the "non-changing genes". These genes are those for which the difference between its highest and lowest expression value ("expression difference") is lower than the median of all the expression differences calculated for each gene, and in addition for which the mean expression signal between samples is lower than the median of all the expression signals calculated for each gene. After removing the "non-changing genes" we obtain an already quantile-normalised dataset with information for 8,585 genes. We do not apply more preprocessing steps to this dataset (we do not quantile-normalise again the data and we do not set the lowest expressed genes from each sample to the lowest expression value).

We calculate the distance correlation matrix (D), the Pearson correlation matrix (P), the signed correlation matrix (S), and the matrix $|P|$ of absolute values of the Pearson correlation. Table S7 and Figure S5 show the summaries and distribution of these matrices.

Correlation matrix	Min	1st q	Median	3rd q	Max	Mean
D	0.00	0.03	0.04	0.05	0.88	0.04
$\|P\|$	0.00	0.01	0.02	0.03	0.91	0.02
P	-0.63	-0.02	0.00	0.02	0.91	0.00
S	-0.71	-0.03	-0.02	0.04	0.88	0.00

Table 7: Summaries for the correlation matrices of the single-cell dataset from human liver

We estimate the optimal threshold values θ^*, $\theta^\star$, and $\theta^\diamond$ to construct the unweighted gene coexpression networks $A_S(\theta^*)$, $A_P(\theta^\star)$, and $A_I(\theta^\diamond)$ using COGENT [2]. We follow a similar pipeline as the one employed in the case of *R. leguminosarum* to calculate and adjust the self-consistency of the networks. The only difference is that as the input expression matrix is already quantile-normalised and low expressed genes have already been filtered out, in each of the 25 iterations the pre-processing steps are omitted. Figure S6 shows the variation of the score function for the correlation matrices across different edge densities. For the three tested methods, there is a edge density value for which the score function reaches its maximum. This value is $d_S = 0.00009$ for signed distance correlation (score of 0.896, which is achieved for $\theta^* = 0.44$), $d_P = 0.000089$ for Pearson correlation (score of 0.781, which is achieved for $\theta^\star = 0.43$), and $d_I = 0.00017$ for mutual information (score of 0.63, which is achieved for $\theta^\star = 0.63$. The results show that for all tested edge densities, the use of signed distance correlation offers more stable networks. We analyse the networks $NS(d_S)$ (edge density d_S), $NS(d_P)$ (edge density d_P), and $NS(d_I)$ (edge density d_I) retrieved from S, the networks $NP(d_S)$ (edge density d_S), $NP(d_P)$ (edge density d_P) and $NP(d_I)$ (edge density d_I) retrieved from P, the networks $NI(d_S)$ (edge density d_S),$NI(d_P)$ (edge density d_P), and $NI(d_I)$ (edge density d_I) retrieved from I, and the networks $NR(d_S)$ (edge density d_S), $NR(d_P)$ (edge density d_P), and $NR(d_I)$ (edge density d_I) obtained using Spearman correlation. The summaries of all the networks are detailed in Table S8. We note that the largest connected components of the networks contain a lower proportion of

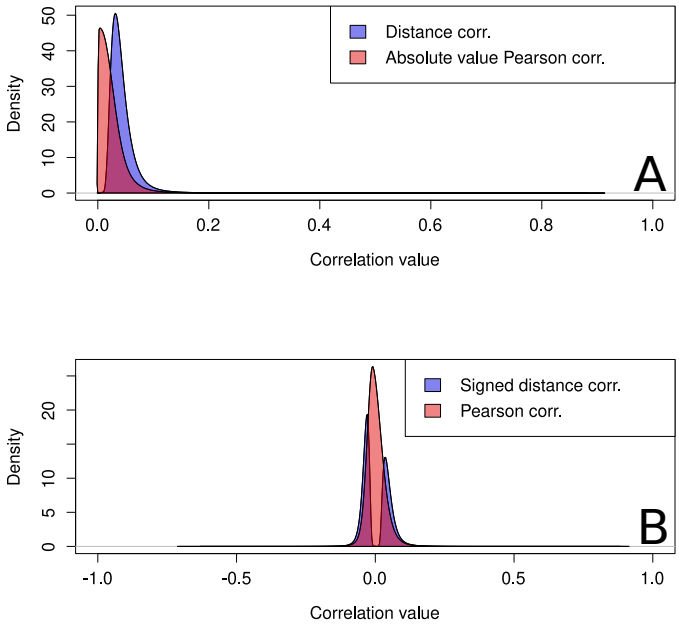


Figure 5: Density plots of the distribution of the values of the correlation matrices from the Yeast dataset. Panel A shows the distribution of values in the distance correlation matrix (blue) and of the absolute value of the values in the Pearson correlation matrix (red). Panel B shows the distribution of values in the signed distance correlation matrix (blue) and of the values in the Pearson correlation matrix (red).

the total vertices than in previous datasets. The networks obtained using signed distance correlation have a smaller and denser largest connected component than those obtained using Pearson correlation. In contrast to the two previous datasets, the largest connected components of the networks obtained using Spearman correlation and mutual information are smaller and denser than those from signed distance correlation and Pearson correlation networks; however, the signed distance correlation networks still have the highest global clustering coefficient. We observe that when increasing the edge density from d_S to d_I, the size of the largest connected component in the signed distance correlation networks increases considerably. This change is more subtle for the mutual information network. This finding may be connected with the fact that the optimal edge density for mutual information is higher than for signed distance correlation and Pearson correlation.

Similarly as for the two previous datasets, we evaluate the constructed networks using three sets of confidence values obtained from STRING. We also evaluate 60 random networks with edge densities d_S (30 networks), d_P (30 networks), and d_I (30 networks). Table S9 and Figure S7 show the results. Overall, the networks obtained using signed distance correlation $NS(d_S)$ and $NS(d_P)$ retrieve a higher score than their competitors based on either Pearson or Spearman correlation or mutual information. This result suggest that among the assessed methods, signed distance correlation captures the broadest range of biological information, which makes it a tool of choice to generating gene coexpression networks from single-cell RNA-Seq data. All the analysed networks obtain a higher score than the random ones. The highest difference to random is obtained for signed distance correlation networks when using only coexpression information ($C^{\dagger}$) to evaluate the networks; for $NS(d_P)$ the score is 92.92 times higher than the mean score obtained by random networks with matching densities; for $NS(d_S)$ it is even 94.28 than the corresponding mean score for the random networks.

We conclude that the liver gene coexpression networks obtained using signed distance correlation are more stable and show recover more biological information than those based on Pearson correlation and that our signed distance correlation method is suitable for studying single cell gene expression

Network	Number of Edges	Number of vertices in LCC	Edge density LCC * 100	Global clustering coeff. LCC
$NS(d_S)$	3320	212	14.61	0.87
$NS(d_P)$	3279	210	14.71	0.87
$NS(d_I)$	6150	398	7.69	0.83
$NP(d_S)$	3320	219	12.42	0.83
$NP(d_P)$	3279	216	12.62	0.83
$NP(d_I)$	6150	455	5.13	0.81
$NI(d_S)$	3320	204	16.00	0.80
$NI(d_P)$	3279	198	16.77	0.81
$NI(d_I)$	6150	279	15.84	0.77
$NR(d_S)$	3320	161	25.54	0.81
$NR(d_P)$	3279	160	25.54	0.81
$NR(d_I)$	6150	273	16.45	0.78

Table 8: Summaries of single cell human liver networks.
LCC denotes largest connected component.

Network	All of STRING information (C)	Only coexpression information ($C^{\dagger}$)	All information except coexpression ($C^{\ddagger}$)
$NS(d_S)$	**212.02**	**185.35**	**203.55**
$NP(d_S)$	196.30	169.84	188.12
$NI(d_S)$	189.05	172.79	184.31
$NR(d_S)$	183.07	173.02	179.42
RE d_S	6.75 ± 0.73	1.97 ± 0.3	5.90 ± 0.73
$NS(d_P)$	**212.35**	**185.80**	**204.00**
$NP(d_P)$	196.93	170.86	188.85
$NI(d_P)$	187.68	171.82	182.96
$NR(d_P)$	184.18	174.13	180.61
RE d_P	6.82 ± 0.82	2.00 ± 0.32	5.98 ± 0.78
$NS(d_I)$	**176.21**	**144.00**	**162.72**
$NP(d_I)$	162.60	132.43	152.03
$NI(d_I)$	140.49	121.55	131.09
$NR(d_I)$	146.91	131.03	136.13
RE d_I	7.03 ± 0.91	2.06 ± 0.25	6.17 ± 0.86

Table 9: Evaluation of the biological content of the networks with STRING. RE indicates the expected (mean) result based on random networks with
the indicated edge density and its standard deviation.

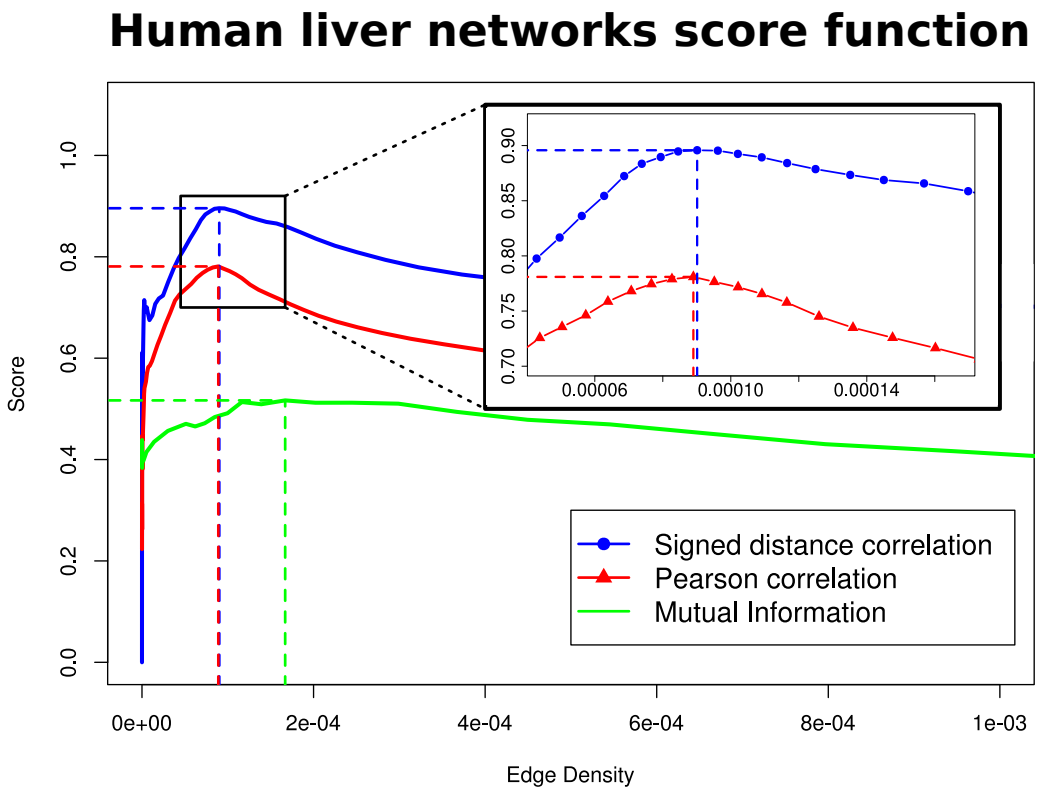


Figure 6: Plot of the score function values for different edge densities using the human liver single-cell RNA-Seq dataset. The blue line shows the scores obtained using signed distance correlation, the red line those obtained using Pearson correlation, and the green line those obtained using mutual information. The dotted lines indicate the position of the highest score point for each line. This value is 0.896 for the signed distance correlation network (giving edge density $9.01e^{-05}$ which is achieved with $\theta^* = 0.44$), 0.781 for the Pearson correlation network (giving edge density $8.90e^{-05}$ which is achieved with $\theta^* = 0.43$) and 0.517 for the mutual information network (giving edge density $1.67e^{-0.4}$.

data.

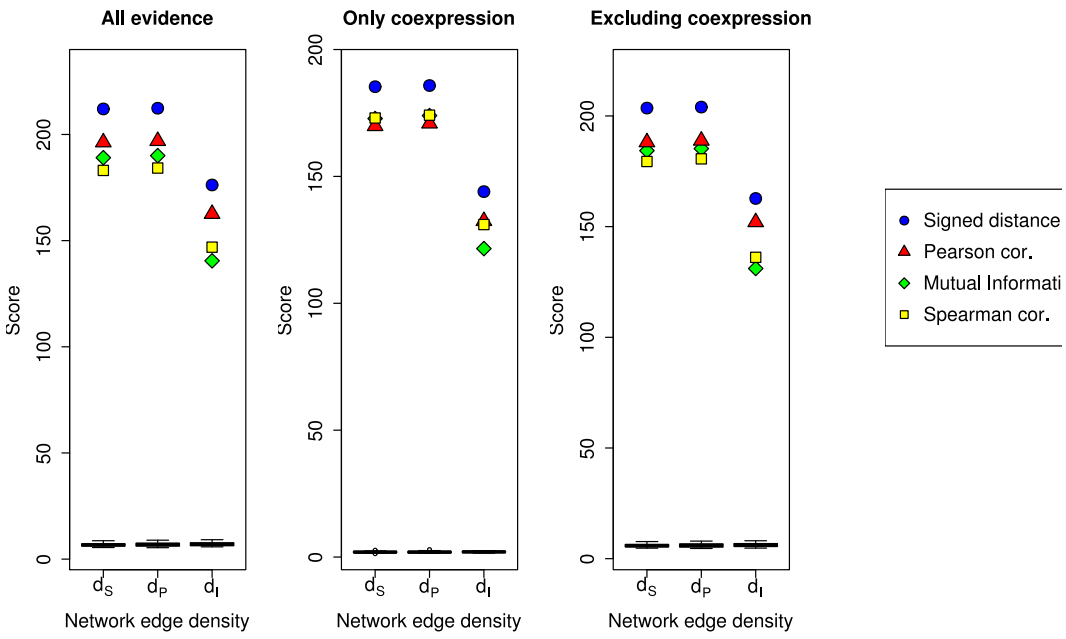


Figure 7: Scores obtained for the single-cell human liver gene coexpression networks using STRING. All panels show the score for the different networks in the y-axis, and the network density on the x-axis. The scores are the result of adding up the confidence scores from STRING associated with the edges in the networks. Each plot corresponds to a different set of values: using all evidence C, only coexpression evidence $C^{\dagger}$ and excluding coexpression evidence $C^{\ddagger}$. The black box plots correspond to the scores obtained by 30 random networks. Blue circles, red triangles and yellow squares represent signed distance correlation, Pearson correlation and Spearman correlation, respectively.

5 Network evaluation using information from STRING

We use the interactions between proteins reported in STRING to evaluate the amount of biological information that the constructed networks capture. The association evidence in STRING is categorized into independent channels, weighted, and integrated, resulting in a confidence score for all recorded protein interactions. For each organism, we use three different sets of interactions and confidence scores: the overall confidence score C provided by STRING, the $C^{\dagger}$ obtained attending only to coexpression information, and the $C^{\ddagger}$ obtained excluding all coexpression information.

The recomputing of the each of the scores was done using the python script located on the STRING webpage (Figure 8), which was accessed on the 20$^{\text{th}}$ of April 2020 using the URL `https://string-db.org/cgi/help.pl?&subpage=faq%23how-are-the-scores-computed`. The script can be found at the end of this section. We commented out lines 96–98, 100–102 in the script to compute the $C^{\dagger}$ values; and the line 99 in the script to obtain the $C^{\ddagger}$ values.

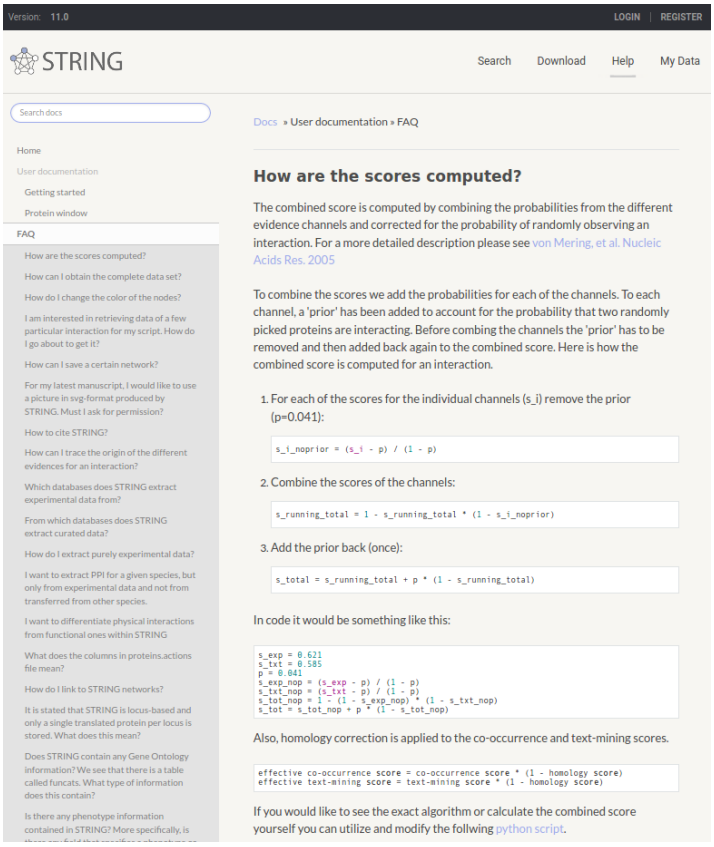


Figure 8: STRING FAQ webpage. Accessed on the 20$^{\text{th}}$ of April 2020. The content describes how the STRING combines the scores of the different channels. At the end of the section there is the link to the script which was modified and used in this work.

```
from __future__ import print_function
import os
import sys

############################################################
## This script combines all the STRING's channels subscores
## into the final combined STRING score.
## It uses unpacked protein.links.full.xx.txt.gz as input
## which can be downloaded from the download subpage:
##      https://string-db.org/cgi/download.pl
############################################################

input_file = "9606.protein.links.full.v10.5.txt"

if not os.path.exists(input_file):
    sys.exit("Can't locate input file %s" % input_file)

prior = 0.041

def compute_prior_away(score, prior):

    if score < prior: score = prior
    score_no_prior = (score - prior) / (1 - prior)

    return score_no_prior

header = True
for line in open(input_file):

    if header:
        header = False
        continue

    l = line.split()

    ## load the line

    (protein1, protein2,
     neighborhood, neighborhood_transferred,
     fusion, cooccurrence,
     homology,
     coexpression, coexpression_transferred,
     experiments, experiments_transferred,
     database, database_transferred,
     textmining, textmining_transferred,
     initial_combined) = l

    ## divide by 1000

    neighborhood = float(neighborhood) / 1000
    neighborhood_transferred = float(neighborhood_transferred) / 1000
    fusion = float(fusion) / 1000
    cooccurrence =  float(cooccurrence) / 1000
    homology = float(homology) / 1000
    coexpression = float(coexpression) / 1000
    coexpression_transferred = float(coexpression_transferred) / 1000
    experiments = float(experiments) / 1000
    experiments_transferred = float(experiments_transferred) / 1000
    database = float(database) / 1000
    database_transferred = float(database_transferred) / 1000
    textmining = float(textmining) / 1000
    textmining_transferred = float(textmining_transferred) / 1000
    initial_combined = int(initial_combined)

    ## compute prior away

```

```
    neighborhood_prior_corrected             = compute_prior_away (neighborhood, prior)
    neighborhood_transferred_prior_corrected = compute_prior_away (neighborhood_transferred, prior)
    fusion_prior_corrected                   = compute_prior_away (fusion, prior)
    cooccurrence_prior_corrected             = compute_prior_away (cooccurrence, prior)
    coexpression_prior_corrected             = compute_prior_away (coexpression, prior)
    coexpression_transferred_prior_corrected = compute_prior_away (coexpression_transferred, prior)
    experiments_prior_corrected              = compute_prior_away (experiments, prior)
    experiments_transferred_prior_corrected  = compute_prior_away (experiments_transferred, prior)
    database_prior_corrected                 = compute_prior_away (database, prior)
    database_transferred_prior_corrected     = compute_prior_away (database_transferred, prior)
    textmining_prior_corrected               = compute_prior_away (textmining, prior)
    textmining_transferred_prior_corrected   = compute_prior_away (textmining_transferred, prior)

    ## then, combine the direct and transferred scores for each category:

    neighborhood_both_prior_corrected = 1.0 - (1.0 - neighborhood_prior_corrected) * (1.0 - neighborhood_transferred_prior_corrected)
    coexpression_both_prior_corrected = 1.0 - (1.0 - coexpression_prior_corrected) * (1.0 - coexpression_transferred_prior_corrected)
    experiments_both_prior_corrected  = 1.0 - (1.0 - experiments_prior_corrected) * (1.0 - experiments_transferred_prior_corrected)
    database_both_prior_corrected     = 1.0 - (1.0 - database_prior_corrected) * (1.0 - database_transferred_prior_corrected)
    textmining_both_prior_corrected   = 1.0 - (1.0 - textmining_prior_corrected) * (1.0 - textmining_transferred_prior_corrected)

    ## now, do the homology correction on cooccurrence and textmining:

    cooccurrence_prior_homology_corrected = cooccurrence_prior_corrected * (1.0 - homology)
    textmining_both_prior_homology_corrected = textmining_both_prior_corrected * (1.0 - homology)

    ## next, do the 1 - multiplication:

    combined_score_one_minus = (
        (1.0 - neighborhood_both_prior_corrected) *
        (1.0 - fusion_prior_corrected) *
        (1.0 - cooccurrence_prior_homology_corrected) *
        (1.0 - coexpression_both_prior_corrected) *
        (1.0 - experiments_both_prior_corrected) *
        (1.0 - database_both_prior_corrected) *
        (1.0 - textmining_both_prior_homology_corrected) *
        1)

    ## and lastly, do the 1 - conversion again, and put back the prior *exactly once*

    combined_score = (1.0 - combined_score_one_minus)            ## 1- conversion
    combined_score *= (1.0 - prior)                              ## scale down
    combined_score += prior                                      ## and add prior.

    ## round

    combined_score = int(combined_score * 1000)
    print(protein1, protein2, combined_score)
```

Listing 1: Combine subscores script

References

[1] Benjamin M Bolstad, Rafael A Irizarry, Magnus Åstrand, and Terence P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

[2] Lyuba V Bozhilova, Javier Pardo-Diaz, Gesine Reinert, and Charlotte M Deane. COGENT: evaluating the consistency of gene co-expression networks. *bioRxiv*, 2020.

[3] R Karunakaran, V K Ramachandran, J C Seaman, A K East, B Mouhsine, T H Mauchline, J Prell, A Skeffington, and P S Poole. Transcriptomic analysis of Rhizobium leguminosarum biovar viciae in symbiosis with host plants Pisum sativum and Vicia cracca. *Journal of Bacteriology*, 191(12):4002–4014, 2009.

[4] Florian Klimm, Enrique M Toledo, Thomas Monfeuga, Fang Zhang, Charlotte M Deane, and Gesine Reinert. Functional module detection through integration of single-cell rna sequencing data with protein–protein interaction networks. *bioRxiv*, page 698647, 2019.

[5] Robert Petryszak, Maria Keays, Y Amy Tang, Nuno A Fonseca, Elisabet Barrera, Tony Burdett, Anja Füllgrabe, Alfonso Munoz-Pomer Fuentes, Simon Jupp, Satu Koskinen, et al. Expression atlas update—an integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Research*, 44(D1):D746–D752, 2016.

[6] Francesco Pini, Alison K East, Corinne Appia-Ayme, Jakub Tomek, Ramakrishnan Karunakaran, Marcela Mendoza-Suarez, Anne Edwards, Jason Terpolilli, Joshua Rowoth, J Allan Downie, et al. Lux bacterial biosensors for in vivo spatiotemporal mapping of root secretion. *Plant Physiology*, pages pp–01302, 2017.

[7] Carlos Prieto, Alberto Risueño, Celia Fontanillo, and Javier De Las Rivas. Human gene coexpression landscape: confident network derived from tissue transcriptomic profiles. *PloS One*, 3(12):e3911, 2008.

[8] Vinoy K Ramachandran, Alison K East, Ramakrishnan Karunakaran, J Allan Downie, and Philip S Poole. Adaptation of Rhizobium leguminosarum to pea, alfalfa and sugar beet rhizospheres investigated by comparative transcriptomics. *Genome Biology*, 12(10):R106, 2011.

[9] Damian Szklarczyk, Annika L Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, Nadezhda T Doncheva, John H Morris, Peer Bork, et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, 47(D1):D607–D613, 2019.

[10] Daniel E Wagner, Caleb Weinreb, Zach M Collins, James A Briggs, Sean G Megason, and Allon M Klein. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science*, 360(6392):981–987, 2018.

[11] Caleb Weinreb, Samuel Wolock, and Allon M Klein. Spring: a kinetic interface for visualizing high dimensional single-cell expression data. *Bioinformatics*, 34(7):1246–1248, 2018.

C

Supplementary Information: COGENT: evaluating the consistency of gene co-expression networks

COGENT: evaluating the consistency of gene co-expression networks

Supplementary Information

Lyuba V. Bozhilova[1], Javier Pardo-Diaz[1], Gesine Reinert[1], Charlotte M. Deane[1,*]
[1]Department of Statistics, University of Oxford, Oxford, OX1 3LB, United Kingdom
*deane@stats.ox.ac.uk

1 Introduction

The COGENT workflow is illustrated and explained in Section 2. Implementation details, as well as a complete list of functions available in COGENT are presented in Section 3. Network consistency measures used by COGENT are explained in detail in Section 4. Finally, an example application of COGENT is presented in Section 5. A similar, interactive application can be found in the tutorial.

2 Workflow

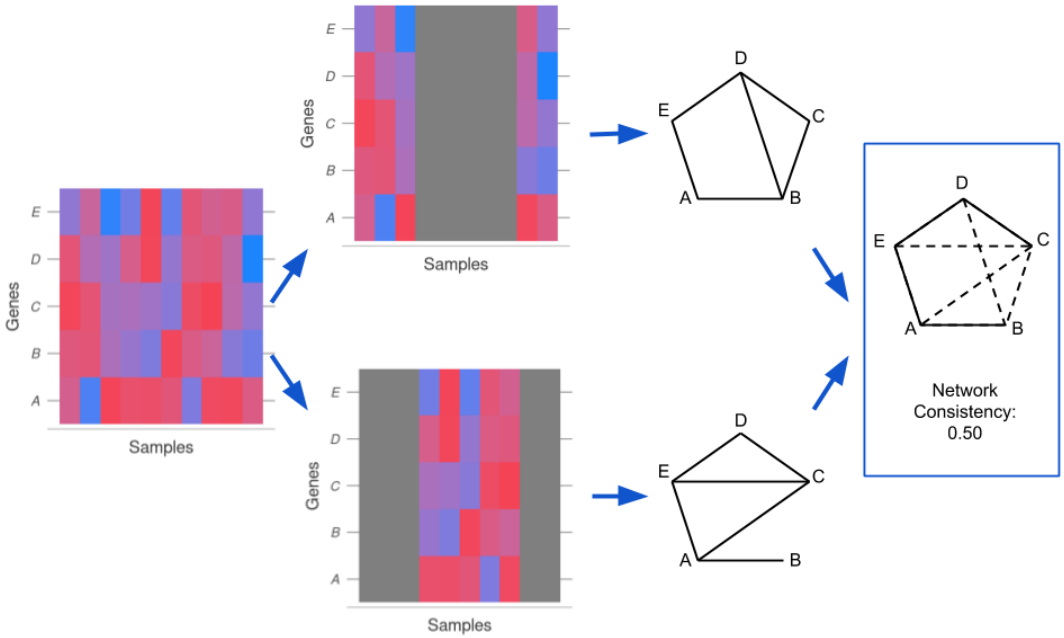


Figure 1: **COGENT workflow schematic.** In this example, the input data is a gene expression matrix with rows corresponding to genes—in this case A, B, C, D, E—and columns corresponding to samples (far left). First, the expression matrix columns are randomly split into two possibly overlapping groups of equal size (left). Then, a network is constructed from each of the sample groups (right). Finally, the two resulting networks are compared and the consistency between them is calculated (far right). In this example, the two networks have six edges each, and overlap at four of these edges. One measure of their consistency is the Jaccard index between their edge sets (see Section 4), which in this case is 0.50.

COGENT evaluates network construction methods through iterative resampling. At each step, the gene expression samples are split into two possibly overlapping sets of equal size. The overlap amount can be controlled by the user, and the sets are generated at random. In cases where a large number of samples are available, the subsets can be left disjoint.

After the two random gene expression subsets are generated, the same network construction method is applied to both in order to obtain two gene co-expression networks. Since the data is split by samples and not by genes, the two resulting networks should have identical or significantly overlapping node

sets. COGENT then calculates several measures of consistency between these two networks. This workflow is illustrated in Figure 1.

The entire process is repeated multiple times in order to obtain robust results (see Algorithm 1). Consistency can be calculated in a number of different ways, chosen and parametrised by the user. Network construction methods which result in highly similar pairs of networks are considered to be consistent. When two or more competing methods are considered, the method exhibiting higher internal consistency should be preferred.

Algorithm 1 COGENT workflow.

Require: Gene expression data M, network construction method $\psi(\cdot)$, number of repetitions n, [optional] node metric $f(\cdot)$
  for $i \in \{1, 2, \ldots, n\}$ **do**
    Randomly split M into M_1, M_2
    Create networks $G_1 = \psi(M_1)$, $G_2 = \psi(M_2)$
    Calculate the edge set consistency EC_i between G_1 and G_2
    if node metric $f(\cdot)$ is provided **then**
      Calculate the node metric consistency NC_i between metric value vectors $\overrightarrow{f(V(G_1))}$ and $\overrightarrow{f(V(G_2)}$
    end if
  end for
  return $\overrightarrow{EC} = \{EC_1, \ldots, EC_n\}$ and [optional] $\overrightarrow{NC} = \{NC_1, \ldots, NC_n\}$

3 Software description

3.1 Implementation

COGENT was built in R v.3.6.1 [11] using devtools [17] and roxygen2 [16]. It imports the core R package parallel (v.3.6.1 and higher), as well as the packages Matrix (v.1.2.17 and higher) and igraph (v.1.2.4.1 and higher), all of which are available on CRAN.

The main COGENT functions take as input gene expression data (an R object of class data frame) and a network construction method (an R object of class function) and return a set of consistency metrics. These consistency metrics aim to capture the suitability of the network construction method when applied to the data. Higher consistency implies a more robust network construction method.

3.2 Function overview

The R package COGENT includes 15 functions, listed in Table 1. They are split into four categories—input checks, internal functions, network similarity (i.e. consistency) functions, and main COGENT functions.

3.3 Input checks

The input check functions are `checkArrayList`, `checkExpressionDF`, `checkFun` and `checkMatrixList`. They are used to check whether the input (and output) of intermediate steps of the COGENT pipeline adhere to the required format.

Different types of common errors are clearly documented, so that input checks can be used for debugging before any computationally demanding analysis is performed.

3.4 Internal functions

The internal functions include alignment functions (`alignArrays` and `alignMatrices`), similarity and distance functions (`calculateEuclideanDistance` and `calculateKSimilarity`) and the random data split function (`splitExpressionData`).

Network comparison in COGENT relies on a one-to-one mapping of nodes in the pair of networks generated at each iteration. If the network construction method does not respect the order in which

Function	Description
alignArrays	Align node metric arrays
alignMatrices	Align adjacency matrices
calculateEuclideanDistance	Euclidean distance
calculateKSimilarity	Rank k-similarity
checkArrayList	Check node metric list
checkExpressionDF	Check a gene expression data frame
checkFun	Check function
checkMatrixList	Check adjacency matrix list
cogentLinear	Multiple COGENT calls, executed on a single thread
cogentParallel	Multiple COGENT calls, executed in parallel
cogentSingle	Single COGENT call
getEdgeSimilarity	Get the edge similarity for two networks
getEdgeSimilarityCorrected	Get the edge similarity with configuration model correction
getNodeSimilarity	Get the node similarity of two networks
splitExpressionData	Randomly subset expression samples

Table 1: **Functions implemented in COGENT.**

genes are labelled in the expression data set, e.g. by removing isolated nodes, an additional mapping step is needed. Similarly, if the node metric function does not respect a fixed node ordering, additional mapping may be required before the node metric comparison step. This mapping, or alignment, is not performed by default, and appropriate error messages are produced if a mismatch is detected.

The distance and similarity functions are, respectively, Euclidean distance with optional rescaling to $[0, 1]$ and parametrised rank k-similarity [13]. For rank k-similarity, the top proportion of ranks to be compared can be given as a quantile (e.g. the top 10% of nodes) or as an integer (e.g. the top 100 nodes). See Section 4.3 for details.

3.5 Network similarity

Network similarity functions include the edge set consistency functions calculated by `getEdgeSimilarity`, the density adjusted edge set consistency in `getEdgeSimilarityCorrected`, and node metric consistency functions calculated by `getNodeSimilarity`. Note that global and local similarity are calculated simultaneously by `getEdgeSimilarity`. Density adjustment in `getEdgeSimilarityCorrected` can be carried either with the fully random correction term, or with semi-random correction. Similarly, `getNodeSimilarity` calculates node metric consistency in any of the three available ways (via a correlation, rank k-similarity, or Euclidean distance). Details on the calculation of all of these are provided in Section 4.

3.6 Main functions

The main COGENT functions `cogentSingle`, `cogentLinear` and `cogentParallel` execute the complete COGENT pipeline. The principal input of these are a gene expression data set, stored in a data frame, and a network construction method, which transforms the data frame to a network adjacency matrix.

The function `cogentSingle` carries out a single iteration of COGENT. It splits the data in two random subset, constructs a network from each dataset, and returns a number of consistency metrics (e.g. global similarity, average local similarity, and rank k-similarity for some user-defined node metric). Optional parameters, e.g. whether to do any alignment or what proportion of the data is shared during the data split step, are passed to the internal and network similarity functions described in Sections 3.4 and 3.5.

The functions `cogentLinear` and `cogentParallel` are wrappers for `cogentSingle` and are used for carrying out multiple COGENT iterations. The former executes COGENT linearly, and is suitable

for cases where network construction is fast or slow, but parallelised. The latter executes COGENT in parallel on multiple threads. It is designed for cases where network construction is slow and not parallelised.

4 Network consistency

Consistency in COGENT is measured through a network comparison step at each iteration. Suppose we are interested in studying some gene expression data set M_0, and wish to construct a co-expression network G_0 from it by applying some function $G_0 = \psi(M_0)$ to the data. For example, $\psi(\cdot)$ may involve first calculating Pearson correlation coefficients between all pairs of gene expression profiles, i.e. all pairs of rows of M_0, and then selecting the top 10% of these as edges. This will result in a co-expression network, where the 10% most strongly co-expressed pairs of genes are linked by edges.

The first step of COGENT is to randomly split the data samples M_0 in two groups of equal size, M_1 and M_2. Note that the split is performed only over the samples, so both groups will contain data for all genes in M_0. By default, M_1 and M_2 do not overlap. However, a level of overlap can be set by the user if the overall number of samples in M_0 is low. The network construction function $\psi(\cdot)$ is then applied to M_1 and M_2 in order to produce two networks, $G_1 = \psi(M_1)$ and $G_2 = \psi(M_2)$. Consistency is measured by comparing G_1 and G_2. We first use edge overlap as a measure of consistency. We calculate a global edge overlap across the networks, and local overlap, which measures how well each gene neighbourhood is preserved across G_1 and G_2 [7]. Both of these can be applied to unweighted, as well as weighted networks. Since edge overlap is expected to be higher for denser networks, we introduce a density adjustment for the global edge overlap measure. The density adjusted consistency measure is only suitable for unweighted networks and can be used to inform threshold choice. Finally, we introduce an optional measure of node metric consistency. If, for example, genes of high betweenness in the co-expression network are of interest, then G_1 and G_2 can be compared by the betweenness centrality of their nodes, rather than by the overlap of their edges.

4.1 Edge set consistency

Suppose that a single iteration of COGENT applied to some some data M_0 and network construction method $\psi(\cdot)$ results in two networks $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$ with a shared node set V and edge sets E_1 and E_2. The two networks G_1 and G_2 can be considered to be similar if $E_1 \approx E_2$. If such similarity is replicated across COGENT iterations, $\psi(\cdot)$ is a consistent method for constructing networks from the data M_0. We measure agreement between E_1 and E_2 using a Jaccard index if the networks are binary (i.e. unweighted), and a weighted Jaccard index if they are weighted.

4.1.1 Unweighted and weighted Jaccard index

Let A and B be two finite non-empty sets. These could be, for example, the edge sets of two unweighted networks. The *Jaccard index* [6] between them is calculated as

$$Jacc(A, B) = \frac{|A \cap B|}{|A \cup B|}. \tag{1}$$

This is a generic metric of set overlap, and is not restricted to network comparison; it is also known as Tanimoto similarity [12, 2]. It is always in the range $[0, 1]$, with zero corresponding to no set overlap $A \cap B = \emptyset$ and one corresponding to perfect overlap $A \equiv B$.

Applied over edge sets, the Jaccard index can be used to provide a measure of both global and local network similarity, for example for community detection in multilayer networks [15, 7]. In COGENT, we generalise these to the weighted case and use them to perform network comparison between G_1 and G_2 at every iteration of our algorithm.

The Jaccard index can be used to evaluate the similarity of two edge sets, but does not take into account edge weights. A measure of set overlap which penalises weight difference is the *weighted Jaccard index*. Suppose that for some finite index set $\mathcal{I}$, we have two non-negative weight vectors $A = \{a_e\}_{e \in \mathcal{I}}$ and $B = \{b_e\}_{e \in \mathcal{I}}$. In the context of weighted networks, a_e could be the edge weight of $e \in V^2$ in the

network G_1 and b_e could be the weight of e in G_2. Non-edges are assumed to have zero weight. The weighted Jaccard index between the weight vectors A and B is

$$Jacc(A,B) = \frac{\sum_{e\in\mathcal{I}} min(a_e, b_e)}{\sum_{e\in\mathcal{I}} max(a_e, b_e)}. \tag{2}$$

Like the Jaccard index (Equation 1), the weighted Jaccard index is in the range $[0,1]$, with higher values corresponding to better agreement between A and B. Further, there is a correspondence between the two indices. If the weights are binary, then the weighted Jaccard index $Jacc(A,B)$ reduces to the unweighted $Jacc(A^*,B^*)$ between $A^* = \{e \in \mathcal{I} : a_e = 1\}$ and $B^* = \{e \in \mathcal{I} : b_e = 1\}$.

4.1.2 Global similarity

Suppose $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$ are binary, i.e. unweighted, and non-empty. The *global similarity* between them is defined as the Jaccard index between their edge sets:

$$\textit{global similarity}\ (G_1, G_2) = \frac{|E_1 \cap E_2|}{|E_1 \cup E_2|}. \tag{3}$$

It can be interpreted as the ratio between the number of edges of the intersection of G_1 and G_2 and the number of edges of their union. If the two networks are similar, then their would not be much larger than their intersection, resulting in a global similarity close to one.

This measure does not take into account edge weights. It would be unsuitable for comparing dense, possibly complete co-expression networks, in which non-negative edge weights correspond to gene co-expression. Suppose G_1 and G_2 are weighted, and that for $e \in V^2$, $w_1(e)$ is the weight of e in G_1, and w_2 is the weight of e in G_2. We treat non-edges as equivalent to zero-weight edges, i.e. $w_i(e) = 0 \iff e \notin E_i$.

Intuitively, the two weighted networks G_1 and G_2 are similar if corresponding edges have similar weights, i.e. if for most $e \in V^2$, $w_1(e) \approx w_2(e)$. To measure similarity between the networks, we use the weighted Jaccard index over the edge weights:

$$\textit{global similarity}\ (G_1, G_2) = \frac{\sum_{e\in V^2} \min(w_1(e), w_2(e))}{\sum_{e\in V^2} \max(w_1(e), w_2(e))}. \tag{4}$$

The unweighted global similarity is equivalent to the weighted one, where edges are all assigned equal weight one. More generally, in the unweighted case, a possible edge $e \in V^2$ contributes to the numerator of the fraction, i.e. the graph intersection, if it is present in both networks, and to the the denominator of the fraction, i.e. the graph union, if it is present in at least one of the networks. In the weighted case, the smaller weight contributes to the numerator—in both networks, the edge appears with weight at least $\min(w_1(e), w_2(e))$—and the larger weight contributes to the denominator—in at least one of the networks, a weight of $\max(w_1(e), w_2(e))$ is observed.

4.1.3 Local similarity

Rather than considering the complete edge sets, we can calculate a *local similarity* specific to the neighbourhood of every node in the network. Such a measure is of interest when heterogeneous noise is present. This can be the case in gene expression data, where noise is sometimes assumed to have gene-specific variance [5]. If a gene v is strongly affected by experimental noise, we would expect that its neighbourhood in a gene co-expression network to be more strongly dependent on the choice of samples used for network construction. In this case, in a single iteration of COGENT, the neighbourhoods $\Gamma_{(1)}(v)$ of v in G_1 and $\Gamma_{(2)}(v)$ of v in G_2 may differ more than the neighbourhoods of other genes. In the unweighted case [7], this is quantified by:

$$\textit{local similarity}\ (v; G_1, G_2) = \frac{|\Gamma_{(1)}(v) \cap \Gamma_{(2)}(v)|}{|\Gamma_{(1)}(v) \cup \Gamma_{(2)}(v)|}. \tag{5}$$

This is related to the global similarity (Equation 3), where instead of the complete edge sets, only the neighbourhoods of v are considered. For the weighted case, the measure can be defined analogously to Equation 4.

$$\textit{local similarity } (v; G_1, G_2) = \frac{\sum_{u \in V, u \neq v} \min(w_1(uv), w_2(uv))}{\sum_{u \in V, u \neq v} \max(w_1(uv), w_2(uv))}. \tag{6}$$

Like global similarity, local similarity for a node is in $[0, 1]$, with higher values corresponding to better preserved networks. Local similarity can be used to identify particularly consistent—or inconsistent—parts of the network, and it can also be used to provide a global network summary by averaging over all nodes [7]. The average local similarity and the global similarity are related but not necessarily identical. Their relationship is similar to that between the average local clustering coefficient and the global clustering coefficient, in that the former treats all nodes equally, and the latter is more strongly influenced by high-degree nodes.

4.1.4 Edge overlap and network density

Both global and local similarities are measured by scaling the edge overlap between two networks. However, edge overlap is affected by network density. Denser networks are more likely to share edges by chance than sparser ones. Consider, for example, two random graphs on 4 nodes, each of which has precisely one edge. Since there are $\binom{4}{2} = 6$ ways of placing the edge, the probability that the networks share an edge is $1/6$, and most of the time the graphs would have a global similarity of zero. If the two graphs instead had five randomly placed edges each, they would always share at least four edges, meaning their global similarity would be at least $2/3$. However, in both cases the overlap is entirely random—so in a sense neither pair should exhibit better "consistency".

Since COGENT works by repeatedly applying a network construction function $\psi(\cdot)$ to random subsets of the data M of equal size, we would expect the pairs of graphs $(G_1, G_2)_j$ obtained across different iterations $j \in \{1, 2, \dots\}$ to have similar density. Hence, global and local similarity can be meaningfully compared across iterations. If we consider an alternative network construction function $\xi(\cdot)$ which produces networks (H_1, H_2) of similar density to the ones produced by $\psi(\cdot)$, then we can also meaningfully compare $\psi(\cdot)$ to $\xi(\cdot)$. For example, if $\psi(\cdot)$ builds edges from the top 10% of highest Pearson correlations between gene expression profiles, and $\xi(\cdot)$ builds edges from the top 10% of highest Kendall correlations, then global similarity would be a good way of identifying which method is more consistent.

If, however, $\chi(\cdot)$ uses the top 20% of Pearson correlation coefficients instead, comparing $\psi(\cdot)$ and $\chi(\cdot)$ would be less straightforward. As $\chi(\cdot)$ produces denser networks, we would expect the global and local similarities at COGENT iterations for it to be higher. If $\chi(\cdot)$ outperformed $\psi(\cdot)$, it would not be clear whether this was due to the method being genuinely more consistent. Higher similarity scores for $\chi(\cdot)$ compared to $\psi(\cdot)$ may just arise from the increase in network density. In order to account for this, we introduce a density adjusted edge set consistency metric.

4.2 Density adjusted edge set consistency

Density adjustment is required in order to compare between network construction methods $\psi(\cdot)$ and $\chi(\cdot)$ which result in networks of considerably different edge densities. This is particularly relevant when COGENT is used to determine a threshold or score cut-off in the construction of binary networks—e.g. whether to take the top 10% or top 20% of most highly co-expressed gene pairs to build a network from. The density adjustment implemented in COGENT relies on a comparison to randomly generated networks. First we explain the fully random density adjustment and study its dependency on network density in fully randomised networks. We then introduce a semi-random adjustment, and show that both adjustments produce similar results when applied to synthetic data. Both procedures are available within COGENT, but we note that the semi-random procedure is more computationally efficient and should generally be preferred. Density adjustment is designed specifically for unweighted networks. In the case of weighted networks, a density adjustment can be carried out by transforming network weights to quantiles, and calculating weighted global and local similarities on the transformed values.

4.2.1 Fully random density adjustment

Consider a single COGENT iteration, resulting in two binary networks $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$. Let d_1 and d_2 be the degree sequences of G_1 and G_2 respectively, so $d_1(v)$ is the degree of v in G_1, and $d_2(v)$ is the degree of the same node v in G_2.

Intuitively, we would like to know how much of the overlap between E_1 and E_2 is "genuine", and how much of it we might expect from any two networks of similar degree sequences. Density adjustment in COGENT corrects for the effect of random overlap by generating two networks, $G_1^* = (V, R_1)$ and $G_2^* = (V, R_2)$ from a type of configuration model, with random permutations d_1^* and d_2^* of the degree sequences d_1 and d_2 respectively. The degree sequences are permuted, since in the original network nodes correspond to genes and are therefore labelled. The networks G_1^* and G_2^* are generated with fixed degree sequences, no loops and no multiple edges, using the implementation by Csardi and Nepusz [4]. We define a correction term based on two pairwise comparisons between the co-expression and randomly generated networks: one between G_1 and G_2^*, and one between G_2 and G_1^*:

$$\alpha(G_1, G_2|G_1^*, G_2^*) = \frac{1}{2}(|E_1 \cap R_2| + |E_2 \cap R_1|). \tag{7}$$

Heuristically, the value of α corresponds to a level of intersection between G_1 and G_2 we may expect at random. Each of the terms in the sum is the overlap between one "real" co-expression network and a random network which resembles its counterpart. To define the *adjusted consistency* between G_1 and G_2 we subtract $\alpha(G_1, G_2|G_1^*, G_2^*)$ from the graph intersection and add it to the graph union of G_1 and G_2 in the unweighted global similarity (Equation 3), to give what we call the *fully random adjusted consistency*:

$$\textit{α-adjusted consistency } (G_1, G_2|G_1^*, G_2^*) = \frac{|E_1 \cap E_2| - \alpha(G_1, G_2|G_1^*, G_2^*)}{|E_1 \cup E_2| + \alpha(G_1, G_2|G_1^*, G_2^*)}. \tag{8}$$

Unlike global and local similarity, the density adjusted consistency measure can be negative. The lowest possible adjusted consistency is obtained when G_1 and G_2 share no edges, but $G_1^* \equiv G_2$ and $G_2^* \equiv G_1$. In this case, the adjusted consistency is $-1/3$. Perfectly overlapping G_1 and G_2 will have adjusted consistency close to 1, with higher consistency for lower density networks. If network overlap is random, then $|E_1 \cap E_2| \approx \alpha(G_1, G_2|G_1^*, G_2^*)$ and the adjusted consistency would be close to 0.

We illustrate that the fully random consistency in Equation 8 does not increase with density like the global and local similarities with a simulation study. We generated pairs networks G_1 and G_2 from the $\mathcal{G}(N, p)$ model on $N = 100$ nodes and with edge probability $p \in \{0.01, 0.02, \ldots, 0.50\}$. For each value of p we generated 100 pairs (G_1, G_2) and calculated the global similarity and the fully random adjusted consistency of each one. This resulted in a total of 5000 pairs of networks. Since the networks are random, we observed adjusted consistencies close to zero at all values of p. Further, there was no relationship between the adjusted consistencies and p (Pearson correlation coefficient 0.025). In contrast, the global similarity increased with p (Pearson correlation coefficient 0.994). See Figure 2 for details.

Generating the pair of random networks G_1^* and G_2^* can be computationally expensive when the number of nodes in the network is large. This is due to the random graph generation algorithm: first, "stubs" are attached to every node in the network, and then pairs of stubs are randomly connected to build edges. However, every time the procedure creates a self-edge (two stubs of the same node are connected), or a multi edge (two stubs are connected between the pair of nodes which already share an edge), the process is restarted. In order to speed up the comparison, we introduce a semi-random density adjustment, based on approximate expectations of edge occurrence rather than on randomly generated networks.

4.2.2 Semi-random density adjustment

In Equation 7 we introduced a density adjustment term by randomising the degree sequences of $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$ and creating two random networks $G_1^* = (V, R_1)$ and $G_1^* = (V, R_2)$ with the same degree distributions as G_1 and G_2 respectively. We then calculated the graph intersection

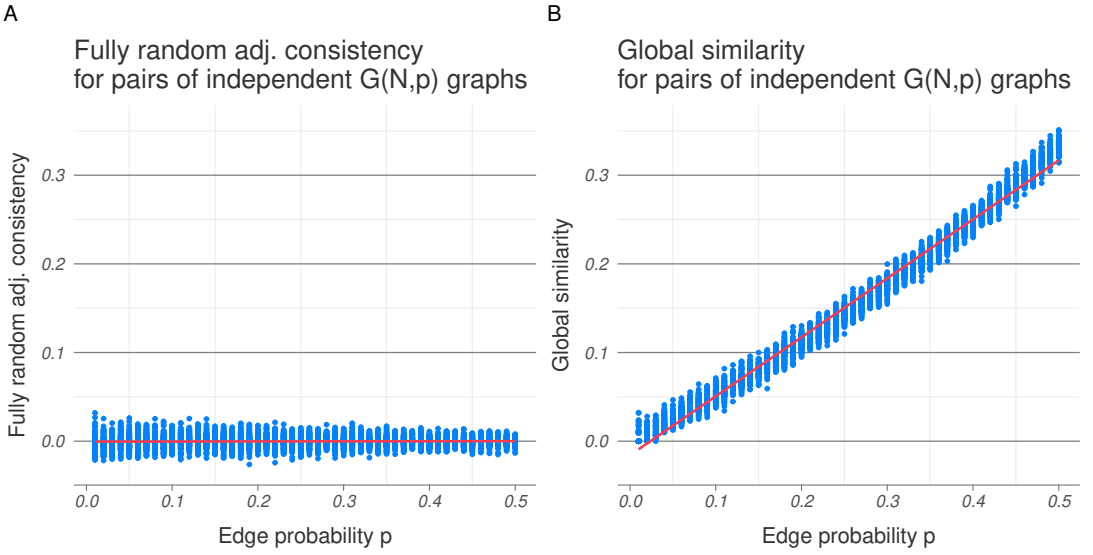


Figure 2: **Fully random adjusted consistency of $\mathcal{G}(N,p)$ pairs.** For different values of p, one hundred pairs of networks were generated from $\mathcal{G}(100,p)$ and the fully random adjusted consistency (A) and global similarity (B) were calculated for each pair. The dots (blue) correspond to the obtained values and the lines (red) were fitted using least squares regression. Consistency remained close to zero far all network densities, while global similarity increased with p.

between the fixed G_1 and the random G_2^*, as well as between G_2 and G_1^*. These intersections are both also random graphs by construction. Based on the intersections, we calculated a correction term α.

Rather than generating the random graph intersections, we can calculate a correction term β analogous to α by approximating their expected size. Without loss of generality, consider the intersection between G_1 and G_2^*. First, note that if an edge $e=(u,v)$ is not present in G_1, i.e. $e \notin E_1$, then it clearly cannot be present in the intersection either. Therefore, with G_1 fixed,

$$\mathbb{P}(e \in E_1 \cap R_2) = \mathbb{1}\{e \in E_1\}\mathbb{P}(e \in R_2). \tag{9}$$

The graph G_2^* is generated from a configuration model, after discarding any networks with self-loops and multi-edges. However, for sparse, large networks, the density of self-loops and multi-edges in the configuration model is low. Therefore, we can approximate the probability of an edge $e=(u,v)$ existing in G_2^* with the probability of edge occurrence in the standard Molloy–Reed configuration model, which is proportional to $d_2^*(u)d_2^*(v)$ [9]. We normalise so that the expected number of edges in the network is $|R_2| = |E_2| = \frac{1}{2}\sum_u d_2^*(u)$ to get

$$\mathbb{P}(e=(u,v) \in R_2 | d_1^*, d_2^*) \approx \frac{d^*(u)d^*(v)}{K} \text{ for } u \neq v, \tag{10}$$

where the normalising constant K is

$$K = \frac{1}{|R_2|}\sum_{u>v} d^*(u)d^*(v) = \frac{1}{|E_2|}\sum_{u>v} d(u)d(v). \tag{11}$$

Given a random degree permutation $\{d_2^*(u)\}_{u\in V}$, the probability in Equation 9 for $u \neq v$ can be approximated by

$$\mathbb{P}(e=(u,v) \in E_1 \cap R_2 | d_1^*, d_2^*) \approx \mathbb{1}\{e \in E_1\}\frac{|E_2| d_2^*(u)d_2^*(v)}{\sum_{u>v} d_2^*(u)d_2^*(v)} =: p_e. \tag{12}$$

This allows us to calculate an approximation to the expected adjacency matrix of the intersection between G_1 and G_2^*, given the degree permutation used for generating G_2^*. This adjacency matrix

can be thought of as a weighted network, where edge weights p_{uv} correspond to probability of edge occurrence. The expected number of edges in the overlap is then approximately the sum of all p_{uv}:

$$O_{12} = \sum_e p_e \propto \sum_{e=(u,v)\in E_1} d_2^*(u) d_2^*(v). \tag{13}$$

We can approximate the overlap O_{21} between G_2 and G_1^* analogously, replacing the degree sequence d_2 with a random permutation d_1 of the degree sequence of G_1. Using both, we define a new correction term β:

$$\beta(G_1, G_2 | d_1^*, d_2^*) = \frac{1}{2}(O_{12} + O_{21}). \tag{14}$$

The new correction term $\beta(G_1, G_2|d_1^*, d_2^*)$ approximates the expectation of $\alpha(G_1, G_2|G_1^*, G_2^*)$ given that the degree sequence permutations d_1^* and d_2^* were used to generate G_1^* and G_2^* respectively. We call this correction term "semi-random" because while it does not require the generation of the graphs G_1^* and G_2^*, it still requires the random permuting of the degree sequences d_1 and d_2. Finally, we use β to define a *semi-random adjusted consistency* measure, similar to the fully random one in Equation 8:

$$\beta\text{-adjusted consistency } (G_1, G_2|d_1^*, d_2^*) = \frac{|E_1 \cap E_2| - \beta(G_1, G_2|d_1^*, d_2^*)}{|E_1 \cup E_2| + \beta(G_1, G_2|d_1^*, d_2^*)}. \tag{15}$$

In its implementation, the semi-random adjusted consistency relies entirely on matrix algebra and array permutations, rather than on network generation. This makes it more computationally efficient than the fully random one.

To evaluate its performance, semi-random adjusted consistency was measured for the set of $\mathcal{G}(N, p)$ pairs of networks discussed in the previous section. Like the fully random adjusted consistency, its values were close to zero and did not increase with network density. Further, they correlated well with the fully random values (Pearson correlation coefficient 0.807). Therefore, the approximations made are appropriate. See Figure 3 for details.

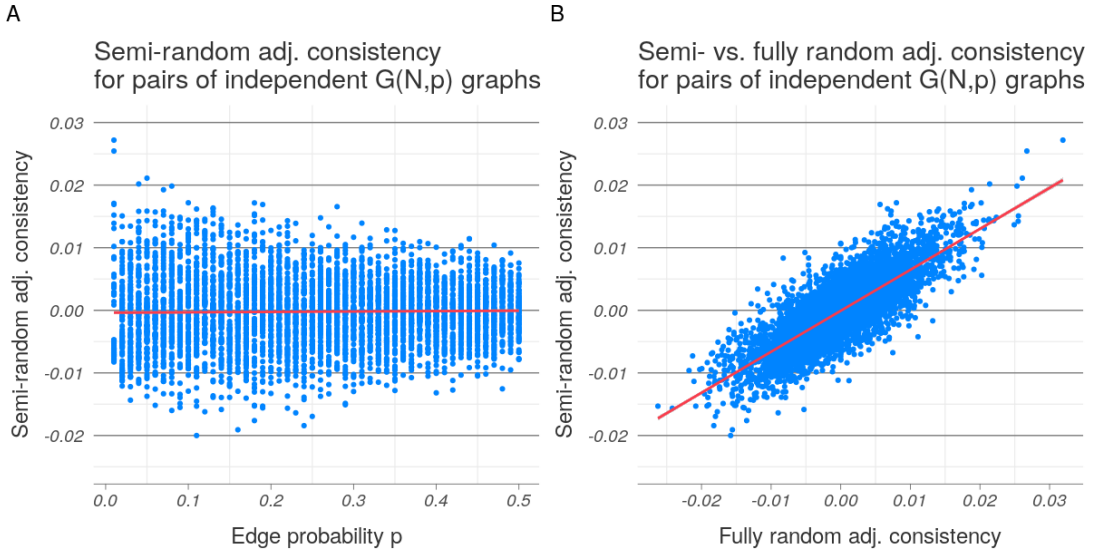


Figure 3: **Semi-random adjusted consistency of $\mathcal{G}(N, p)$ pairs.** (A) Like the fully random adjusted consistency (Figure 2A), the semi-random measure remained close to zero as the edge probability p increased. (B) The fully and semi-random measures were well-correlated. The dots (blue) correspond to the obtained values and the lines (red) were fitted using least squares regression.

Overall, we have demonstrated that both the fully random and the semi-random density adjusted consistency measures can be used to measure the agreement between two binary networks, controlling for their densities. This makes them suitable for comparing network construction algorithms resulting

in binary networks of significantly different densities. Like global and local similarity, density adjusted consistency is a measure of edge overlap. Within COGENT, we introduce a further way of evaluating network consistency using node metric agreement.

4.3 Node metric consistency

Edge overlap is not the only way to think about network similarity. For example, consider two transport networks of Britain, where nodes correspond to villages, towns and cities, and edges are either roads or direct railway connections. We may argue that the railway network is in a sense functionally similar to the road network, since both identify the same large cities as major transport hubs. This similarity is between the node properties of the two networks, and is not based on edge overlap. It also depends on the node metric of choice—while the high-degree nodes might be the same across the two networks, the high betweenness nodes may be different.

Both degree and betweenness centrality as well as other node metrics are used in co-expression network analysis to identify key functionally relevant genes, e.g. [14]. It therefore may be desirable to study the consistency of a network construction method with respect to these metrics. That is, we may ask how much we can trust node metric values obtained from the co-expression network, rather than how reliable the network itself is. Node metric consistency can be calculated in three different ways within COGENT.

Suppose again that a COGENT iteration results in a pair of networks $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$, and let $f : V(G) \to \mathbb{R}$ be a node metric of interest. A network construction method may be considered consistent *with respect to the node metric* f if $f(v|G_1) \approx f(v|G_2)$ for all $v \in V$. COGENT can measure the agreement between $f(v|G_2)$ and $f(v|G_2)$ in three different ways:

- through Euclidean distance,
- through a correlation coefficient, and
- through rank k-similarity [13, 3].

The Euclidean distance measures the distance between the vectors $f(V(G_1))$ and $f(V(G_2))$ in $\mathbb{R}^V$ space. The obtained value is non-negative, with *lower* values corresponding to better network consistency. Optionally, a scaled Euclidean distance can be calculated, for which metric values are first scaled to the range $[0, 1]$ using

$$f^*(u) = \frac{f(u) - \min_{v \in V}\{f(v)\}}{\max_{v \in V}\{f(v)\} - \min_{v \in V}\{f(v)\}} \tag{16}$$

and then the Euclidean distance between $f(V(G_1))^*$ and $f(V(G_2))^*$ is itself scaled between 0 and 1 using

$$d(f^*(V(G_1)), f^*(V(G_2))) = \frac{||f^*(V(G_1)) - f^*(V(G_1))||}{\sqrt{|V|}}. \tag{17}$$

This results in a measure between 0 and 1, where again *lower* values correspond to higher network consistency.

The correlation between the node metric vectors $f(V(G_1))$ and $f(V(G_2))$ can also be calculated. By default, Pearson correlation coefficients are calculated, and missing data is handled by pairwise deletions. That is, only nodes where a metric value was obtained in both G_1 and G_2 are used. However, in our implementation options are directly passed to the R-base correlation function $cor(\cdot)$, which allows Pearson, Spearman and Kendall correlation coefficients to be calculated, and missing data to be handled in different ways.

Rank k-similarity [13] can also be used to evaluate network consistency within COGENT. Briefly, rank k-similarity measures how well the sets of highest ranking nodes with respect to the metric $f(\cdot)$ in G_1 and G_2 agree. For example, it can be used to say what proportion of the top 50 highest degree nodes in G_1 are also among the highest degree nodes in G_2. The default value of k in COGENT is 10% of the number of genes $|V|$.

Note that, since k-similarity is rank-based, it is also not dependent on the network density. More generally, if the node metric $f(\cdot)$ used or the consistency measure applied to it is not dependent on network density, node metric consistency can be used instead of density adjusted edge set consistency to choose between network construction methods of different densities.

Together, the different edge set and node metric consistency measures provide a profile of the network similarity between G_1 and G_2 at every COGENT iteration. By aggregating across multiple iterations, COGENT outputs a consistency profile of a network construction method $\psi(\cdot)$ when applied to a gene expression data set M. In the next section we illustrate how COGENT can be used to inform network construction in different cases. We focus on unweighted networks; applications with weighted networks are analogous.

5 Example application

The network comparison measures described in Section 4 can be used to evaluate the consistency of a network construction method $\psi(\cdot)$ as applied to a gene expression data set M. Comparing the consistency of different methods $\psi(\cdot)$ and $\xi(\cdot)$ allows the user to prioritise a construction method for further analysis. This prioritisation can happen with respect to overall network consistency, or with respect to a particular application. For example, if genes of high betweenness centrality are of interest, network consistency can be measured with respect to betweenness centrality. Our method allows the user to assess network quality without the use of any additional data, such as protein interaction or functional annotation data.

In this section we illustrate how COGENT can be used in two different scenarios—to choose between measures of gene co-expression for network construction and to set a co-expression score cut-off. In both applications we focus on a gene expression dataset obtained from the Expression Atlas [10].

5.1 Gene expression data

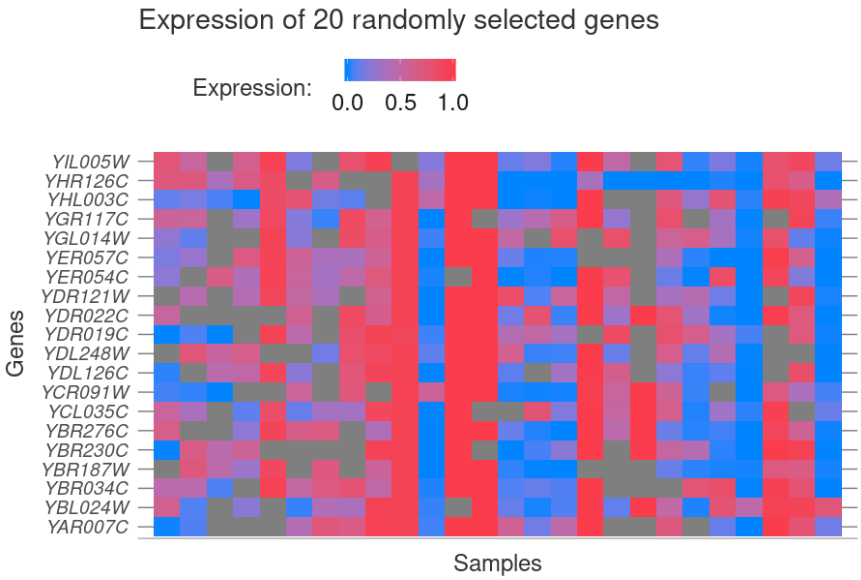


Figure 4: **Sample gene expression data for yeast.** Rows correspond to genes and columns correspond to samples. Expression values are scaled between 0 and 1, with low expression shown in blue and high expression shown in red. Missing values are shown in grey. Rows with similar colour patterns indicate high gene co-expression.

RNA-seq data of *Saccharomyces cerevisiae* (yeast) expressing pathways designed to increase ATP or GTP consumption was obtained from the Expression Atlas (accession number E-MTAB-5174). The

original experiment as deposited in ArrayExpress [1] contains 156 samples, sequenced using Illumina HiSeq 2500. The curated data obtained from the Expression Atlas was reduced to expression profiles for 3103 genes across 26 samples. We further removed any genes where expression data was missing for over 25% of all samples, resulting in a dataset of 1920 gene expression profiles. Finally, we took the 500 genes with highest variation in gene expression as measured by the mean absolute deviation. This resulted in a final dataset M of 500 gene expression profiles across 26 samples. Expression values for 20 randomly selected genes are shown in Figure 4.

5.2 Choosing between measures of co-expression

We consider two simple ways of building gene co-expression networks from the data described in Section 5.1. The first builds a network by calculating Pearson correlation coefficients between expression profiles and creating edges between the gene pairs with highest co-expression. This is sometimes known as hard thresholding [8]. In the first instance, we take the top 5% of the co-expression values between pairs of genes, which results in a gene co-expression network on 500 nodes and 6238 edges. Note that 5% of all possible gene pairs is $5\%\binom{500}{2} = 6237.5$. A total of 49 nodes in the network are isolated.

The second construction method uses Kendall correlation coefficients instead of Pearson correlation coefficients. Again, we take the top 5% of obtained values to correspond to edges, resulting in a second network on 500 nodes and 6238 edges. Of the 58 isolated edges in the Kendall network, 43 are also isolated in the Pearson network.

We can use COGENT both to examine the consistency of a network construction method and to compare the Pearson and Kendall networks. The global similarity between the two networks is 0.57, corresponding to an overlap of 4512 edges, or 72% of the edges in each network. Edge differences are not spread uniformly across the networks—higher degree nodes generally have higher local similarity and nodes of extremely low degree in either network account for peaks at 0 and 1 (see Figure 5).

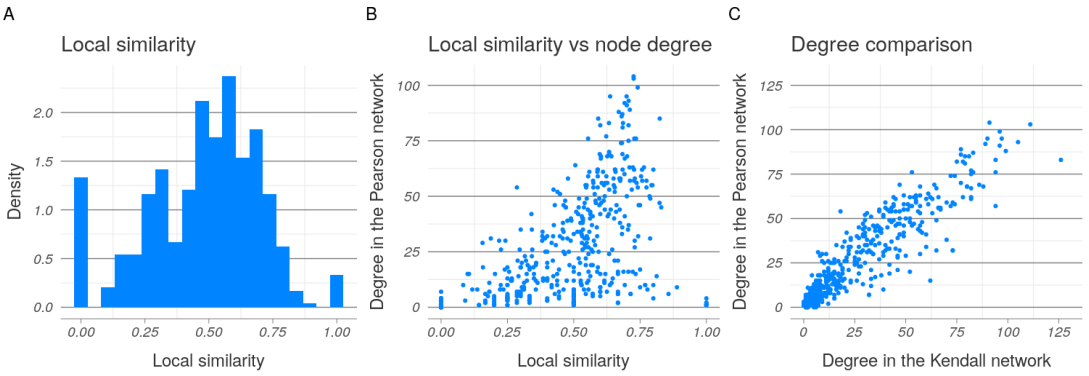


Figure 5: **Similarities between the Pearson and Kendall co-expression networks.** (A) Local similarity for each node (mean 0.48, st. dev. 0.22). Most values are symmetric around the mean, with additional peaks around 0 and 1. Binwidth is set to 0.05. (B) Local similarity plotted against node degree in the Pearson network. The peaks at 0 and 1 correspond to nodes of low degree. However, higher degree nodes tend to exhibit higher similarity. (C) The two degree distributions are positively correlated (Pearson correlation coefficient 0.92).

While the Pearson and Kendall networks share a significant number of edges, they are considerably different. We use COGENT to determine which of the two network correlation calculations is more consistent with respect to data resampling and which network should therefore be prioritised for further analysis. Since the networks are of the same density 0.05 by construction, density adjustment is not required. For illustration purposes, we evaluate edge set consistency through global similarity and node metric consistency for the degree through Pearson's correlation coefficient.

Since the dataset contains only 26 samples, we run COGENT for both methods with a sample overlap of 50% of the data at each iteration. Thus, when the data M is resampled and split into subsets M_1 and M_2, 13 samples are randomly selected and included in both sets, and the remaining samples are randomly allocated to only one of the sets. One hundred COGENT iterations are used

to evaluate the stability of each method. The computational complexity of the COGENT pipeline is linear in the complexity of the network construction method (and the node metric calculation). Thus, medium-to-large Pearson correlation networks can be analysed without any parallelisation on a standard desktop computer. COGENT has an additional inbuilt parallel mode, which is employed for the Kendall networks, where computing correlation matrices is slower.

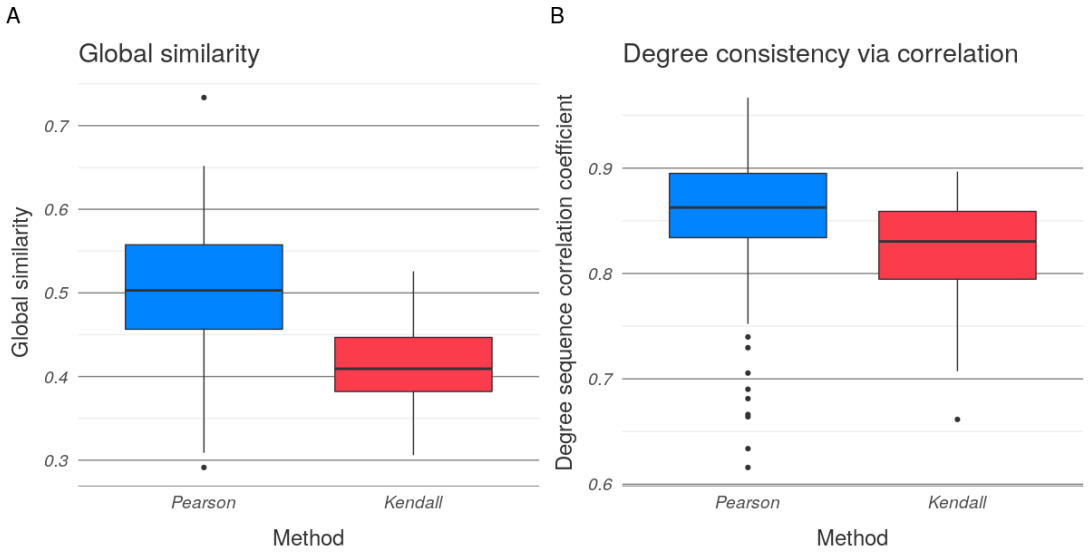


Figure 6: **COGENT analysis for Pearson and Kendall co-expression networks.** (A) Global similarity is visibly higher for the Pearson network construction method (mean 0.50, st. dev. 0.08) than for the Kendall method (mean 0.41, st. dev. 0.05). (B) While Pearson (mean 0.85, st. dev. 0.07) still outperforms Kendall (mean 0.82, st. dev. 0.04) in rank k-similarity for the degree, the difference is slightly less pronounced. This may be expected, since the two networks have similar degree distributions (Figures 5B). Note the difference in scales on the y-axis.

Edge set consistency as measured by global similarity is higher for the Pearson than the Kendall method (Figure 6). Pearson also outperforms Kendall in rank k-similarity for the degree with the default settings. While the difference is less pronounced for k-similarity than it is for global similarity, both are statistically significant. A two-sided Wilcoxon rank sum test for global similarity has p-value $p \approx 2.3 \times 10^{-15}$, and the same test for degree correlations has p-value $p \approx 4.2 \times 10^{-7}$.

The comparison between network consistency for both Pearson and Kendall correlation networks therefore indicates that the Pearson network should be preferred. However, this analysis was performed only on networks thresholded to ensure a 5% network density. In the next section, we use COGENT to study threshold choice.

5.3 Imposing a co-expression score cut-off

In the previous section, we used COGENT to compare the consistency of two methods which both produced networks of 5% edge density. However, the value of 5% was arbitrary. We can also use COGENT in order to determine whether a different value may result in more consistent networks. We illustrate this by exploring different thresholds for a Pearson correlation network constructed using the same gene expression data.

As discussed in Figure 2, global similarity cannot be used to distinguish between methods with different network densities, because denser networks are more likely to have higher edge overlap by chance. In order to correct this, we use the semi-random density-adjusted edge set consistency measure in COGENT. We consider 50 density thresholds between 0.01 and 0.50, i.e. from taking only the top 1% of highest pairwise gene expression correlations down to taking the top half of these. At each threshold, we run the COGENT pipeline for 100 iterations, and split the data the same sample overlap as before (13 out of 26 shared samples in each sample subset).

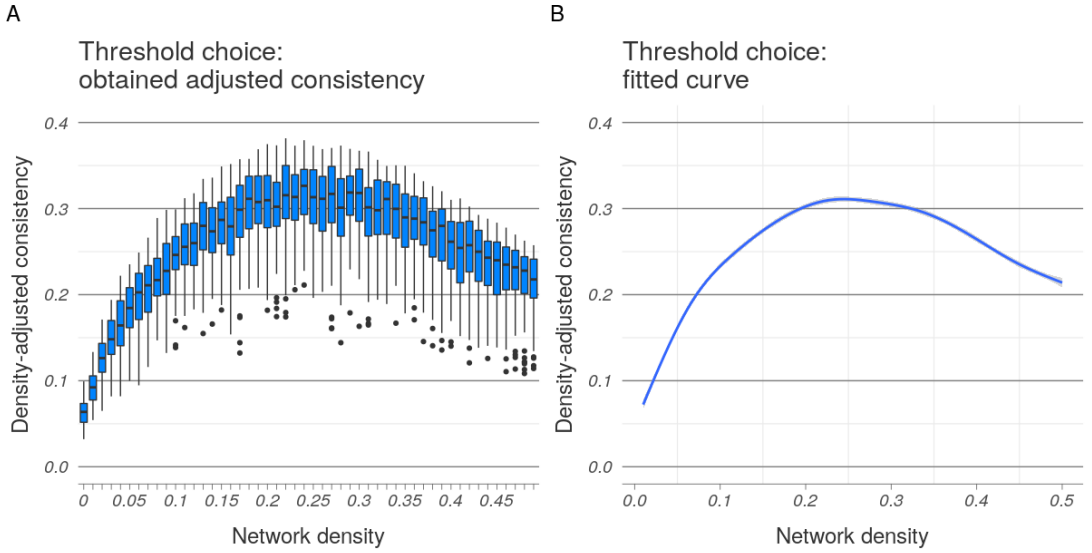


Figure 7: **COGENT analysis for Pearson thresholds.** (A) Obtained density adjusted edge set consistencies at different network densities. (B) Curve fitted to the data using a generalised additive model. The curve reaches its maximum around density 0.25, suggesting taking the top 25% highest correlation values may be optimal.

The density adjusted edge set consistency for the Pearson network at the 5% density threshold has mean 0.16 and standard deviation 0.03. As expected due to the semi-random correction term β, this is lower than the global similarity calculated using the same parameters (mean 0.50, standard deviation 0.08). As the network density increases, the adjusted similarity also increases until it plateaus for densities around 25%. For higher densities, the adjusted consistency then decreases (Figure 7). Fixing the network density around 25% may therefore be most preferable.

The highest observed adjusted consistency is at network density 25% (mean 0.32 and std. dev. 0.04). This is a significant improvement over the value at 5% density. However, differences in the plateau region are not necessarily statistically significant—a one-sided Wilcoxon's rank test ($\alpha = 5\%$) showed six other thresholds between 0.23 and 0.31 resulted in consistency values which were not significantly lower than those at the optimal threshold 0.25. Therefore, while 25% density results in networks of the highest consistency, COGENT can be used to identify a range of permissible thresholds.

We have illustrated the two main expected uses of COGENT with respect to gene co-expression data: to choose between different measures of co-expression, and to inform threshold choice for binary networks. Both of these problems arise when networks are constructed from other types of occurrence data.

References

[1] Awais Athar, Anja Füllgrabe, Nancy George, Haider Iqbal, Laura Huerta, Ahmed Ali, Catherine Snow, Nuno A Fonseca, Robert Petryszak, Irene Papatheodorou, et al. ArrayExpress update–from bulk to single-cell expression data. *Nucleic Acids Research*, 47(D1):D711–D715, 2018.

[2] Dávid Bajusz, Anita Rácz, and Károly Héberger. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics*, 7(1):20, 2015.

[3] Lyuba V Bozhilova, Alan V Whitmore, Jonny Wray, Gesine Reinert, and Charlotte M Deane. Measuring rank robustness in scored protein interaction networks. *BMC Bioinformatics*, 20(1):446, 2019.

[4] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006.

[5] Chuan Gao, Ian C McDowell, Shiwen Zhao, Christopher D Brown, and Barbara E Engelhardt. Context specific and differential gene co-expression networks via Bayesian biclustering. *PLoS Computational Biology*, 12(7):e1004791, 2016.

[6] Paul Jaccard. Nouvelles recherches sur la distribution florale. *Bulletin de la Societe Vaudoise des Sciences Naturelles*, 44:223–270, 1908.

[7] Ta-Chu Kao and Mason A Porter. Layer communities in multiplex networks. *Journal of Statistical Physics*, 173(3-4):1286–1302, 2018.

[8] Peter Langfelder and Steve Horvath. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1):559, 2008.

[9] Mark EJ Newman. *Networks*. Oxford University Press, Second edition, 2018.

[10] Irene Papatheodorou, Nuno A Fonseca, Maria Keays, Y Amy Tang, Elisabet Barrera, Wojciech Bazant, Melissa Burke, Anja Füllgrabe, Alfonso Muñoz-Pomer Fuentes, Nancy George, et al. Expression atlas: gene and protein expression across multiple studies and organisms. *Nucleic Acids Research*, 46(D1):D246–D251, 2017.

[11] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019.

[12] Taffee T Tanimoto. IBM internal report. *Nov*, 17:1957, 1957.

[13] Stojan Trajanovski, Javier Martín-Hernández, Wynand Winterbach, and Piet Van Mieghem. Robustness envelopes of networks. *Journal of Complex Networks*, 1(1):44–62, 2013.

[14] Sipko van Dam, Urmo Vosa, Adriaan van der Graaf, Lude Franke, and Joao Pedro de Magalhaes. Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in Bioinformatics*, 19(4):575–592, 2017.

[15] András Vörös and Tom AB Snijders. Cluster analysis of multiplex networks: Defining composite network measures. *Social Networks*, 49:93–112, 2017.

[16] Hadley Wickham, Peter Danenberg, and Manuel Eugster. *roxygen2: In-Line Documentation for R*, 2018. R package version 6.1.1.

[17] Hadley Wickham, Jim Hester, and Winston Chang. *devtools: Tools to Make Developing R Packages Easier*, 2019. R package version 2.2.1.

D

Supplementary Information: Analysis and evaluation of gene coexpression networks for *Rhizobium leguminosarum*

Supplementary Information: Extracting information from gene coexpression networks of *Rhizobium leguminosarum*

SI A Datasets used to construct networks

In this paper we analyse two different gene coexpression networks: standard network and complete network, constructed from two different microarray datasets. Table SI 1 shows some main network summaries of both networks. Table SI 2 presents the conditions considered for the standard network. Table SI 3 gives the conditions considered for the complete network.

Network	Edges	Average weight	Average degree	Average strength
Standard	388374	0.679	109.75	74.51
Complete	687497	0.604	194.29	117.47

Table SI 1: Network summaries of the two networks.

Microarray (channel)	**Description**	**Type**
kk21 (Cy3), kk41 (Cy5), kk45 (Cy3)	Succinate NH_4	F
Jay10 (Cy5), Jay11 (Cy5), Jay12 (Cy3)	Glucose Glutamate	FL
Jay10 (Cy3), Jay11 (Cy3), Jay12 (Cy5)	Glucose Aspartate	FL
kk46 (Cy3), kk48 (Cy3), kk50 (Cy3)	Pyruvate NH_4	FL
Vinoy43 (Cy5), Vinoy44 (Cy5), Vinoy45 (Cy3)	Pyruvate NH_4 Hespertin	FL
Ade7 (Cy3), Ade9 (Cy3), Ade18 (Cy5)	Pyruvate NH_4 IAA	FL
Ade8 (Cy5), Ade11 (Cy3), Ade15 (Cy5)	Pyruvate NH_4 Kinetin	FL
kk39 (Cy5), kk43 (Cy5), kk51 (Cy3)	Inositol NH_4	FL
Vinoy26 (Cy5), Vinoy31 (Cy5), Vinoy34 (Cy3)	7 days pea, 1 day PI	R
Vinoy27 (Cy5), Vinoy32 (Cy5), Vinoy35 (Cy3)	7 days pea, 3 day PI	R
Vinoy28 (Cy5), Vinoy33 (Cy5), Vinoy36 (Cy3)	7 days pea, 7 day PI	R

Microarray (channel)	**Description**	**Type**
Vinoy29 (Cy5), Vinoy37 (Cy5), Vinoy39 (Cy3)	14 days pea, 1 day PI	R
Vinoy30 (Cy5), Vinoy38 (Cy5), Vinoy40 (Cy3)	21 days pea, 1 day PI	R
kk67 (Cy5), kk71 (Cy3), kk72 (Cy5)	1 week pea, 7 days PI	B
kk65 (Cy5), kk68 (Cy3), kk69 (Cy3)	1 week pea, 15 days PI	B
kk73 (Cy5), kk74 (Cy3), kk75 (Cy5)	1 week pea, 21 days PI	B
kk64 (Cy5), Vinoy22 (Cy3), Vinoy24 (Cy3)	1 week pea, 28 days PI	B
kk70 (Cy5), kk60 (Cy3), kk61 (Cy5)	Vetch seed, 28 days PI	B

Table SI 2: List of microarrays included in the *standard dataset*. The second column in the table indicates the growth condition of the bacteria; for example, "7 days pea, 1 day PI" stands for inoculation of seven days old pea plant sampled one day post-inoculation. The last column indicates the type of sample, which can be "FL" (free-living), "R" (rhizosphere), or "B" (bacteroid).

Microarray (channel)	**Description**	**Type**
kk21 (Cy3), kk41 (Cy5), kk45 (Cy3)	Succinate NH_4	FL
Jay10 (Cy5), Jay11 (Cy5), Jay12 (Cy3)	Glucose Glutamate	FL
Jay10 (Cy3), Jay11 (Cy3), Jay12 (Cy5)	Glucose Aspartate	FL
kk46 (Cy3), kk48 (Cy3), kk50 (Cy3), Nefali1 (Cy3), Nefali2 (Cy3), Nefali3 (Cy3),Vinoy47 (Cy3), Vinoy48 (Cy3)	Pyruvate NH_4	FL
Vinoy43 (Cy5), Vinoy44 (Cy5), Vinoy45 (Cy3)	Pyruvate NH_4 Hespertin	FL
Ade7 (Cy3), Ade9 (Cy3), Ade18 (Cy5)	Pyruvate NH_4 IAA	FL
Ade8 (Cy5), Ade11 (Cy3), Ade15 (Cy5)	Pyruvate NH_4 Kinetin	FL
kk39 (Cy5), kk43 (Cy5), kk51 (Cy3)	Inositol NH_4	FL
Ade5 (Cy5), Ade16 (Cy5)	ABA Pyruvate NH_4	FL
Alastair1 (Cy5)	Formate Pyruvate	FL
Alastair2 (Cy3)	Acetoacetate NH_4	FL
kk19 (Cy3), kk20 (Cy3)	Succinate GABA	FL
kk98 (Cy5)	Acetate NH_4	FL
kk100 (Cy3)	Proline Pyruvate	FL

Microarray (channel)	Description	Type
kk101 (Cy3)	Phenylalanine Pyruvate	FL
kk103 (Cy3)	Arabinogalactan Pyruvate NH_4	FL
kk104 (Cy3)	Arabinose NH_4	FL
kk105 (Cy3)	Galactose NH_4	FL
kk110 (Cy5)	TY	FL
Nefali1 (Cy5), Nefali2 (Cy5), Nefali3 (Cy5)	Iron limited	FL
Vinoy25 (Cy5), Vinoy41 (Cy5), Vinoy42 (Cy3)	Root juice Pyruvate NH_4	FL
Vinoy47 (Cy5)	Protocatechuate	FL
Vinoy48 (Cy5)	4-hydroxybenzoate	FL
Vinoy71 (Cy5), Vinoy72 (Cy3), Vinoy75 (Cy5)	7 days alfalfa, 7 day PI	R
Vinoy73 (Cy3), Vinoy74 (Cy3), Vinoy76 (Cy5)	7 days sugar beet, 7 day PI	R
Vinoy26 (Cy5), Vinoy31 (Cy5), Vinoy34 (Cy3)	7 days pea, 1 day PI	R
Vinoy27 (Cy5), Vinoy32 (Cy5), Vinoy35 (Cy3)	7 days pea, 3 day PI	R
Vinoy28 (Cy5), Vinoy33 (Cy5), Vinoy36 (Cy3)	7 days pea, 7 day PI	R
Vinoy29 (Cy5), Vinoy37 (Cy5), Vinoy39 (Cy3)	14 days pea, 1 day PI	R
Vinoy30 (Cy5), Vinoy38 (Cy5), Vinoy40 (Cy3)	21 days pea, 1 day PI	R
kk67 (Cy5), kk71 (Cy3), kk72 (Cy5)	1 week pea, 7 days PI	B
kk65 (Cy5), kk68 (Cy3), kk69 (Cy3)	1 week pea, 15 days PI	B
kk73 (Cy5), kk74 (Cy3), kk75 (Cy5)	1 week pea, 21 days PI	B
kk64 (Cy5), Vinoy22 (Cy3), Vinoy24 (Cy3)	1 week pea, 28 days PI	B
kk70 (Cy5), kk60 (Cy3), kk61 (Cy5)	Vetch seed, 28 days PI	B

Table SI 3: List of microarrays included in the *complete dataset*. The second column in the table indicates the growth condition of the bacteria; for example, "7 days pea, 1 day PI" stands for inoculation of seven days old pea plant sampled one day post-inoculation. The last column indicates the type of sample, which can be "FL" (free-living), "R" (rhizosphere), or "B" (bacteroid).

SI B Ribosomal genes

The complete list of the 49 ribosomal genes used as control genes in Section 4.2 and as seed list in Section 4.3 is given in Table SI 4.

Gene ID	Gene Name	Gene Description	FL	R	B
RL0267	*rpmI*	putative 50S ribosomal protein L35	ES	ES	NE
RL0268	*rplT*	putative 50S ribosomal protein L20	ES	ES	NE
RL0374	*rpsT*	putative 30S ribosomal protein S20	ES	ES	NE
RL1552	*rplI*	putative 50S ribosomal protein L9	NE	NE	DE
RL1554	*rpsR*	putative 30S ribosomal protein S18	ES	ES	DE
RL1555	*rpsF*	putative 30S ribosomal protein S6	ES	ES	DE
RL1672	*rpsI*	putative 30S ribosomal protein S9	DE	DE	ES
RL1731		putative 50S ribosomal protein L33	NE	NE	DE
RL1761	*rplK*	putative 50S ribosomal protein L11	ES	ES	ES
RL1762	*rplA*	putative 50S ribosomal protein L1	ES	ES	ES
RL1764	*rplJ*	putative 50S ribosomal protein L10	ES	ES	ES
RL1765	*rplL*	putative 50S ribosomal protein L7/L12 (L8)	ES	ES	ES
RL1769	*rpsL*	putative 30S ribosomal protein S12	ES	ES	ES
RL1770	*rpsG*	putative 30S ribosomal protein S7	ES	ES	ES
RL1773	*rpsJ*	putative 30S ribosomal protein S10	ES	ES	ES
RL1774	*rplC*	putative 50S ribosomal protein L3	ES	ES	ES
RL1775	*rplD*	putative 50S ribosomal protein L4	ES	ES	ES
RL1776	*rplW*	putative 50S ribosomal protein L23	ES	ES	ES
RL1777	*rplB*	putative 50S ribosomal protein L2	ES	ES	ES
RL1778	*rpsS*	putative 30S ribosomal protein S19	ES	ES	ES
RL1779	*rplV*	putative 50S ribosomal protein L22	ES	ES	ES
RL1780	*rpsC*	putative 30S ribosomal protein S3	ES	ES	ES
RL1781	*rplP*	putative 50S ribosomal protein L16	ES	ES	ES
RL1782	*rpmC*	putative 50S ribosomal protein L29	ES	ES	ES
RL1783	*rpsQ*	putative 30S ribosomal protein S17	ES	NE	ES
RL1784	*rplN*	putative 50S ribosomal protein L14	DE	ES	ES
RL1785	*rplX*	putative 50S ribosomal protein L24	ES	ES	ES
RL1786	*rplE*	putative 50S ribosomal protein L5	ES	ES	ES
RL1787	*rpsN*	putative 30S ribosomal protein S14	ES	ES	ES
RL1788	*rpsH*	putative 30S ribosomal protein S8	ES	ES	ES
RL1789	*rplF*	putative 50S ribosomal protein L6	ES	ES	ES
RL1790	*rplR*	putative 50S ribosomal protein L18	DE	ES	ES
RL1791	*rpsE*	putative 30S ribosomal protein S5	DE	ES	ES
RL1792	*rpmD*	putative 50S ribosomal protein L30	DE	ES	ES
RL1793	*rplO*	putative 50S ribosomal protein L15	DE	ES	ES
RL1796	*rpsM*	putative 30S ribosomal protein S13 (BS14)	ES	ES	ES
RL1797	*rpsK*	putative 30S ribosomal protein S11	ES	ES	ES
RL1799	*rplQ*	putative 50S ribosomal protein L17	ES	ES	ES

Gene ID	Gene Name	Gene Description	FL	R	B
RL2221	*rpsB*	putative 30S ribosomal protein S2	ES	ES	ES
RL2624	*rpsD*	putative 30S ribosomal protein S4	ES	DE	ES
RL3471		putative 50S ribosomal protein L25	DE	ES	NE
RL4017	*rpmE*	putative 50S ribosomal protein L31	NE	NE	NE
RL4130	*rpsU2*	putative 30S ribosomal protein S21	DE	ES	ES
RL4549		putative 30S ribosomal protein S16	DE	ES	ES
RL4552	*rplS*	putative 50S ribosomal protein L19	ES	ES	NE
RL4624	*rpmF*	putative 50S ribosomal protein L32	DE	ES	NE
RL4676	*rplU*	putative 50S ribosomal protein L21	ES	ES	ES
RL4677	*rpmA*	putative 50S ribosomal protein L27	ES	ES	ES

Table SI 4: Genes included in the control seed. All of them are genes that code for ribosomal proteins. The last three columns show whether the genes are essential (ES, in red) or defective (DE, in purple), neutral (NE, in black) in free-living conditions, the rhizosphere and the bacteroid, according to Wheatley *et al.* (2020).

SI C Analysis of other seed list sizes

Fig. SI 1 shows the effect of using different control seed list sizes. From the list of 49 control genes, in the main text, 5 are chosen at random and the other genes are ranked according to the scores from Subsections 3.2.1-3.2.3. The remaining 44 ribosomal genes are used to assess the power of the scores as a classifier; thus, there are exactly 44 true positives in the dataset. Here we assess the robustness of the choice of taking 5 genes, instead of 1, 16, 25, 33, or 48 genes. The more genes are taken, the higher the AUROC tends to be. All three scores, CD, PR and EW, outperform the random (RO) scores, but there is no clear best performing score among the three scores. This observation warrants the use of procedures such as those from Subsections 3.4 and 3.5 for combining these scores.

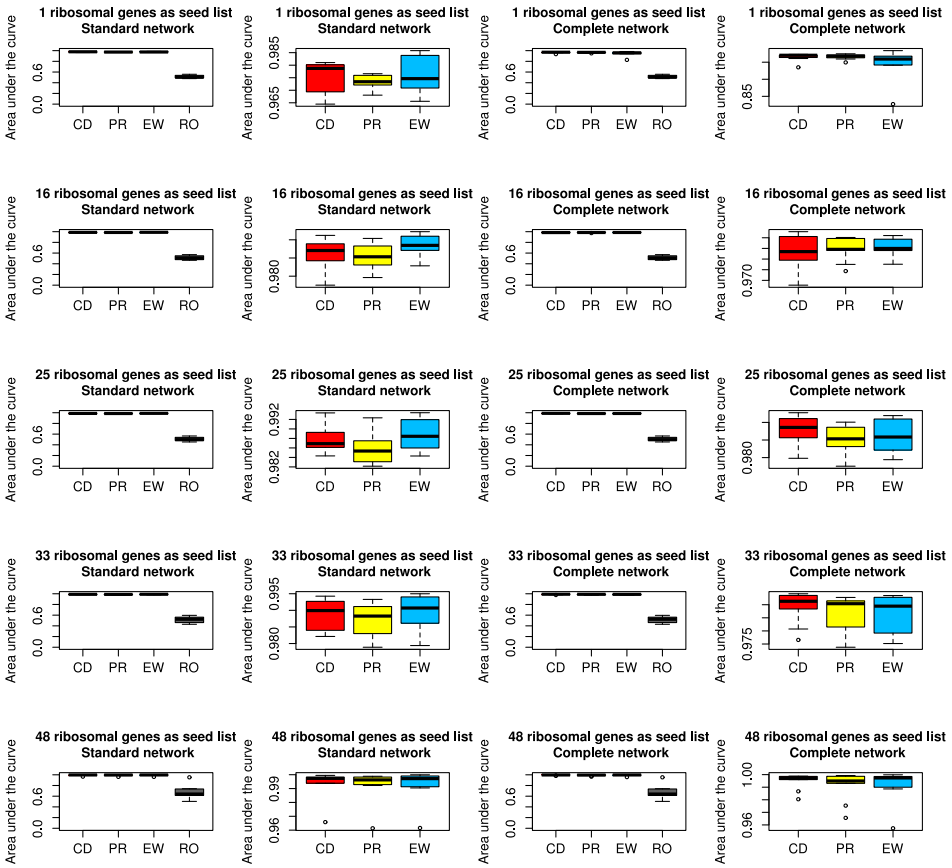


Figure SI 1: Effect of using different control seed list sizes. Which of the scores gives the highest AUROC is not independent of the size of the control seed list. In the main text, 5 control genes were chosen.

SI D Poisson approximation for identifying potentially relevant genes

Suppose we are given K random permutations $\pi^{(1)}, \ldots, \pi^{(K)}$ of $\{1, \ldots, N\}$; for us there would be N genes and K ranking methods. First we pick an $M \leq n$ and select all $i \in \{1, \ldots, N\}$ such that $\pi_i^j \leq M$ for at least one $j = 1, \ldots, K$. From this set we take all i such that $\pi_i^{(j)} \leq T$ for all $j = 1, \ldots, K$, or $\pi_i^j \leq S$ for at least one $j = 1, \ldots, K$. Here, typically, $S \leq M \leq T$.

Picking the top S elements for each ranking gives around KS genes. Here we are interested in number of genes which are selected in addition to those genes, through the condition that $\pi_i^{(j)} \leq T$ for all $j = 1, \ldots, K$. For notation, let $\mathbb{1}$ denote the indicator function such that for an event A, $\mathbb{1}(A) = 1$ if A is the case, and 0 otherwise. Then for $i \in \{1, \ldots, n\}$ set

$$X_i = \mathbb{1}\left(\sum_{j=1}^{K} \mathbb{1}(\pi_i^{(j)} \leq M) \geq 1\right) \mathbb{1}\left(\prod_{j=1}^{K} \mathbb{1}(\pi_i^{(j)} \leq T)\right).$$

Then we let

$$W = \sum_{i=1}^{n} X_i$$

denote the size of the set of such indices. We can calculate its mean;

$$\begin{aligned}
&\mathbb{E}(X_i) \\
&= \mathbb{P}\left(\sum_{j=1}^{K} \mathbb{1}(\pi_i^{(j)} \leq M) \geq 1; \prod_{j=1}^{K} \mathbb{1}(\pi_i^{(j)} \leq T) = 1\right) \\
&= \mathbb{P}\left(\prod_{j=1}^{K} \mathbb{1}(\pi_i^{(j)} \leq T) = 1\right) \mathbb{P}\left(\sum_{j=1}^{K} \mathbb{1}(\pi_i^{(j)} \leq M) \geq 1 \Big| \prod_{j=1}^{K} \mathbb{1}(\pi_i^{(j)} \leq T) = 1\right).
\end{aligned}$$

Now, by the independence of the permutations,

$$\mathbb{P}\left(\prod_{j=1}^{K} \mathbb{1}(\pi_i^{(j)} \leq T) = 1\right) = \left(\frac{T}{N}\right)^K$$

and for the conditional probability, for $T > M$,

$$\begin{aligned}
&\mathbb{P}\left(\sum_{j=1}^{K}\mathbb{1}(\pi_i^{(j)} \le M) \ge 1 \Big| \prod_{j=1}^{K}\mathbb{1}(\pi_i^{(j)} \le T) = 1\right) \\
&= \quad 1 - \mathbb{P}\left(\sum_{j=1}^{K}\mathbb{1}(\pi_i^{(j)} \le M) = 0 | \pi_i^{(j)} \le T, i = 1, \ldots, K\right) \\
&= \quad 1 - \mathbb{P}\left(\prod_{j=1}^{K}\mathbb{1}(\pi_i^{(j)} > M) = 1 \Big| \pi_i^{(j)} \le T, i = 1, \ldots, K\right) \\
&= \quad 1 - \left(\frac{T-M}{T}\right)^K .
\end{aligned}$$

For $T \le M$, this probability is trivially 1. Hence

$$\mathbb{E}(X_i) \quad = \quad \left(\frac{T}{N}\right)^K - \left(\frac{\max(T-M,0)}{N}\right)^K .$$

Summing up these expectations yields

$$\mathbb{E}(W) =: \lambda = \lambda(M,T;K,N) = N\left\{\left(\frac{T}{N}\right)^K - \left(\frac{\max(T-M,0)}{N}\right)^K\right\}.$$

Moreover for $i \neq j$, X_i and X_j are negatively related; if $X_i = 1$ then i's entries take up some of the top ranks, and hence $X_j = 1$ is less likely than if $X_i = 0$. For sums W of negatively related Bernoulli random variables, with $\lambda = \mathbb{E}(W)$, combining Theorem 2.R of Barbour *et al.* (1992), p.43 with Proposition A.2.3 of Barbour *et al.* (1992), p.259, the following Poisson approximation result holds (for more detail about how to obtain this result see also Chapter 11.4 of Waterman (1995)).

Proposition 1. *Let I be a finite index set, let X_α be Bernoulli random variables such that $(X_\alpha, \alpha \in I)$ are negatively related, and let $W = \sum_{\alpha\in I} X_\alpha$ have mean λ. Then, for $m \ge 3\lambda$,*

$$\mathbb{P}(W \ge m) \le 2\frac{m+1}{m+1-\lambda}\frac{1}{\sqrt{2\pi m}}\exp\left\{-\frac{(m-\lambda)^2}{2(m+\lambda)}\right\}.$$

SI E Correlations between scores

Fig. SI 2 shows plots of the different scores when using the ribosomal genes as seed genes against each other. There is some correlation but the plots also show that the scores differ from each other considerably and may score highly for different signals.

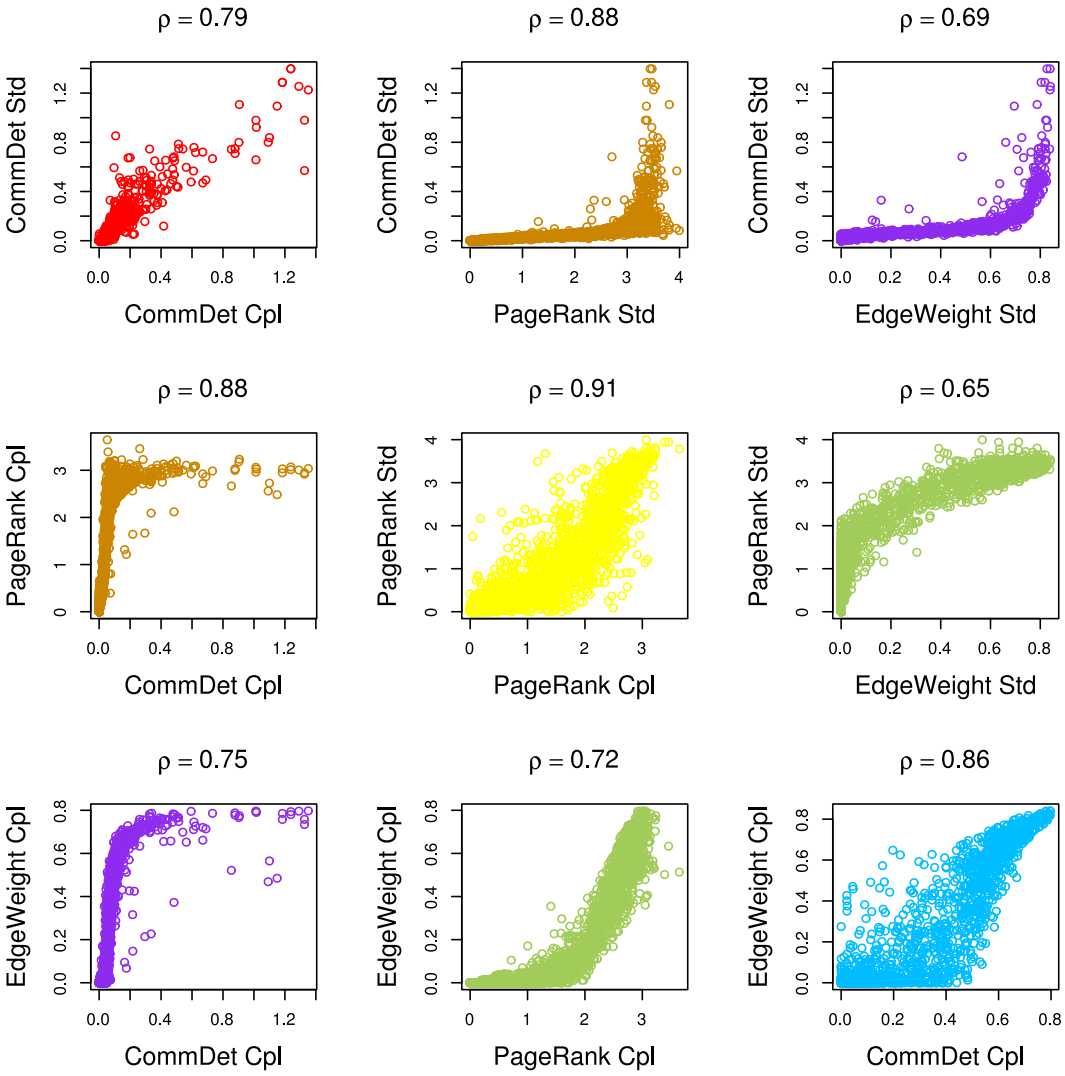


Figure SI 2: Correlation of different scores, in the standard dataset as well as in the complete dataset. The Spearman's rank correlation coefficients are indicated above the scatterplots.

SI F Selected genes - Ribosomal genes

Table SI 5 presents the *R. leguminosarum* genes selected when using the ribosomal genes as seed list. We include the genes among the top $M = 50$ for at least one of the three scores considered and in the top $T = 100$ for all three of them. We also include the genes in the top $S = 10$ for at least one score.

Std	Cpl	Gene ID	Gene Name	Gene Description	FL	R	B
	x	pRL100260		putative alfa-L-rhamnosidase	NE	NE	NE
	x	pRL120235		putative fatty aldehyde dehydrogenase	NE	NE	NE
	x	pRL120521	*etfB2*	putative electron transfer flavoprotein beta-subunit	NE	NE	NE
	x	pRL120522	*etfA2*	putative electron transfer flavoprotein alpha-subunit	NE	NE	NE
x	x	RL0883	*groLch1*	putative 60 kda chaperonin (protein cpn60)	ES	GD	GD
	x	RL0926	*atpE*	putative ATP synthase subunit C	ES	GD	ES
	x	RL0927	*atpG2*	putative ATP synthase B' chain (subunit II)	ES	GD	ES
x	x	RL1022	*ctaD1*	putative cytochrome c oxidase polypeptide I (cytochrome aa3 subunit 1)	NE	GD	ES
x	x	RL1023	*ctaB*	putative protoheme IX farnesyltransferase (heme o synthase)	NE	GD	ES
x	x	RL1260	*cysN*	putative sulfate adenylyltransferase subunit 1	NE	NE	NE
x		RL1261	*cysD*	putative sulfate adenylyltransferase subunit 2	NE	NE	NE
x		RL1499	*ropA2*	putative outer membrane porin protein RopA	NE	GD	ES
x	x	RL1551	*dnaC*	putative DNA helicase	ES	GD	ES
x	x	RL1559	*acpP*	putative acyl carrier protein (fatty acid synthase acyl carrier protein)	GD	GD	ES
x	x	RL1691	*hupA1*	putative DNA-binding protein HU	GD	GD	NE
	x	RL1709	*nuoI*	putative NADH-quinone oxidoreductase subunit I	ES	GD	ES
	x	RL1710	*nuoJ*	putative NADH-quinone oxidoreductase subunit J	ES	GD	ES
x	x	RL1719	*proS*	putative prolyl-tRNA synthetase	ES	GD	ES
x	x	RL1757	*tufB1*	putative elongation factor Tu (EF-Tu)	ES	GD	ES
x	x	RL1763		hypothetical protein	ES	GD	ES

Std	Cpl	Gene ID	Gene Name	Gene Description	FL	R	B
x	x	RL1766	*rpoB*	putative DNA-directed RNA polymerase beta chain	ES	GD	ES
x	x	RL1767	*rpoC*	putative DNA-directed RNA polymerase beta' chain	ES	GD	ES
x	x	RL1772	*tufB2*	putative elongation factor Tu (EF-Tu)	ES	GD	ES
	x	RL1798	*rpoA*	putative DNA-directed RNA polymerase alpha chain	ES	GD	ES
x	x	RL1948	*gpt*	putative xanthine-guanine phosphoribosyltransferase	NE	NE	NE
x		RL2062	*tig*	putative chaperone trigger factor	NE	GD	NE
x	x	RL2222	*tsf*	putative elongation factor Ts (EF-Ts)	GD	GD	ES
x		RL2228		putative outer membrane protein	ES	GD	ES
x	x	RL2239	*eno*	putative enolase	NE	GD	ES
x		RL2291		conserved hypothetical protein	NE	NE	ES
	x	RL2526		putative oxidoreductase	NE	NE	NE
x		RL2573	*gcvP*	putative glycine dehydrogenase [decarboxylating]	NE	NE	NE
x		RL2606	*purC1*	putative phosphoribosylaminoimidazole-succinocarboxamide synthase (saicar synthetase)	NE	NE	ES
	x	RL2686	*aroG*	putative phospho-2-dehydro-3-deoxyheptonate aldolase	NE	NE	NE
x		RL2697		conserved hypothetical exported protein	ES	NE	NE
	x	RL2829	*cobO*	putative cob(I)yrinic acid a,c-diamide adenosyltransferase	NE	ES	ES
	x	RL2964	*cspA3*	putative cold shock protein CspA	GD	GD	NE
x	x	RL3333		putative ATP-binding:ATP-binding (ABC:ABC) component of ABC transporter Unclass	NE	GD	ES
x	x	RL3540	*braC3*	putative SBP of ABC transporter HAAT orphan braC3	NE	NE	NE
x	x	RL3617	*araA*	putative SBP of ABC transporter CUT2 (elevated on MA L-arabinose and galactose Rlv3841)	NE	NE	NE
x		RL4007	*gap*	putative glyceraldehyde-3-phosphate dehydrogenase	NE	GD	ES
x		RL4065		conserved hypothetical protein	NE	GD	NE
x		RL4261	*nrdH*	putative glutaredoxin-like protein	NE	NE	NE
x		RL4295		putative acetyltransferase	NE	NE	ES

Std	Cpl	Gene ID	Gene Name	Gene Description	FL	R	B
x	x	RL4407	*atpD*	putative ATP synthase beta chain	ES	ES	ES
	x	RL4408	*atpG*	putative ATP synthase gamma chain	ES	ES	ES
	x	RL4409	*atpA*	putative ATP synthase alpha chain	ES	ES	ES
x	x	RL4617	*hslR*	putative heat shock protein 15	NE	GD	NE
x	x	RL4634		putative SBP of ABC transporter PAAT	NE	NE	NE
	x	RL4635		putative permease component of ABC transporter PAAT	NE	NE	NE
x	x	RL4715	*asd*	putative aspartate-semialdehyde dehydrogenase	GD	ES	ES

Table SI 5: Genes selected when using the ribosomal genes as seed list. The first two columns indicate whether the genes are selected for the standard dataset (Std) and the complete dataset (Cpl). The last three columns show whether the genes are essential (ES, in red), defective (DE, in purple) or neutral (NE, in black) in free-living conditions, the rhizosphere and the bacteroid, according to Wheatley *et al.* (2020).

SI G Selected genes - Comparing conditions

Table SI 5 presents the *R. leguminosarum* genes selected when studying different growth conditions. We include the genes among the top $S = 5$ for at least one of the three scores considered or in the top $T = 20$ for all three of them.

Std	Cpl	Gene ID	Gene Name	Gene Description	FL	R	B
x		pRL100163		hypothetical protein with homology to asparagine synthase (glutamine hyrdrolysing) at C-term	NE	GD	NE
x	x	pRL100164	*rhiI*	putative enzyme to synthesise (autoinducer) AHL, RhiI	NE	NE	NE
x	x	pRL100169	*rhiA*	rhizosphere-induced protein RhiA	NE	NE	NE
x	x	pRL100170	*rhiB*	rhizosphere-induced protein RhiB	NE	NE	NE
x		pRL100171	*rhiC*	rhizosphere-induced protein RhiC	NE	NE	NE
x		pRL100359		putative permease component of ABC transporter CUT2	NE	NE	NE
	x	pRL100385	*ecfM*	putative RNA polymerase ECF sigma factor, family ECF20/ECF01	NE	GD	NE
x	x	pRL100390		putative NAD-binding D-isomer specific 2-hydroxyacid dehydrogenase	NE	NE	NE
x		pRL100401	*ntaA1*	putative nitrilotriacetate monooxygenase component A	NE	NE	NE
x	x	pRL100422		conserved hypothetical protein	NE	NE	NE
x	x	pRL110005		putative transmembrane YccA-like protein (role in maintaining membrane integrity) BI-1 (BAX-Inhibitor) with anti-apoptotic effect	NE	NE	NE
x	x	pRL110006	*asfN*	putative transmembrane anti-sigma factor (COG5662), AsfN	NE	NE	NE
	x	pRL110007	*ecfN*	putative RNA polymerase ECF sigma factor, EcfN, family ECF26	NE	NE	NE
x	x	pRL110086	*pcaG*	putative protocatechuate 3,4-dioxygenase alpha chain	NE	NE	NE
	x	pRL110087	*pcaH1*	putative protocatechuate 3,4-dioxygenase beta chain	NE	NE	NE
	x	pRL110088	*pcaC*	putative carboxymuconolactone decarboxylase	NE	NE	NE
x	x	pRL110089		putative beta-ketoadipate enol-lactone hydrolase	NE	NE	NE
x		pRL110160		putative LysR family transcriptional regulator	NE	NE	NE

Std	Cpl	Gene ID	Gene Name	Gene Description	FL	R	B
x		pRL110167		putative permease component of ABC transporter CUT2	NE	NE	NE
x	x	pRL110242		putative glycosiside hydrolase fam4	NE	NE	NE
x	x	pRL110247		putative alpha-galactosidase	NE	NE	NE
x	x	pRL110287	*pcaI*	putative 3-oxoadipate CoA-transferase subunit A	NE	NE	NE
x	x	pRL110288	*pcaJ*	putative 3-oxoadipate CoA-transferase subunit B	NE	NE	NE
	x	pRL110289	*pcaF*	putative beta-ketoadipyl CoA thiolase	NE	NE	NE
x	x	pRL110304	*treX*	putative glycosylhydrolase	NE	NE	NE
	x	pRL110374	*sulA*	putative SBP of ABC transporter SulT (S. mel SBP homologue SMb21133 induced by sulphate limitation)	NE	NE	NE
x		pRL110401		putative ATP-binding component of ABC transporter HAAT (S. mel SBP homologue SMc03121 induced by methionine, hydroxybenzoate and elevated by acetoacetate in MA on 3841)	NE	NE	NE
x		pRL110404		putative permease component of ABC transporter HAAT (S. mel SBP homologue SMc03121 induced by methionine, hydroxybenzoate and elevated by acetoacetate in MA on 3841)	NE	AD	NE
x	x	pRL110484		putative dienelactone hydrolase family protein	NE	NE	NE
	x	pRL110568		putative transmembrane protein	NE	NE	NE
x		pRL120052	*tdcB2*	putative threonine dehydratase catabolic	NE	NE	NE
x		pRL120058		hypothetical protein	NE	NE	NE
	x	pRL120171		putative ATP-binding component of ABC transporter POPT	NE	NE	NE
x		pRL120191		putative SBP of ABC transporter CUT2	NE	NE	NE
	x	pRL120211		hypothetical protein	NA	NA	NA
x		pRL120260		conserved hypothetical protein	NE	NE	NE
x	x	pRL120342	*hspD*	putative small heat shock protein	NE	NE	NE
x	x	pRL120416	*dadX*	putative alanine racemase	NE	NE	NE
x	x	pRL120417	*dadA*	D-amino acid dehydrogenase	NE	NE	NE
x		pRL120418		putative AraC family transcriptional regulator (activator)	NE	NE	NE

Std	Cpl	Gene ID	Gene Name	Gene Description	FL	R	B
x		pRL120641		conserved hypothetical protein	NE	NE	NE
x	x	pRL120694		putative LacI family transcriptional regulator (repressor)	NE	NE	GD
	x	pRL120750	*qatX5*	putative SBP of ABC transporter QAT	NE	NE	NE
x		pRL120757		putative ATP-binding:ATP-binding (ABC:ABC) component of ABC transporter CUT2	NE	AD	NE
x		pRL120768		conserved hypothetical protein	NE	NE	NE
x		pRL120778		putative dehydratase/epimerase	NE	NE	NE
x		pRL120780		putative SBP of ABC transporter PepT	NE	NE	NE
x		pRL120781		putative permease component of ABC transporter PepT	NE	NE	NE
x		pRL120789	*nolR2*	putative nodulation protein NolR	NE	NE	NE
x	x	pRL120792		putative two-component sensor/regulator; NodW-like nodulation transcriptional regulator	NE	NE	NE
x	x	pRL120793		putative repetitive two-component sensor/regulator; histidine kinase	NE	NE	NE
x		pRL70002	*repBp7a*	putative replication protein RepB	NE	NE	NE
x		pRL70160		conserved hypothetical protein	NE	NE	NE
	x	pRL90219		putative TenA family transcriptional regulator	NE	NE	NE
x		pRL90241		hypothetical protein	NE	NE	NE
x		pRL90276		putative ATP-binding :permease (ABC:IMP) component of ABC transporter Export cytochrome bd-related	NE	NE	NE
x	x	RL0037	*pckA*	putative phosphoenolpyruvate carboxykinase	NE	NE	NE
x	x	RL0132		conserved hypothetical protein	NE	NE	NE
x		RL0158		putative GGDEF/EAL domain transcriptional regulator	NE	NE	NE
x	x	RL0324		conserved hypothetical protein	NE	NE	NE
x	x	RL0381	*hrcA*	putative heat-inducible transcription repressor	NE	NE	GD
x		RL0394	*phoH*	putative phosphate starvation-induced protein	GD	GD	GD
	x	RL0441		putative vitamin B12 binding methyltransferase	NE	NE	NE

Std	Cpl	Gene ID	Gene Name	Gene Description	FL	R	B
x		RL0518		putative SBP of ABC transporter CUT2 (S. mel SBP homologue SMc02774 induced by D-fucose, pyruvic acid)	NE	NE	NE
x		RL0522		conserved hypothetical protein	NE	NE	NE
x	x	RL0760		conserved hypothetical protein	NA	NA	NA
x	x	RL0761	*aceA*	putative isocitrate lyase	NE	NE	NE
x	x	RL0776		malonic semialdehyde oxidative decarboxylase	NE	NE	GD
	x	RL0852		conserved hypothetical protein	NE	NE	NE
x		RL0871	*tag*	putative DNA-3-methyladenine glycosylase I	NE	NE	NE
	x	RL0914		conserved hypothetical protein	NE	NE	NE
x	x	RL0915	*dgoA*	putative 2-dehydro-3-deoxy-6-phosphogalactonate aldolase	NE	NE	NE
	x	RL0949	*mcpJ*	putative methyl-accepting chemotaxis protein McpJ	NE	NE	NE
x		RL1074		conserved hypothetical protein	NE	NE	NE
x	x	RL1089		conserved hypothetical protein	NE	NE	NE
	x	RL1101		hypothetical protein	NE	NE	NE
x	x	RL1121		hypothetical protein	NE	NE	NE
x		RL1122		putative LuxR/GerE family transcriptional regulator	NE	NE	NE
x		RL1123		conserved hypothetical protein	NE	NE	NE
x		RL1133		putative ArsR family transcriptional regulator	NE	NE	NE
x		RL1138	*ecfC*	putative RNA polymerase ECF sigma factor, family ECF26	NE	NE	NE
	x	RL1299		putative cold shock protein	NE	NE	NE
x		RL1328	*salR*	putative LysR family transcriptional regulator of MFS family efflux transporter (+ve regulator)	NE	NE	NE
	x	RL1486		putative LysR family transcriptional regulator	NE	NE	NE
x	x	RL1491		conserved hypothetical protein	NE	NE	NE
	x	RL1492	*iolB*	putative myo-inositol catabolism protein B	NE	NE	NE
	x	RL1493	*iolE*	putative rhizopine catabolism protein	NE	NE	NE
x	x	RL1494	*iolD*	putative malonate semialdehyde dehydrogenase	NE	NE	NE
x	x	RL1495	*iolC*	putative myo-inositol catabolism protein C	NE	NE	NE

Std	Cpl	Gene ID	Gene Name	Gene Description	FL	R	B
	x	RL1496	*iolR*	putative SIS (Sugar ISomerase)-RpiR family transcriptional regulator	NE	NE	NE
x	x	RL1497		putative dehydrogenase/oxidoreductase	NE	NE	NE
x	x	RL1498		putative flavoprotein	GD	ES	GD
x		RL1525		putative aminomethyltransferase	GD	GD	GD
	x	RL1527		conserved hypothetical protein	GD	GD	GD
x		RL1895		hypothetical protein	NE	NE	ES
x	x	RL1912		conserved hypothetical protein with homology to BatD (but lacks membrane spanning domains) BatD is involved in tolerance to oxygen in Bacteroides species	NE	NE	NE
	x	RL1966	*aldA*	putative alanine dehydrogenase	NE	GD	NE
x	x	RL2411		putative transmembrane protein	NE	NE	NE
x		RL2588	*tyrS1*	putative tyrosyl-tRNA synthetase	NE	NE	ES
x		RL2637	*recA*	putative recombinase	NE	GD	ES
x		RL2869		hypothetical exported protein	NE	NE	NE
x		RL2879		putative AraC family transcriptional regulator (activator)	NE	NE	NE
x		RL3037		hypothetical exported protein	NA	NA	NA
	x	RL3172		hypothetical protein	NE	NE	NE
	x	RL3297	*lpxC*	putative UDP-3-O-[3-hydroxymyristoyl] N-acetylglucosamine deacetylase	ES	GD	ES
x	x	RL3326	*btaA*	putative S-adenosylmethionine:diacylglycerol 3-amino-3-carboxypropyl transferase	NE	NE	NE
	x	RL3374		conserved hypothetical protein	NE	NE	NE
x	x	RL3424	*dctA*	putative C4-dicarboxylate transport protein	NE	NE	GD
	x	RL3613	*araE*	putative arabinose dehydrogenase (elevated on MA L-arabinose and galactose Rlv3841)	NE	GD	NE
x	x	RL3616	*araB*	putative ATP-binding:ATP-binding (ABC:ABC) component of ABC transporter protein CUT2 (elevated on MA L-arabinose and galactose Rlv3841)	NE	NE	NE
x	x	RL3686		putative transporter component protein	NE	NE	NE

Std	Cpl	Gene ID	Gene Name	Gene Description	FL	R	B
x	x	RL3705	*phyR*	putative anti-anti-sigma factor; two-component sensor/regulator; transcriptional regulator, PhyR, ECF15 gene organisation (histidine kinase, response regulator, asf, ecf)	NE	GD	NE
x		RL3706		putative two-component sensor/regulator; histidine kinase, ECF15 gene organisation (histidine kinase, response regulator, asf, ecf)	ES	GD	ES
	x	RL3906		putative SBP of ABC transporter HAAT (S. mel SBP homologue SMb20568 induced by protocatechuate, hydroxybenzoate and elevated by hydroxybenzoate in MA on 3841)	NE	NE	NE
	x	RL3907		putative permease component of ABC transporter HAAT (S. mel SBP homologue SMb20568 induced by protocatechuate, hydroxybenzoate and elevated by hydroxybenzoate in MA on 3841)	NE	NE	NE
	x	RL3908		putative permease component of ABC transporter HAAT (S. mel SBP homologue SMb20568 induced by protocatechuate, hydroxybenzoate and elevated by hydroxybenzoate in MA on 3841)	NE	NE	NE
	x	RL4012		putative fructose-bisphosphate aldolase	NE	NE	NE
	x	RL4062		putative amidohydrolase	NE	NE	ES
x		RL4195		putative transmembrane protein	NE	NE	NE
x		RL4266	*acoR*	putative IclR family transcriptional regulator, putative AcoR activator of acetoin/glycerol metabolism	NE	NE	NE
x		RL4446	*sdhC*	putative succinate dehydrogenase cytochrome b556 subunit	GD	GD	ES
x		RL4509		putative transmembrane efflux protein	NE	NE	NE
x	x	RL4614	*rpoH2*	putative RNA polymerase sigma factor (heat shock)	NE	NE	NE

Std	**Cpl**	**Gene ID**	**Gene Name**	**Gene Description**	**FL**	**R**	**B**
x	x	RL4653	*intC*	putative permease component of ABC transporter CUT2 inositol transporter (S. mel SBP homologue SMb20712 induced by inositol)	NE	NE	NE
x	x	RL4654	*intB*	putative ATP-binding:ATP-binding (ABC:ABC) component of ABC transporter CUT2 inositol transporter (S. mel SBP homologue SMb20712 induced by inositol)	NE	NE	NE
x	x	RL4655	*intA*	putative SBP of ABC transporter CUT2 inositol transporter (S. mel SBP homologue SMb20712 induced by inositol)	NE	NE	NE
x	x	RL4709		putative component of MFS transporter, homology with shikimate transporter	NE	NE	NE

Table SI 6: Genes selected when studying different growth conditions. The first two columns indicate whether the genes are selected for the standard dataset (Std) and the complete dataset (Cpl). The last three columns show whether the genes are essential (ES, in red), defective (DE, in purple), advantageous (AD, in green), or neutral (NE, in black) in free-living conditions, the rhizosphere and the bacteroid, according to Wheatley *et al.* (2020).

E

Supplementary Tables

Table E.1: Genes upregulated in the bacteroid

Gene Information				RNA-Seq			MA	INSeq				
Gene ID	**Old ID**	**Dsg**	**Description**	**B**	**FL**	**FC**	**FC**	**FL**	**Rh**	**R**	**NB**	**Bd**
RL_RS39930			hypothetical protein	4786.32	0.42	37345.67						
RL_RS28380	pRL100200	fixA	electron transfer protein FixA	4411.44	0.53	24293.98	132.73	NE	NE	NE	DE	DE
RL_RS28185	pRL100162	nifH	nitrogenase iron protein NifH	5009.22	2.36	6625.47	148.68	NE	NE	NE	NE	DE
RL_RS28375	pRL100199	fixB	electron transfer protein FixB	4382.61	2.27	6147.6	43.44	NE	NE	NE	DE	DE
RL_RS28180	pRL100161	nifD	nitrogenase molybdenum-iron protein alpha chain NifD	4279.51	3.08	4442.61	129.9	NE	NE	NE	ES	DE
RL_RS27915	pRL100103		putative hydroxyvalerate/ alcohol dehydrogenase	3717.03	4.79	2446.88	15.85	ES	NE	NE	NE	NE
RL_RS28175	pRL100160	nifK	nitrogenase molybdenum-iron protein beta chain NifK	5037.38	6.9	2313.25	131.58	NE	NE	NE	ES	DE
RL_RS39925	pRL100100		*pseudogene*	4159.04	6.18	2145.87	64.84	NE	NE	NE	NE	NE
RL_RS27920	pRL100104	phaE	putative phaE protein, subunit of polyhydroxyalkanoate (PHA) synthase	3050.09	4.5	2145.4	38.75	NE	NE	NE	NE	NE

Continued on next page

Table E.1 – continued from previous page

Gene Information				RNA-Seq			MA	INSeq				
Gene ID	**Old ID**	**Dsg**	**Description**	**B**	**FL**	**FC**	**FC**	**FL**	**Rh**	**R**	**NB**	**Bd**
RL_RS27975	pRL100119		putative propionate CoA-transferase	1282.71	2.15	1879.49	41.34	NE	NE	NE	NE	NE
RL_RS28420			hypothetical protein*	4627.4	10.7	1338.46						
RL_RS28170	pRL100159	nifE	putative nitrogenase iron-molybdenum cofactor biosynthesis protein NifE	1877.73	4.69	1281.84	50.74	NE	NE	NE	NE	DE
RL_RS28165	pRL100158	nifN	putative nitrogenase iron-molybdenum cofactor biosynthesis protein NifN	1396.1	3.88	1147.78	38.61	NE	NE	NE	DE	DE
RL_RS27960	pRL100114		hypothetical protein	60.01	0.17	1118.44	2.36	ES	NE	NE	NE	NE
RL_RS27925	pRL100105	phaC2	putative polyhydroxyalkanoate (PHA) synthase subunit C	2220.59	7.43	950.55	16.37	NE	NE	NE	NE	NE
RL_RS28385	pRL100201		conserved hypothetical protein	1609.32	5.76	879.99	121.81	NE	AD	AD	NE	NE
RL_RS28370	pRL100198	fixC	nitrogen fixation protein FixC	2113.01	7.77	859.19	71.52	NE	NE	NE	NE	DE
RL_RS28415	pRL100207	fixP1	putative cytochrome oxidase subunit	3042.75	11.48	835.06	85.28	NE	NE	NE	AD	NE

Continued on next page

Table E.1 – continued from previous page

Gene Information				RNA-Seq			MA	INSeq				
Gene ID	**Old ID**	**Dsg**	**Description**	**B**	**FL**	**FC**	**FC**	**FL**	**Rh**	**R**	**NB**	**Bd**
RL_RS27980	pRL100120		phasin-like protein, often associated with PHB or other polyhydroxyalkanoate (PHA)	982.25	4.7	659.74	15.53	NE	NE	NE	NE	NE
RL_RS27910	pRL100101		*pseudogene*, acetolactate synthase	8137.18	39.06	659	105.08	NE	NE	NE	NE	NE
RL_RS28410	pRL100206A	fixQ1	putative component of cytochrome oxidase	2417.11	11.55	648.57	71.34	NE	NE	NE	AD	NE
RL_RS38405	pRL100156	fdxB1	putative ferredoxin	419.05	2.19	627.43	9.46	NE	DE	NE	NE	NE
RL_RS28365	pRL100197	fixX	ferredoxin-like protein FixX	1608.23	9.55	531.88	29.44	NE	NE	NE	NE	DE
RL_RS28405	pRL100206	fixO1	putative cytochrome oxidase subunit	2886.91	17.18	529.74	69.33	NE	NE	NE	AD	NE
RL_RS28335	pRL100192		hypothetical protein	493.85	3.01	518.11	25.52	NE	NE	NE	NE	NE
RL_RS28350			putative nitrogen fixation protein NifT	1647.94	10.4	499.63						
RL_RS28400	pRL100205	fixN1	putative transmembrane cytochrome oxidase subunit	2741.86	17.44	497.43	119.68	NE	NE	NE	DE	DE
RL_RS27985	pRL100121	acsA3	putative acetyl-coA synthetase	644.37	4.12	496.23	13.33	NE	NE	NE	NE	NE
RL_RS38450			4Fe-4S binding protein	1379.44	8.95	485.84						

Continued on next page

Table E.1 – continued from previous page

Gene Information				RNA-Seq			MA	INSeq				
Gene ID	**Old ID**	**Dsg**	**Description**	**B**	**FL**	**FC**	**FC**	**FL**	**Rh**	**R**	**NB**	**Bd**
RL_RS28360	pRL100196	nifA	transcriptional regulator NifA	1627.89	12.7	406.98	11.88	NE	NE	NE	NE	DE
RL_RS28340	pRL100193		conserved hypothetical protein	800.31	7.53	336.88	39.16	NE	NE	NE	NE	NE
RL_RS27900	pRL100099		conserved hypothetical protein	6331.65	59.61	332.56	35.98	NE	NE	NE	NE	NE
RL_RS27930	pRL100107		hypothetical protein	680.61	7.57	287.91	7.9	NE	NE	NE	NE	NE
RL_RS28390	pRL100202		conserved hypothetical protein	123.86	1.62	240.67	9.16	NE	NE	AD	NE	NE
RL_RS28355	pRL100195	nifB	FeMo cofactor biosynthesis protein NifB	1155.71	16.3	225.62	35	NE	NE	NE	NE	DE
RL_RS11745	RL2270		putative nodulin-related transmembrane protein	1318.04	20.1	207.73	2.77	NE	NE	NE	NE	NE
RL_RS27990	pRL100122	adk2	putative adenylate kinase	268.77	4.05	205.03	3.06	NE	NE	NE	NE	NE
RL_RS28150	pRL100154		hypothetical protein	153.78	2.86	172.14	13.05	NE	NE	NE	NE	NE
RL_RS05975	RL1145		putative conjugated bile salt hydrolase	395.91	7.53	167.06	34.66	NE	NE	NE	NE	NE
RL_RS09770	RL1876	adh	putative alcohol dehydrogenase	296.95	5.85	162.83	5.9	NE	AD	NE	NE	NE
RL_RS27895	pRL100098		conserved hypothetical protein	2246.85	44.55	158.65	14.59	NE	NE	NE	NE	NE

Continued on next page

Table E.1 – continued from previous page

Gene Information				RNA-Seq			MA	INSeq				
Gene ID	**Old ID**	**Dsg**	**Description**	**B**	**FL**	**FC**	**FC**	**FL**	**Rh**	**R**	**NB**	**Bd**
RL_RS27935	pRL100108		hypothetical protein	55.04	1.19	145.63	1.23	NE	NE	NE	NE	NE
RL_RS27890	pRL100097		conserved hypothetical protein	1334.45	29.06	145.12	6.38	NE	NE	NE	NE	NE
RL_RS27965	pRL100115		*pseudogene*, acetolactate synthase	129.87	2.85	144.27	2.77	NE	NE	NE	NE	NE
RL_RS25965	pRL90016A	fixQ2	putative component of cytochrome oxidase	701.21	16.02	137.78	53.34	NE	NE	NE	NE	NE
RL_RS25960	pRL90016	fixP2	putative cytochrome oxidase component	866.54	20.22	135.45	63.95	NE	NE	NE	NE	NE
RL_RS27970	pRL100117		cytochrome c protein *pseudogene*, iron-containing alcohol dehydrogenase	129.31	3.06	133.53	6.73	ES	NA	NA	NA	NA
RL_RS25970	pRL90017	fixO2	putative transmembrane cytochrome c oxidase fixO protein	841.4	22.34	119.55	54.75	NE	NE	NE	NE	NE
RL_RS28095	pRL100143		*pseudogene*, acetolactate synthase	136.51	4.23	99.45	1.75	NE	NE	NE	NE	NE
RL_RS05120	RL0984		putative ATP-binding component of ABC transporter PAAT	23.28	0.69	98.57	2.55	AD	NE	NE	NE	NE

List of the 50 genes most highly upregulated in the bacteroid, compared to the free-living conditions and according to the

RNA-Seq dataset. “Dsg” indicates the gene name. The columns under the “RNA-Seq” title indicate the expression values (TPM), according to the RNA-Seq dataset, of the gene in the bacteroid (B), in free-living conditions (FL), and the Fold Change (FC) value as their quotient. The column under the “MA” title indicate the FC values for the genes accoording to the microarray dataset. The last five columns indicate the INSeq value for the genes in free-living conditions (FL), the rhizosphere (Rh), the root (R), nodule bacteria (NB), and bacteroid (Bd), according to (Wheatley *et al.*, 2020); possible values for these columns are essential (ES), deffective (DE), neutral (NE), and advantaged (AD).

Table E.2: Genes coexpressed with the studied transcriptional regulators

Gene Information			Scores			INSeq				
Gene ID	**Dsg**	**Description**	**CD**	**PR**	**EW**	**FL**	**Rh**	**R**	**NB**	**Bd**
Genes coexpressed with pRL80079										
pRL100462		putative permease component of ABC transporter CUT2	1617	6	829	NE	NE	NE	NE	NE
pRL120221		conserved hypothetical protein	1392	9	773	NE	NE	NE	NE	NE
pRL120270		putative exonuclease	7	229	9	NE	NE	NE	NE	NE
RL0417	ihfB	putative integration host factor beta subunit	9	167	14	ES	NE	NE	NE	NE
RL0493	pyrC	putative dihydroorotase	126	310	3	NE	NE	NE	ES	DE
RL0494		putative transmembrane GGDEF sensory box protein	5	228	43	NE	NE	NE	NE	NE
RL0677		putative LuxR/GerE family transcriptional regulator	8	170	10	NE	NE	NE	NE	NE
RL0801		putative transmembrane glycosyl transferase	1097	10	204	NE	NE	NE	NE	NE
RL1319		putative DNA topology modulation protein	1514	8	813	NE	NE	NE	NE	NE
RL1817		putative lipase/esterase	213	388	7	NE	NE	NE	NE	NE
RL1935		hypothetical protein with pentapeptide motif	6	113	101	NE	NE	NE	NE	NE
RL2210		conserved hypothetical protein	1760	4	686	NE	DE	DE	NE	NE
RL2395		putative transmembrane protein	241	220	2	NE	NE	NE	NE	NE

Continued on next page

Table E.2 – continued from previous page

Gene Information			Scores			INSeq				
Gene ID	**Dsg**	**Description**	**CD**	**PR**	**EW**	**FL**	**Rh**	**R**	**NB**	**Bd**
RL2880		putative two-component sensor/regulator; transcriptional regulator with GGDEF sensory box domain, diguanylate cyclase	2	177	1	NE	NE	NE	NE	NE
RL2929		hypothetical protein	10	210	27	NE	NE	NE	NE	NE
RL3053		conserved hypothetical protein	1550	5	710	NE	NE	NE	AD	NE
RL3481		putative MaoC family protein	19	27	20	NE	NE	NE	NE	NE
RL3482		putative MaoC family protein	1520	7	707	NE	NE	NE	NE	NE
RL3837		conserved hypothetical protein	1	145	13	AD	NE	NE	NE	NE
RL3891		conserved hypothetical exported protein	1351	3	224	NE	NE	NE	NE	NE
RL3962		putative heat resistant agglutinin 1	316	248	6	NE	NE	NE	NE	NE
RL4050	ggt1	putative gamma-glutamyltranspeptidase precursor	1815	2	656	NE	NE	NE	NE	NE
RL4115			329	298	8					
RL4294		putative NTP pyrophosphohydrolase	74	233	5	NE	NE	NE	NE	NE
RL4604		putative GntR family transcriptional regulator	1697	1	754	NE	NE	NE	NE	NE
RL4627	mtgA	putative monofunctional biosynthetic peptidoglycan transglycosylase	3	115	4	NE	NE	NE	NE	NE
RL4651		putative solute binding-protein of ABC transporter PepT	4	45	11	AD	NE	NE	NE	NE
Genes coexpressed with pRL90058										
pRL110257A		conserved hypothetical protein	18	99	68	NE	NE	NE	NE	NE

Continued on next page

Table E.2 – continued from previous page

Gene Information			Scores			INSeq				
Gene ID	**Dsg**	**Description**	**CD**	**PR**	**EW**	**FL**	**Rh**	**R**	**NB**	**Bd**
pRL110321		conserved hypothetical protein	3	212	1	NE	NE	NE	NE	NE
pRL110424		hypothetical protein	9	48	32	NE	NE	NE	NE	NE
pRL110425		conserved hypothetical protein	2527	4	474	NE	NE	NE	NE	NE
pRL120335		putative acylase	1662	5	415	NE	NE	NE	NE	NE
pRL120433		putative permease component of ABC transporter PepT	478	1	267	NE	NE	NE	NE	NE
pRL120717		conserved hypothetical protein	34	94	67	NE	NE	NE	NE	NE
pRL120756		putative solute binding-protein of ABC transporter CUT2	487	6	228	NE	NE	NE	NE	NE
pRL90198		putative curved DNA-binding protein	563	2	385	NE	NE	NE	NE	NE
pRL90212	msrA	putative peptide methionine sulfoxide reductase MsrA	6	60	34	NE	NE	NE	NE	NE
RL0190		putative transmembrane protein	10	218	82	NE	NE	NE	NE	NE
RL0667		conserved hypothetical protein	47	49	61	NE	NE	NE	NE	NE
RL0913		putative PRC family protein	30	138	7	NE	NE	NE	AD	NE
RL1306		putative glyoxalase	515	8	487	NE	NE	NE	NE	NE
RL1365		putative glyoxalase/bleomycin-resistance protein/dioxygenase	2	122	10	NE	NE	NE	NE	NE
RL1366	uspF	putative universal stress protein	54	89	33	NE	NE	NE	NE	NE
RL1485		conserved hypothetical protein	22	175	8	NE	NE	NE	NE	NE
RL1888		hypothetical exported protein	8	194	15	NE	AD	NE	NE	NE
RL2391		conserved hypothetical protein	5	177	35	NE	NE	NE	NE	NE
RL2969		putative transmembrane protein	36	81	30	NE	NE	NE	NE	NE

Continued on next page

Table E.2 – continued from previous page

Gene Information			Scores			INSeq				
Gene ID	**Dsg**	**Description**	**CD**	**PR**	**EW**	**FL**	**Rh**	**R**	**NB**	**Bd**
RL3178	znuA	putative solute binding-protein of ABC transporter MZT zinc transporter	678	10	472	NE	NE	NE	NE	NE
RL3186		putative transmembrane protein	27	95	5	NE	NE	NE	NE	NE
RL3273		putative von Willebrand factor type A	7	149	2	NE	NE	NE	NE	NE
RL3457		conserved hypothetical protein	15	52	70	NE	NE	NE	NE	NE
RL3701		putative transmembrane protein	28	97	6	NE	NE	NE	NE	NE
RL3704	asfZ	putative anti-sigma factor (no membrane-spanning domains), NepR-like AsfZ, ECF15 gene organisation (histidine kinase, response regulator, asf, ecf)	472	7	295	NE	NE	NE	NE	NE
RL3766	rpoH1	putative RNA polymerase sigma-32 factor (heat shock)	456	9	211	NE	NE	NE	ES	NE
RL3867		conserved hypothetical protein	17	36	18	NE	NE	NE	NE	NE
RL3982		conserved hypothetical protein	29	169	9	NE	NE	NE	NE	NE
RL4099		hypothetical protein	1	40	3	NE	NE	NE	NE	NE
RL4122	soxG1	putative sarcosine oxidase gamma subunit	474	3	231	NE	NE	NE	NE	NE
RL4535		conserved hypothetical exported protein	4	152	4	NE	NE	NE	ES	NE
Genes coexpressed with pRL100133										
pRL100071		putative permease component of ABC transporter PAAT	27	21	82	NE	NE	NE	NE	NE
pRL100072		putative solute binding-protein of ABC transporter PAAT	28	4	11	NE	NE	NE	ES	NE

Continued on next page

Table E.2 – continued from previous page

Gene Information			Scores			INSeq				
Gene ID	**Dsg**	**Description**	**CD**	**PR**	**EW**	**FL**	**Rh**	**R**	**NB**	**Bd**
pRL100255		conserved hypothetical protein	51	13	5	NE	NE	NE	NE	NE
pRL110184		putative LysR family transcriptional regulator	531	18	10	NE	NE	NE	NE	NE
pRL110198		hypothetical protein	3	2	8	NE	NE	NE	NE	NE
pRL110488		putative solute binding-protein of ABC transporter CUT1 (S. mel solute binding-protein homologue SMb21151 induced by galactosamine)	50	7	14	NE	NE	NE	NE	NE
pRL120637		putative transmembrane protein	21	9	12	NE	NE	NE	NE	NE
pRL70029		putative effector protein	6	51	1869	NE	NE	NE	NE	NE
pRL90280		conserved hypothetical protein	1500	14	7	NE	ES	NE	NE	NE
RL0017		putative polyhydroxybutyrate depolymerase	41	6	19	NE	NE	NE	NE	NE
RL0212	cpaA	putative pilus assembly protein	135	11	6	NE	NE	NE	NE	NE
RL0602		putative transmembrane ionic channel	4479	15	9	NE	NE	NE	NE	NE
RL0853		conserved hypothetical protein	1962	20	2	NE	NE	NE	NE	NE
RL1012	cbpA	putative curved DNA-binding protein	5	5	17	NE	NE	NE	NE	NE
RL2107		conserved hypothetical protein	96	10	21	NE	NE	NE	NE	NE
RL2509	citZ	putative citrate synthase I	1	1	1	NE	NE	NE	NE	NE
RL3197		putative short-chain dehydrogenase/oxidoreductase	3174	8	16	NE	NE	NE	NE	NE
RL3251		conserved hypothetical protein	18	12	15	NE	NE	NE	NE	NE
RL3269		putative RND family efflux transporter (inner membrane protein)	8	52	5625	NE	NE	NE	NE	NE

Continued on next page

Table E.2 – continued from previous page

Gene Information			Scores			INSeq				
Gene ID	**Dsg**	**Description**	**CD**	**PR**	**EW**	**FL**	**Rh**	**R**	**NB**	**Bd**
RL3654	pssD	putative polysaccharide biosynthesis protein	2	3	4	NE	NE	AD	ES	ES
RL3703	ecfZ	putative RNA polymerase ECF sigma factor, EcfZ, ECF15 gene organisation (histidine kinase, response regulator, asf, ecf) RpoE2 in S.meliloti	4	32	6055	NE	NE	NE	NE	NE
RL4384	arcA	putative arginase	7	31	6727	NE	NE	NE	NE	NE
RL4402		putative solute binding-protein of ABC transporter NitT	9	28	6745	NE	DE	DE	ES	ES
RL4505		putative acetyltransferase	3754	17	3	NE	NE	NE	NE	NE
RL4534		conserved hypothetical protein	10	68	6871	NE	NE	NE	AD	NE
Genes coexpressed with pRL100196										
pRL100097		conserved hypothetical protein	5	39	66	NE	NE	NE	NE	NE
pRL100098		conserved hypothetical protein	6	98	24	NE	NE	NE	NE	NE
pRL100099		conserved hypothetical protein	7	171	30	NE	NE	NE	NE	NE
pRL100103		putative hydroxyvalerate/ alcohol dehydrogenase	8	91	109	ES	NE	NE	NE	NE
pRL100104	phaE	putative phaE protein, subunit of polyhydroxyalkanoate (PHA) synthase	9	139	40	NE	NE	NE	NE	NE
pRL100105	phaC2	putative polyhydroxyalkanoate (PHA) synthase subunit C	10	65	27	NE	NE	NE	NE	NE
pRL100107		hypothetical protein	12	78	22	NE	NE	NE	NE	NE
pRL100111		hypothetical protein	3	20	19	DE	NE	DE	ES	ES
pRL100112		putative dehalogenase/hydrolase	13	24	21	NE	NE	NE	NE	NE
pRL100121	acsA3	putative acetyl-coA synthetase	16	88	45	NE	NE	NE	NE	NE

Continued on next page

Table E.2 – continued from previous page

Gene Information			Scores			INSeq				
Gene ID	**Dsg**	**Description**	**CD**	**PR**	**EW**	**FL**	**Rh**	**R**	**NB**	**Bd**
pRL100138		putative MerR family transcriptional regulator	18	40	11	NE	NE	NE	NE	NE
pRL100154		hypothetical protein	19	44	18	NE	NE	NE	NE	NE
pRL100156	fdxB1	putative ferredoxin	20	69	3	NE	NE	NE	NE	NE
pRL100162	nifH	nitrogenase iron protein NifH	25	136	7	NE	NE	NE	NE	DE
pRL100190		conserved hypothetical protein	1	4	14	NE	NE	NE	NE	NE
pRL100192		hypothetical protein	26	87	15	NE	NE	NE	NE	NE
pRL100193		conserved hypothetical protein	27	36	5	NE	NE	NE	NE	NE
pRL100194		hypothetical protein	28	32	1	NE	NE	NE	NE	NE
pRL100195	nifB	FeMo cofactor biosynthesis protein NifB	29	37	4	NE	NE	NE	NE	DE
pRL100197	fixX	ferredoxin-like protein FixX	30	74	13	NE	NE	NE	NE	DE
pRL100199	fixB	electron transfer protein FixB	32	76	6	NE	NE	NE	DE	DE
pRL100202		conserved hypothetical protein	35	50	9	NE	NE	AD	NE	NE
pRL100205	fixN1	putative transmembrane cytochrome oxidase subunit	36	22	2	NE	NE	NE	DE	DE
pRL100211		putative FNR/CRP family transcriptional regulator	41	84	36	NE	NE	NE	NE	NE
pRL90012		putative FNR/CRP family transcriptional regulator	78	18	73	NE	NE	NE	NE	NE
pRL90018	fixN2	putative cytochrome oxidase transmembrane component FixN	47	26	10	NE	ES	ES	ES	ES
RL0384	hspH	putative small heat shock protein	50	82	51	NE	NE	NE	NE	NE
RL1149		putative exported arylsulfatase	53	143	8	NE	NE	NE	NE	NE

Continued on next page

Table E.2 – continued from previous page

Gene Information			Scores			INSeq				
Gene ID	**Dsg**	**Description**	**CD**	**PR**	**EW**	**FL**	**Rh**	**R**	**NB**	**Bd**
RL1310		putative ATP-binding:ATP-binding (ABC:ABC) component of ABC transporter Export	215	2	98	NE	NE	NE	NE	NE
RL1419		putative serine protease	90	79	28	NE	NE	NE	NE	NE
RL1682		putative solute binding-protein of ABC transporter PhoT	169	9	117	NE	NE	NE	NE	NE
RL1801		putative transmembrane protein	151	5	108	NE	NE	NE	NE	NE
RL2213	clpA	putative ATP-dependent Clp protease ATP-binding subunit	212	8	115	NE	NE	NE	ES	NE
RL2531	folE	putative GTP cyclohydrolase I	69	29	25	DE	DE	ES	DE	ES
RL2580	sufC	putative ATP-binding component of ABC transporter Unclass	2	10	12	ES	ES	ES	ES	ES
RL2581		hypothetical protein	4	52	33	DE	ES	ES	ES	ES
RL3214		putative transmembrane NADH dehydrogenase	89	47	65	NE	NE	NE	NE	NE
RL3256	thyA	putative thymidylate synthase	182	3	126	ES	ES	ES	DE	ES
RL3424	dctA	putative C4-dicarboxylate transport protein	717	1	172	NE	NE	NE	NE	NE
RL3715		putative TetR family transcriptional regulator	281	7	163	AD	NE	NE	NE	NE
RL4100		putative transmembrane protein	75	61	50	NE	NE	NE	AD	NE
RL4612		putative transmembrane MFS family transport	144	6	70	NE	NE	NE	NE	NE
Genes coexpressed with pRL120291										
pRL100319		conserved hypothetical protein	2583	29	4	NE	NE	NE	NE	NE

Continued on next page

Table E.2 – continued from previous page

Gene Information			Scores			INSeq				
Gene ID	**Dsg**	**Description**	**CD**	**PR**	**EW**	**FL**	**Rh**	**R**	**NB**	**Bd**
pRL110465		conserved hypothetical protein	11	2	19	ES	ES	ES	NE	NE
pRL120196		putative fructose-bisphosphate aldolase	25	23	12	NE	NE	NE	NE	NE
pRL120238	ilvD	putative dihydroxy-acid dehydratase	12	9	7	NE	NE	NE	NE	NE
pRL120363		putative PIN domain signalling protein	421	8	30	NE	NE	NE	NE	NE
pRL120571		putative transmembrane tryptophan-rich protein	14	7	44	NE	AD	AD	NE	NE
pRL120619		putative aldolase	10	11	34	NE	NE	NE	NE	NE
pRL120720		conserved hypothetical protein	476	43	5	NE	AD	AD	NE	NE
pRL120738		putative transmembrane protein	1	1	25	NE	NE	NE	NE	NE
RL0232	optF	putative aminotransferase	40	18	18	NE	NE	NE	NE	NE
RL0256		putative XRE family (HipB) transcriptional regulator with cupin2 domian	52	22	22	NE	NE	NE	NE	NE
RL0586		putative transmembrane protein	4	10	50	NE	NE	NE	NE	NE
RL0604		putative UDP-glucose 4-epimerase	63	27	48	NE	NE	NE	NE	NE
RL0658		conserved hypothetical protein	6	17	9	NE	NE	NE	NE	NE
RL0914		conserved hypothetical protein	75	31	41	NE	NE	NE	AD	NE
RL1008		putative transmembrane GGDEF/EAL sensory box protein	81	12	2	NE	NE	NE	NE	NE
RL1434	feuQ	putative two-component sensor/regulator; histidine kinase	940	45	10	NE	DE	DE	ES	ES
RL1630		conserved hypothetical protein	5	14	31	NE	NE	NE	NE	NE
RL2129		conserved hypothetical protein	97	34	47	NE	NE	NE	NE	NE
RL2481		putative transmembrane GGDEF sensory box protein	42	13	33	NE	NE	NE	NE	NE

Continued on next page

Table E.2 – continued from previous page

Gene Information			Scores			INSeq				
Gene ID	**Dsg**	**Description**	**CD**	**PR**	**EW**	**FL**	**Rh**	**R**	**NB**	**Bd**
RL2529		conserved hypothetical protein	9	59	4907	NE	NE	NE	NE	NE
RL2535		putative rhomboid (intramembrane protease) family transmembrane protein	3	6	15	NE	NE	NE	NE	NE
RL2630		conserved hypothetical exported protein	1367	30	1	NE	NE	NE	NE	AD
RL2677		hypothetical protein	1391	50	3	NE	NE	NE	NE	NE
RL2885		putative acetyltransferase	2	4	11	NE	NE	NE	NE	NE
RL2891		putative cationic efflux system protein	7	5	23	NE	NE	NE	NE	NE
RL3265		putative AraC family transcriptional regulator (activator)	138	39	8	NE	NE	NE	NE	NE
RL3675		conserved hypothetical exported protein	8	3	21	NE	NE	NE	NE	NE
RL4168	abg	putative beta-glucosidase	1836	36	6	NE	NE	NE	NE	NE
RL4381		putative cell surface protein	29	37	35	NE	NE	NE	NE	NE
Genes coexpressed with pRL120397										
pRL110628	cobJ	precorrin-3b c(17)-methyltransferase (precorrin-3 methyltransferase) (precorrin-3 methylase).	4	174	10	DE	ES	ES	ES	DE
pRL110629	cobI	precorrin-2 c(20)-methyltransferase	2	122	2	DE	ES	ES	ES	DE
pRL110630	cobH	putative precorrin-8x methylmutase	1	178	1	DE	ES	ES	ES	DE
pRL120091		putative transmembrane protein	30	329	7	AD	NE	NE	NE	NE
pRL70002	repBp7a	putative replication protein RepB	1395	7	977	NE	NE	DE	NE	NE
RL0003		putative shikimate dehydrogenase	2023	10	945	NE	ES	ES	ES	ES

Continued on next page

Table E.2 – continued from previous page

Gene Information			Scores			INSeq				
Gene ID	**Dsg**	**Description**	**CD**	**PR**	**EW**	**FL**	**Rh**	**R**	**NB**	**Bd**
RL0199		putative permease component of ABC transporter CUT2	7	322	43	AD	NE	NE	NE	NE
RL0553	apaG	putative methioine biosynthesis related protein	1045	9	720	NE	NE	NE	NE	NE
RL1551	dnaC	putative DNA helicase	19	229	5	ES	ES	ES	ES	ES
RL1638	plsX	putative fatty acid/phospholipid synthesis cluster protein	8	230	14	ES	ES	ES	DE	DE
RL1645		conserved hypothetical protein	6	199	8	NE	NE	NE	NE	NE
RL1763		hypothetical protein	20	97	62	ES	ES	ES	ES	ES
RL1783	rpsQ	putative 30S ribosomal protein S17	21	98	19	ES	NE	DE	ES	ES
RL1948	gpt	putative xanthine-guanine phosphoribosyltransferase	66	176	4	NE	NE	NE	NE	NE
RL1954		conserved hypothetical protein	9	105	61	NE	NE	NE	NE	NE
RL2405		putative peptidyl-prolyl cis-trans isomerase B (cyclophilin-related protein)	76	348	6	NE	NE	NE	ES	DE
RL2472		putative DNAse	10	186	80	NE	NE	NE	NE	NE
RL2692		putative DAHP synthetase protein	1101	4	721	NE	NE	NE	NE	NE
RL2811		conserved hypothetical protein	1243	1	803	NE	NE	NE	NE	NE
RL2829	cobO	putative cob(I)yrinic acid a,c-diamide adenosyltransferase	583	8	289	NE	ES	ES	ES	ES
RL3191		putative exo/endo phosphatase	1173	5	812	NE	NE	NE	NE	NE
RL3309	murD	putative UDP-N-acetylmuramoylalanine–D-glutamate ligase	5	255	27	ES	ES	ES	ES	ES

Continued on next page

Table E.2 – continued from previous page

Gene Information			Scores			INSeq				
Gene ID	**Dsg**	**Description**	**CD**	**PR**	**EW**	**FL**	**Rh**	**R**	**NB**	**Bd**
RL3333		putative ATP-binding:ATP-binding (ABC:ABC) component of ABC transporter Unclass	23	90	42	NE	NE	NE	DE	ES
RL3422	greA	putative transcription elongation factor	899	3	441	NE	ES	ES	ES	ES
RL3732	ureY	putative urease operon 23 kda accessory protein (orf6)	1272	2	869	AD	NE	NE	NE	NE
RL4236	cyaB	putative adenylate/guanylate cyclase	3	96	3	NE	NE	NE	NE	NE
RL4634		putative solute binding-protein of ABC transporter PAAT	1103	6	861	NE	NE	NE	NE	NE
RL4734		conserved hypothetical protein	16	366	9	ES	DE	DE	ES	ES
Genes coexpressed with pRL120694										
pRL100001	repAp10a	putative replication protein RepA	188	516	8	ES	ES	ES	ES	ES
pRL100002	repBp10	putative replication protein RepB	189	376	9	ES	ES	ES	ES	ES
pRL100003	repCp10	putative replication protein RepC	190	2943	10	ES	ES	ES	ES	ES
pRL100031		hypothetical protein	66	30	37	NE	NE	NE	NE	NE
pRL100298		conserved hypothetical protein	4	2	7	NE	NE	NE	NE	NE
pRL120004		putative dihydrodipicolinate synthase	1187	10	1061	NE	NE	NE	AD	NE
pRL120211			6	103	1263					
pRL120362	katG	putative catalase-peroxidase	2	3	3	NE	NE	NE	NE	NE
pRL90270		putative solute binding-protein of ABC transporter Unclass	7	89	2365	NE	NE	NE	NE	NE
RL0425	ptsN	putative nitrogen regulatory IIA protein	2842	7	4	NE	NE	NE	DE	DE
RL0940	mntH	putative manganese transport protein	8	98	3345	NE	NE	NE	NE	NE

Continued on next page

Table E.2 – continued from previous page

Gene Information			Scores			INSeq				
Gene ID	**Dsg**	**Description**	**CD**	**PR**	**EW**	**FL**	**Rh**	**R**	**NB**	**Bd**
RL1302	osmC2	putative OsmC-like organic hydroperoxide resistance protein	33	5	6	NE	NE	NE	NE	NE
RL1966	aldA	putative alanine dehydrogenase	9	105	4357	NE	NE	NE	NE	NE
RL2444		conserved hypothetical protein with DUF1445	10	104	4818	NE	NE	NE	NE	NE
RL2508	gltA2	putative citrate synthase II	4823	6	5	NE	NE	NE	NE	NE
RL2604		conserved hypothetical protein	5	9	4969	NE	NE	NE	NE	NE
RL3441		conserved hypothetical protein	3	8	5795	NE	NE	NE	NE	NE
RL3869	rLuB	putative ribosomal large subunit pseudouridine synthase B	14	4	2	NE	NE	NE	NE	NE
RL3887	sitD	putative permease component of ABC transporter MZT (S. mel solute binding-protein homologue SMc02509 induced by manganese limitation)	1	1	1	NE	NE	NE	NE	NE
Genes coexpressed with pRL120695										
pRL100330		conserved hypothetical protein	895	22	6	NE	NE	NE	NE	NE
pRL110163		conserved hypothetical protein	14	7	12	NE	NE	NE	NE	NE
pRL110169		putative solute binding-protein of ABC transporter CUT2	19	17	18	NE	NE	NE	NE	NE
pRL110217		conserved hypothetical protein	61	18	15	NE	NE	NE	NE	NE
pRL110241		putative AraC family transcriptional regulator (activator)	6	45	687	NE	NE	NE	NE	NE
pRL110307		putative glutathione S-transferase	8	752	749	NE	NE	NE	NE	NE
pRL110329		conserved hypothetical protein	9	86	772	NE	NE	NE	NE	NE

Continued on next page

Table E.2 – continued from previous page

Gene Information			Scores			INSeq				
Gene ID	**Dsg**	**Description**	**CD**	**PR**	**EW**	**FL**	**Rh**	**R**	**NB**	**Bd**
pRL110375	ropA3	putative outer membrane porin protein RopA	270	20	8	NE	NE	NE	NE	NE
pRL120362	katG	putative catalase-peroxidase	24	6	22	NE	NE	NE	NE	NE
pRL120719		putative repetitive protein with two-component regulator histidine kinase and response regulator motifs	66	11	23	NE	NE	NE	NE	NE
pRL120743		putative LysR family transcriptional regulator	1277	19	10	NE	NE	NE	NE	NE
pRL80010		hypothetical protein	1308	25	9	NE	NE	NE	NE	NE
RL0355		putative pirin-like protein	32	10	26	NE	NE	NE	NE	NE
RL1036		conserved hypothetical protein	17	12	4	NE	NE	NE	NE	NE
RL1227		conserved hypothetical exported protein	100	16	2	NE	NE	NE	NE	NE
RL1302	osmC2	putative OsmC-like organic hydroperoxide resistance protein	101	14	7	NE	NE	NE	NE	NE
RL1463		conserved hypothetical protein	10	5	28	NE	NE	NE	NE	NE
RL3269		putative RND family efflux transporter (inner membrane protein)	12	2	20	NE	NE	NE	NE	NE
RL3442		conserved hypothetical protein	7	111	5801	NE	NE	NE	NE	NE
RL4078		putative transmembrane protein	5	112	6431	NE	NE	NE	NE	NE
RL4101		putative transmembrane protein	3	4	17	NE	NE	NE	AD	NE
RL4179		putative MFS family efflux transporter (inner membrane protein)	1	1	11	NE	NE	NE	NE	NE
RL4314		putative transmembrane ribonuclease protein	22	9	16	NE	NE	NE	NE	NE

Continued on next page

Table E.2 – continued from previous page

Gene Information			Scores			INSeq				
Gene ID	**Dsg**	**Description**	**CD**	**PR**	**EW**	**FL**	**Rh**	**R**	**NB**	**Bd**
RL4428		putative transmembrane cytochrome b562	4	8	3	NE	NE	NE	NE	NE
RL4481		putative GntR family transcriptional regulator	46	13	5	AD	NE	NE	NE	NE
RL4559		putative two-component sensor/regulator; transcriptional regulator	2	3	1	NE	ES	ES	ES	ES
Genes coexpressed with RL0149										
pRL100455		putative conserved hypothetical ATPase	16	16	45	NE	NE	NE	NE	NE
pRL110273		conserved hypothetical protein	9	12	60	NE	NE	NE	NE	NE
pRL120013		putative permease component of ABC transporter CUT1	66	10	29	NE	NE	NE	NE	NE
pRL120131		putative solute binding-protein of ABC transporter PepT	7	4	64	NE	NE	NE	NE	NE
pRL120448		putative AraC family transcriptional regulator (activator)	2	1	13	NE	NE	NE	NE	NE
RL0007		putative transmembrane FxsA family protein	799	25	3	NE	NE	NE	DE	NE
RL0049	hslV	putative heat-shock ATP-dependent protease	2204	55	10	AD	NE	NE	NE	NE
RL0101	gabD2	putative succinate-semialdehyde dehydrogenase [NAD(P)+]	3	15	9	NE	NE	NE	NE	NE
RL0480		putative TetR family transcriptional regulator	343	50	1	NE	NE	NE	NE	NE

Continued on next page

Table E.2 – continued from previous page

Gene Information			Scores			INSeq				
Gene ID	**Dsg**	**Description**	**CD**	**PR**	**EW**	**FL**	**Rh**	**R**	**NB**	**Bd**
RL0523		putative purine deaminase	1	3	61	NE	NE	NE	NE	NE
RL1239		hypothetical protein	20	14	11	NE	NE	NE	NE	NE
RL1311		conserved hypothetical protein	411	22	4	NE	NE	NE	NE	NE
RL1690	lon	putative ATP-dependent protease	5	18	23	DE	ES	ES	ES	ES
RL1998		putative permease component of ABC transporter CUT2	26	9	8	NE	NE	NE	NE	NE
RL2064		putative methyltransferase	469	26	5	NE	AD	NE	NE	NE
RL2303	ccdA	putative cytochrome c-type biogenesis protein	481	8	73	NE	DE	ES	ES	ES
RL2358		putative short-chain dehydrogenase/reductase	14	20	67	NE	NE	NE	NE	NE
RL2398	uvrA	putative UvrABC system protein A UvrA	8	7	7	NE	NE	NE	NE	NE
RL3140		putative methyltransferase	11	2	57	NE	NE	NE	NE	NE
RL3212		putative alcohol dehydrogenase	23	6	18	NE	NE	NE	NE	NE
RL3420		putative helicase/glycosylase	10	118	5785	NE	NE	NE	AD	NE
RL3498		conserved hypothetical protein	4	11	6	NE	NE	NE	AD	NE
RL3681		putative ATP-binding:ATP-binding (ABC:ABC) component of ABC transporter Export	21	21	55	NE	NE	NE	NE	NE
RL3995	lpiA	putative transmembrane acid tolerance protein	17	5	54	NE	NE	NE	NE	NE
RL4180		putative MFS family efflux transporter (membrane fusion protein)	6	24	31	NE	NE	NE	NE	NE

Continued on next page

Table E.2 – continued from previous page

Gene Information			Scores			INSeq				
Gene ID	**Dsg**	**Description**	**CD**	**PR**	**EW**	**FL**	**Rh**	**R**	**NB**	**Bd**
RL4322	tlpA	putative thiol:disulfide interchange protein	620	38	2	NE	NE	ES	NE	NE
RL4560		putative cNMP-binding protein	39	23	38	NE	NE	NE	NE	NE
Genes coexpressed with RL0652										
pRL100046		putative transposase-related protein	77	37	43	ES	ES	ES	ES	ES
RL0007		putative transmembrane FxsA family protein	38	50	18	NE	NE	NE	DE	NE
RL0162		conserved hypothetical protein	24	20	38	NE	NE	NE	NE	NE
RL0173	phnK	putative ATP-binding component of ABC transporter PhnP phosphate uptake	44	52	26	NE	NE	NE	NE	NE
RL0208		putative transmembrane protein	39	44	58	NE	NE	NE	NE	NE
RL0221		conserved hypothetical protein atu0215 (agr_c_369p)	19	31	17	NE	NE	NE	NE	NE
RL0232	optF	putative aminotransferase	9	25	19	NE	NE	NE	NE	NE
RL0245		putative CDP-glucose 4,6-dehydratase	20	36	12	NE	NE	NE	NE	NE
RL0256		putative XRE family (HipB) transcriptional regulator with cupin2 domian	22	11	13	NE	NE	NE	NE	NE
RL0292		conserved hypothetical protein	16	14	4	NE	NE	NE	NE	NE
RL0404	mviN	putative transmembrane MviN virulence factor homologue	43	34	36	DE	ES	ES	ES	ES
RL0586		putative transmembrane protein	7	5	7	NE	NE	NE	NE	NE
RL0593		conserved hypothetical protein	82	62	34	NE	NE	NE	NE	NE

Continued on next page

Table E.2 – continued from previous page

Gene Information			Scores			INSeq				
Gene ID	**Dsg**	**Description**	**CD**	**PR**	**EW**	**FL**	**Rh**	**R**	**NB**	**Bd**
RL0594		putative fatty oxidation complex subunit	25	28	47	NE	NE	NE	NE	NE
RL0604		putative UDP-glucose 4-epimerase	18	12	8	NE	NE	NE	NE	NE
RL0610		hypothetical exported protein	12	26	50	NE	NE	NE	NE	NE
RL0617		putative Maf septum inhibitor protein	26	21	24	NE	NE	NE	NE	NE
RL0628		putative AraC family transcriptional regulator (activator)	13	18	29	NE	NE	NE	NE	NE
RL0634		putative short-chain dehydrogenase	3	7	1	AD	NE	NE	NE	NE
RL0641		putative ROK family transcriptional regulator	17	40	11	NE	NE	NE	NE	NE
RL0647		putative GFO/IDH/MocA family oxidoreductase	23	19	20	NE	NE	NE	NE	NE
RL0658		conserved hypothetical protein	6	24	6	NE	NE	NE	NE	NE
RL0666		putative exported lipase	27	30	70	NE	NE	NE	NE	NE
RL0690	cheR1	putative chemotaxis protein methyltransferase	28	23	41	NE	NE	NE	ES	ES
RL0914		conserved hypothetical protein	29	17	33	NE	NE	NE	AD	NE
RL0949	mcpJ	putative methyl-accepting chemotaxis protein McpJ	64	1	73	NE	NE	NE	NE	NE
RL0960		conserved hypothetical protein (Sua5 family)	30	38	30	DE	ES	ES	ES	ES
RL0973	dxs	putative 1-deoxy-d-xylulose-5-phosphate synthase	31	35	63	DE	ES	ES	ES	ES
RL0974		conserved hypothetical protein	32	29	39	NE	NE	NE	NE	NE

Continued on next page

Table E.2 – continued from previous page

Gene Information			Scores			INSeq				
Gene ID	**Dsg**	**Description**	**CD**	**PR**	**EW**	**FL**	**Rh**	**R**	**NB**	**Bd**
RL0990	matA	putative malonyl-CoA decarboxylase	4	2	10	NE	NE	NE	NE	NE
RL1008		putative transmembrane GGDEF/EAL sensory box protein	8	8	5	NE	NE	NE	NE	NE
RL1014	pdxH	putative pyridoxamine 5'-phosphate oxidase	33	22	22	DE	ES	ES	ES	ES
RL1032	rnhA1	putative ribonuclease HI	14	32	21	NE	ES	ES	ES	ES
RL1038		conserved hypothetical protein	10	3	27	NE	NE	NE	NE	NE
RL1334		putative TetR family transcriptional regulator	21	47	65	NE	NE	NE	NE	NE
RL1339		conserved hypothetical protein	2	9	3	NE	NE	NE	NE	NE
RL1363		putative oxidoreductase	1	4	2	NE	NE	NE	NE	NE
RL1369		conserved hypothetical protein with pentapeptide repeat	34	15	9	NE	NE	NE	NE	NE
RL1393	pbpF	putative peptidoglycan biosynthesis/penicillin binding protein	62	45	71	NE	NE	NE	NE	NE
RL1417		putative IolH myo-inositol catabolism protein	5	6	15	NE	NE	NE	NE	NE
RL1503		putative RNA binding protein	35	33	51	NE	DE	ES	ES	ES
RL1725		putative DinB DNA polymerase IV	36	48	49	NE	NE	NE	NE	NE
RL2129		conserved hypothetical protein	11	27	14	NE	NE	NE	NE	NE
RL2173		hypothetical protein	86	41	72	NE	NE	NE	NE	NE
RL2458		putative LysR family transcriptional regulator	72	49	59	NE	NE	NE	NE	NE
RL2481		putative transmembrane GGDEF sensory box protein	15	13	28	NE	NE	NE	NE	NE

Continued on next page

Table E.2 – continued from previous page

Gene Information			Scores			INSeq				
Gene ID	**Dsg**	**Description**	**CD**	**PR**	**EW**	**FL**	**Rh**	**R**	**NB**	**Bd**
RL2535		putative rhomboid (intramembrane protease) family transmembrane protein	69	10	74	NE	NE	NE	NE	NE
Genes coexpressed with RL0727										
pRL110001	repAp11	putative replication protein RepA	6	27	5	DE	ES	ES	ES	DE
pRL110096		conserved hypothetical protein	15	28	13	NE	NE	NE	NE	NE
pRL110115		putative DNA ligase	28	48	58	NE	NE	NE	NE	NE
pRL110239		conserved hypothetical protein	1914	4	93	NE	NE	NE	NE	NE
pRL110398		hypothetical protein	19	34	34	NE	NE	NE	NE	NE
pRL110428		conserved hypothetical protein	3	36	9	NE	NE	NE	NE	NE
pRL120057		putative plasmid stability protein	71	76	21	NE	NE	NE	NE	NE
pRL120410		putative LysR family transcriptional regulator	6321	1	114	NE	NE	NE	NE	NE
RL0159		putative MarR family transcriptional regulator	32	26	33	NE	NE	NE	NE	NE
RL0212	cpaA	putative pilus assembly protein	14	23	27	NE	NE	NE	NE	NE
RL0266	infC	putative translation initiation factor	35	13	71	ES	ES	NE	NE	NE
RL0401		putative universal stress protein	18	58	6	NE	NE	NE	NE	DE
RL0425	ptsN	putative nitrogen regulatory IIA protein	8	44	19	NE	NE	NE	DE	DE
RL0490	frcC	putative permease component of ABC transporter CUT2 fructose transporter from homology with S.mel SMc02170	84	3	109	NE	NE	NE	NE	NE
RL0602		putative transmembrane ionic channel	41	43	98	NE	NE	NE	NE	NE
RL0703	motA	putative chemotaxis motility protein	1	5	3	NE	NE	NE	NE	DE

Continued on next page

Table E.2 – continued from previous page

Gene Information			Scores			INSeq				
Gene ID	**Dsg**	**Description**	**CD**	**PR**	**EW**	**FL**	**Rh**	**R**	**NB**	**Bd**
RL0708	flgC	putative flagellar basal-body (proximal) rod protein FlgC	65	9	110	NE	NE	NE	NE	NE
RL0731	flaF1	putative flagellar synthesis related protein FlaF	66	8	52	NE	NE	NE	DE	DE
RL0732	flbT	putative flagellar biosynthesis repressor protein, binds to and promotes degradation of flagellar mRNA (post-transcriptional regulation)	40	37	4	NE	NE	NE	DE	DE
RL0733	flgD	putative flagellar hook formation protein FlgD	552	81	2	NE	NE	NE	DE	NE
RL0803		conserved hypothetical protein	69	47	44	NE	NE	NE	NE	NE
RL0958		conserved hypothetical protein	25	52	38	NE	NE	NE	NE	NE
RL1379	rosR	putative MucR/RosR family transcriptional regulator involved in nodulation competitiveness RosR	27	2	17	NE	ES	ES	ES	DE
RL1634	hppA	putative transmembrane pyrophosphate-energized proton pump	5	11	39	NE	NE	NE	NE	NE
RL1729	divK	response regulator receiver protein	13	14	8	DE	ES	ES	NE	NE
RL2102	cspA5	putative cold shock protein	4	55	7	NE	NE	NE	ES	NE
RL2132		conserved hypothetical protein	31	64	54	NE	NE	NE	NE	NE
RL2508	gltA2	putative citrate synthase II	26	31	28	NE	NE	NE	NE	NE
RL2527		conserved hypothetical exported protein	7	16	16	NE	NE	ES	DE	NE

Continued on next page

Table E.2 – continued from previous page

Gene Information			Scores			INSeq				
Gene ID	**Dsg**	**Description**	**CD**	**PR**	**EW**	**FL**	**Rh**	**R**	**NB**	**Bd**
RL2653		putative HPt (Histidine Phosphotransfer) domain protein of two-component system (multistage phosphorelay can result)	12	30	25	NE	NE	NE	NE	NE
RL2705		putative GGDEF sensory box protein	11	39	68	NE	NE	NE	NE	NE
RL3150		conserved hypothetical protein	2	19	1	NE	NE	NE	NE	NE
RL3323		putative transmembrane efflux protein	10	29	11	NE	NE	NE	NE	NE
RL3613	araE	putative arabinose dehydrogenase (elevated on MA L-arabinose and galactose Rlv3841)	23	10	72	NE	NE	NE	DE	NE
RL3792		putative hydroxypyruvate isomerase	39	45	81	NE	NE	NE	NE	NE
RL3964	glmM	putative phosphoglucosamine mutase	1703	6	43	ES	ES	ES	ES	ES
RL4082		putative peptidoglycan binding protein	9	21	20	NE	NE	NE	NE	NE
RL4124	soxD1	putative sarcosine oxidase delta subunit	29	12	62	NE	NE	NE	NE	NE
RL4167		putative transmembrane GGDEF/EAL sensory box protein	20	42	10	NE	NE	NE	NE	NE
RL4226	osmC	putative osmotically inducible protein C OsmC	34	53	85	NE	NE	NE	NE	NE
RL4320	etfA1	putative electron transfer flavoprotein alpha-subunit	22	41	63	NE	NE	NE	NE	NE
RL4386	cheW2	putative chemotaxis protein	16	17	47	NE	NE	NE	DE	NE
RL4427		conserved hypothetical exported protein	17	24	92	NE	NE	NE	NE	NE
RL4528		putative 2'-5' RNA ligase	50	7	40	NE	NE	NE	NE	NE

Continued on next page

Table E.2 – continued from previous page

Gene Information			Scores			INSeq				
Gene ID	**Dsg**	**Description**	**CD**	**PR**	**EW**	**FL**	**Rh**	**R**	**NB**	**Bd**
Genes coexpressed with RL2924										
pRL110173		putative transmembrane protein	96	12	14	NE	AD	AD	NE	NE
pRL110180		putative 2-methylcitrate dehydratase	97	11	6	NE	NE	NE	NE	NE
pRL110204		putative formiminoglutamase	65	7	10	NE	NE	NE	NE	NE
pRL120085		conserved hypothetical protein	8	100	1147	NE	NE	NE	NE	NE
pRL120277	ligI	putative 2-pyrone-4,6-dicarboxylic acid hydrolase	70	8	8	NE	NE	NE	NE	NE
pRL120301	yagS	putative xanthine dehydrogenase YagS, FAD-binding subunit	3	15	1356	NE	NE	NE	NE	NE
pRL90059	rmrA	putative MFS family efflux transporter (membrane fusion protein)	1466	14	4	NE	NE	NE	NE	NE
pRL90170		putative major facilitator superfamily (MFS) transporter	91	10	5	NE	NE	NE	NE	NE
RL0289		conserved hypothetical exported protein	9	22	2704	NE	NE	NE	NE	NE
RL0484		putative short-chain oxidoreductase	5	194	2898	NE	NE	NE	NE	NE
RL2646		putative transmembrane protein	10	5	11	NE	NE	NE	NE	NE
RL2900	dhaK1	putative dihydroxyacetone kinase	100	13	3	NE	NE	NE	NE	NE
RL2906	dhaK2	putative PTS-dependent dihydroxyacetone kinase, dihydroxyacetone binding subunit	83	9	7	NE	NE	NE	NE	NE
RL2925	mutT	putative mutator MutT protein	1	6	1	NE	NE	NE	NE	NE
RL3002	dapB2	putative dihydrodipicolinate reductase	69	4	9	NE	NE	NE	NE	NE
RL3216	asr	putative aspartate racemase	4	126	5573	NE	NE	NE	NE	NE

Continued on next page

Table E.2 – continued from previous page

Gene Information			Scores			INSeq				
Gene ID	**Dsg**	**Description**	**CD**	**PR**	**EW**	**FL**	**Rh**	**R**	**NB**	**Bd**
RL3784		putative MFS family efflux transporter (membrane fusion protein)	7	2	12	AD	NE	NE	NE	NE
RL4302		putative oxidoreductase	2	1	13	NE	NE	NE	NE	NE
RL4332		putative transmembrane protein	6	288	6677	NE	NE	NE	NE	NE
RL4378		putative LacI family transcriptional regulator (repressor)	76	3	2	NE	NE	NE	AD	NE
Genes coexpressed with RL3426										
pRL100017		conserved hypothetical protein	17	68	39	NE	AD	AD	NE	NE
pRL110219		conserved hypothetical exported protein	21	24	75	NE	NE	NE	NE	NE
pRL110273		conserved hypothetical protein	2	11	9	NE	NE	NE	NE	NE
pRL110343		conserved hypothetical protein	7	31	44	NE	NE	NE	NE	NE
pRL110501		putative ribonuclease	1715	9	182	NE	AD	AD	NE	NE
pRL120019		putative permease component of ABC transporter CUT1	822	8	245	NE	NE	NE	NE	NE
pRL120061		hypothetical protein	9	32	76	NE	NE	NE	NE	NE
pRL120137		conserved hypothetical protein	41	92	12	NE	AD	AD	NE	NE
pRL120395		conserved hypothetical protein	216	10	235	NE	NE	NE	NE	NE
pRL120401		putative small permease (DctQ-like) component of TRAP transporter (S. mel solute binding-protein homologue SMb20442 induced by fucose, mannose)	42	35	57	NE	NE	NE	NE	NE
pRL80066		hypothetical protein	12	34	53	NE	NE	NE	NE	NE
RL0079	acsAB	putative transmembrane cellulose synthase	30	26	48	NE	NE	NE	NE	NE

Continued on next page

Table E.2 – continued from previous page

Gene Information			Scores			INSeq				
Gene ID	**Dsg**	**Description**	**CD**	**PR**	**EW**	**FL**	**Rh**	**R**	**NB**	**Bd**
RL0458		putative adenylate cyclase	59	185	10	NE	NE	NE	NE	NE
RL0551	hslO	putative Hsp33-like chaperonin	152	119	8	NE	NE	NE	NE	NE
RL0837		hypothetical protein	8	19	16	NE	NE	NE	NE	NE
RL1193		putative short-chain reductase	70	187	3	NE	NE	NE	NE	NE
RL1211		conserved hypothetical protein	1	15	4	NE	NE	NE	NE	NE
RL1217		putative TetR family transcriptional regulator	71	102	1	NE	NE	NE	NE	ES
RL1257	qor2	putative quinone oxidoreductase	16	75	99	NE	NE	NE	NE	NE
RL1572		putative DNA polymerase III subunit	76	64	7	NE	NE	DE	ES	ES
RL1684		putative permease component of ABC transporter PhoT	23	22	54	NE	NE	NE	NE	NE
RL1952		putative transmembrane protein	6	80	36	NE	NE	NE	NE	NE
RL1970		putative flavin reductase	5	38	18	NE	NE	NE	NE	NE
RL2070		putative TetR family transcriptional regulator	22	51	61	NE	NE	NE	NE	NE
RL2286		putative MazG family protein	3	41	2	NE	NE	NE	NE	NE
RL2352		conserved hypothetical protein	4	36	5	NE	NE	NE	NE	NE
RL2422		putative LacI family transcriptional regulator (repressor)	82	95	6	NE	NE	NE	NE	NE
RL2447		putative LacI family transcriptional regulator (repressor)	157	5	64	NE	NE	NE	NE	NE
RL2452		conserved hypothetical protein	24	86	29	NE	NE	NE	NE	NE
RL2499	iorA	putative isoquinoline 1-oxidoreductase alpha subunit	10	169	79	NE	NE	NE	NE	NE
RL2999		putative glyoxylase	13	69	22	NE	NE	NE	NE	NE

Continued on next page

Table E.2 – continued from previous page

Gene Information			Scores			INSeq				
Gene ID	**Dsg**	**Description**	**CD**	**PR**	**EW**	**FL**	**Rh**	**R**	**NB**	**Bd**
RL3118		putative GntR family transcriptional regulator	902	2	149	NE	NE	NE	NE	NE
RL3201		putative PIN/PilT family protein	1178	7	213	NE	NE	NE	NE	NE
RL3420		putative helicase/glycosylase	3658	1	129	NE	NE	NE	AD	NE
RL3498		conserved hypothetical protein	31	42	49	NE	NE	NE	AD	NE
RL3806		putative permease component of ABC transporter CUT1	158	4	70	NE	NE	NE	NE	NE
RL3943		hypothetical protein	1115	3	242	NE	NE	NE	NE	NE
RL4660		putative acetyltransferase	99	100	15	AD	NE	NE	NE	NE
RL4666	pinE	putative DNA-invertase from lambdoid prophage E14	1116	6	136	AD	NE	NE	NE	NE

List of the genes coexpressed with the studied transcriptional regulator. We use the methodoly described in Chapter 4 to retreieve them. "Dsg" indicates the gene name. The columns under the "Scores" title indicate the rank of the genes in for the three methods employed in the coexpresssoin analysis: community detection co-occurrence (CD), personalised Pagerank (PR), and edge weights (EW). The last five columns indicate the INSeq value for the genes in free-living conditions (FL), the rhizosphere (Rh), the root (R), nodule bacteria (NB), and bacteroid (Bd), according to (Wheatley *et al.*, 2020); possible values for these columns are essential (ES), deffective (DE), neutral (NE), and advantaged (AD).

Table E.3: Genes differentially expressed in the mutants

Gene Information				Mutant vs WT			INSeq					DE B
Gene ID	**Old ID**	**Dsg**	**Description**	**FC**	**m**	**WT**	**FL**	**Rh**	**R**	**NB**	**Bd**	**FC**
Genes upregulated in pRL80079 mutant												
RL_RS02350	RL0451	cspA2	putative cold shock protein	3.27	163.04	49.80	NE	NE	NE	NE	NE	1.13
RL_RS07335	RL1409		conserved hypothetical protein	3.85	8.58	2.23	NE	AD	AD	NE	NE	3.72
RL_RS22690	RL4399		conserved hypothetical protein	3.69	17.60	4.77	NE	DE	DE	ES	ES	1.40
RL_RS22700	RL4401		YKOF super family putative ATP-binding component of ABC transporter NitT	3.11	14.05	4.51	NE	DE	DE	ES	ES	0.48
RL_RS25535	pRL80077		putative molybdenum-binding oxidoreductase	4.15	20.82	5.01	NE	NE	NE	NE	NE	0.82
RL_RS25550	pRL80080		putative fructokinase	136.35	781.49	5.73	NE	NE	NE	DE	DE	-0.14
RL_RS25555	pRL80081		putative hydrolase	141.54	1204.84	8.51	NE	NE	NE	DE	NE	-0.09
RL_RS25560	pRL80082		putative biotin-binding protein	116.55	1815.24	15.57	NE	NE	NE	NE	NE	0.08
RL_RS25565	pRL80083		putative dehydrogenase, fusion	81.22	1531.43	18.86	NE	NE	NE	NE	NE	1.18
RL_RS25570	pRL80084		putative epimerase/isomerase	37.28	665.95	17.87	NE	NE	NE	NE	NE	2.00

Continued on next page

Table E.3 – continued from previous page

Gene Information				Mutant vs WT			INSeq					DE B
Gene ID	**Old ID**	**Dsg**	**Description**	**FC**	**m**	**WT**	**FL**	**Rh**	**R**	**NB**	**Bd**	**FC**
RL_RS25575	pRL80085		putative solute binding-protein of ABC transporter CUT2 homoserine transporter (Hynes)	49.81	466.90	9.37	NE	NE	NE	NE	NE	0.89
RL_RS25580	pRL80086		putative ATP-binding:ATP-binding (ABC:ABC) component of ABC transporter CUT2 homoserine transporter (Hynes)	35.87	167.20	4.66	NE	NE	NE	NE	NE	0.68
RL_RS25585	pRL80087		putative permease component of ABC transporter CUT2 homoserine transporter (Hynes)	5.06	88.88	17.57	NE	NE	NE	NE	NE	1.51
RL_RS25590	pRL80088		putative permease component of ABC transporter CUT2 homoserine transporter (Hynes)	23.02	66.33	2.88	NE	AD	NE	NE	NE	0.06
RL_RS25595	pRL80089		pseudogene	13.08	63.06	4.82						0.34
RL_RS25600	pRL80090		putative fucose operon protein	11.58	33.95	2.93	NE	DE	NE	NE	NE	0.47

Continued on next page

Table E.3 – continued from previous page

Gene Information				Mutant vs WT			INSeq					DE B
Gene ID	**Old ID**	**Dsg**	**Description**	**FC**	**m**	**WT**	**FL**	**Rh**	**R**	**NB**	**Bd**	**FC**
RL_RS25605	pRL80091		pseudogene	23.67	35.28	1.49						-0.09
RL_RS25610	pRL80092		putative permease component of ABC transporter CUT1	11.93	24.46	2.05	NE	NE	NE	NE	NE	0.77
RL_RS25615	pRL80093		putative permease component of ABC transporter CUT1	8.69	18.96	2.18	NE	NE	NE	NE	NE	0.66
RL_RS25620	pRL80094		putative solute binding-protein of ABC transporter CUT1	11.11	66.57	5.99	NE	NE	NE	NE	NE	-0.42
RL_RS25625	pRL80095		putative ATP-binding component of ABC transporter CUT1	12.51	13.52	1.08	NE	NE	NE	NE	NE	-0.81
Genes downregulated in pRL80079 mutant												
RL_RS03195	RL0610		hypothetical exported protein	0.33	7.39	22.70	NE	NE	NE	NE	NE	-0.17
RL_RS05150	RL0990	matA	putative malonyl-CoA decarboxylase	0.26	5.14	19.47	NE	NE	NE	NE	NE	2.58
RL_RS05740	RL1105		putative TetR family transcriptional regulator	0.29	6.70	22.91	NE	DE	ES	ES	ES	0.39

Continued on next page

Table E.3 – continued from previous page

Gene Information				Mutant vs WT			INSeq					DE B
Gene ID	**Old ID**	**Dsg**	**Description**	**FC**	**m**	**WT**	**FL**	**Rh**	**R**	**NB**	**Bd**	**FC**
RL_RS05745	RL1106	pspA	putative (phage shock protein A) PspA family regulator by protein-protein interactions	0.18	53.14	293.48	NE	NE	NE	NE	NE	3.07
RL_RS05750	RL1107		putative transmembrane protein YiaAB domain protein	0.30	31.03	104.25	NE	NE	NE	NE	NE	2.22
RL_RS06380	RL1227		conserved hypothetical exported protein	0.16	1.55	9.93	NE	NE	NE	NE	NE	0.25
RL_RS06385	RL1228		putative TetR family transcriptional regulator	0.17	5.16	29.69	NE	NE	NE	NE	NE	1.24
RL_RS13765	RL2667		putative RND family efflux transporter (membrane fusion protein)	0.26	8.83	34.56	NE	NE	NE	NE	NE	-0.30
RL_RS15065	RL2912		conserved hypothetical exported protein	0.25	8.31	33.69	NE	NE	NE	NE	NE	3.04
RL_RS15070	RL2913		putative GntR family transcriptional regulator	0.27	3.27	11.92	NE	NE	NE	NE	NE	-0.53

Continued on next page

Table E.3 – continued from previous page

Gene Information				Mutant vs WT			INSeq					DE B
Gene ID	**Old ID**	**Dsg**	**Description**	**FC**	**m**	**WT**	**FL**	**Rh**	**R**	**NB**	**Bd**	**FC**
RL_RS19295	RL3741		putative solute binding-protein of ABC transporter HAAT (S. mel solute binding-protein homologue SMb20605 induced by N-limitation, MA Rlv3841 elevated on N-limit. and Phe) urea/short chain amide	0.12	1.75	14.04	NE	NE	NE	NE	NE	0.73
RL_RS36390	pRL120695		putative TetR family transcriptional regulator of RND family efflux transporter	0.26	10.10	39.55	NE	NE	NE	DE	DE	2.04
RL_RS36395	pRL120696		putative RND family efflux transporter (membrane fusion protein)	0.26	16.06	62.77	NE	NE	NE	DE	NE	1.37
RL_RS36400	pRL120697		putative RND family efflux transporter (membrane fusion protein)	0.23	4.63	19.96	NE	NE	NE	NE	NE	0.01

Continued on next page

Table E.3 – continued from previous page

Gene Information				Mutant vs WT			INSeq					DE B
Gene ID	**Old ID**	**Dsg**	**Description**	**FC**	**m**	**WT**	**FL**	**Rh**	**R**	**NB**	**Bd**	**FC**
RL_RS26180	pRL90060	rmrB	putative MFS family efflux transporter (inner membrane protein)	0.33	12.85	38.91	NE	NE	NE	NE	NE	1.74
RL_RS25475	pRL80063		putative ATP-binding component of ABC transporter PAAT closest homol mimosine transporter Rhizobium sp. TAL1145	0.29	2.71	9.31	NE	NE	NE	NE	NE	2.18
RL_RS32680	pRL110578		putative 4-oxalocrotonate tautomerase	0.19	3.38	17.98	NE	NE	NE	NE	NE	-4.71
Genes upregulated in pRL90058 mutant												
RL_RS02350	RL0451	cspA2	putative cold shock protein	4.44	221.04	49.80	NE	NE	NE	NE	NE	1.13
RL_RS07335	RL1409		conserved hypothetical protein	5.54	12.35	2.23	NE	AD	AD	NE	NE	3.72
RL_RS10905	RLt26	tRNA-Glu	tRNA-Glu	3.44	9.81	2.85	NE	NE	NE	NE	NE	4.17
RL_RS28610	pRL100242		conserved hypothetical protein	3.38	42.58	12.60	NE	NE	ES	NE	NE	1.74
RL_RS24465	pRL70014		conserved hypothetical protein	3.73	12.25	3.28	NE	NE	NE	NE	NE	-0.49

Continued on next page

Table E.3 – continued from previous page

Gene Information				Mutant vs WT			INSeq					DE B
Gene ID	**Old ID**	**Dsg**	**Description**	**FC**	**m**	**WT**	**FL**	**Rh**	**R**	**NB**	**Bd**	**FC**
RL_RS30170	pRL110082		putative TetR family transcriptional regulator	3.59	19.65	5.47	NE	NE	NE	NE	NE	-0.80
Genes downregulated in pRL90058 mutant												
RL_RS05745	RL1106	pspA	putative (phage shock protein A) PspA family regulator by protein-protein interactions	0.26	77.71	293.48	NE	NE	NE	NE	NE	3.07
RL_RS06380	RL1227		conserved hypothetical exported protein	0.20	2.03	9.93	NE	NE	NE	NE	NE	0.25
RL_RS19295	RL3741		putative solute binding-protein of ABC transporter HAAT (S. mel solute binding-protein homologue SMb20605 induced by N-limitation, MA Rlv3841 elevated on N-limit. and Phe) urea/short chain amide	0.23	3.29	14.04	NE	NE	NE	NE	NE	0.73
RlC57				0.33	17.14	52.12						-1.76

Continued on next page

Table E.3 – continued from previous page

Gene Information				Mutant vs WT			INSeq					DE B
Gene ID	**Old ID**	**Dsg**	**Description**	**FC**	**m**	**WT**	**FL**	**Rh**	**R**	**NB**	**Bd**	**FC**
RL_RS26180	pRL90060	rmrB	putative MFS family efflux transporter (inner membrane protein)	0.20	7.86	38.91	NE	NE	NE	NE	NE	1.74
RL_RS25345	pRL80036		conserved hypothetical protein	0.30	9.28	31.00	NE	NE	NE	NE	NE	4.75
RL_RS25350	pRL80037		putative short-chain dehydrogenase	0.33	5.03	15.36	NE	NE	NE	NE	NE	3.03
RL_RS32680	pRL110578		putative 4-oxalocrotonate tautomerase	0.15	2.75	17.98	NE	NE	NE	NE	NE	-4.71
Genes upregulated in pRL100133 mutant												
RL_RS28015	pRL100127		putative ATP-binding component of ABC transporter Unclass	5.03	33.93	6.74	NE	NE	NE	NE	NE	0.03
RL_RS28020	pRL100128		putative ATP-binding component of ABC transporter Unclass	6.72	32.42	4.82	NE	NE	NE	NE	NE	0.87
RL_RS28025	pRL100129		putative permease component of ABC transporter Unclass	37.16	40.15	1.08	NE	NE	NE	NE	NE	-0.97
RL_RS28030	pRL100130		putative permease component of ABC transporter Unclass	56.79	53.98	0.95	NE	NE	NE	NE	NE	-1.28

Continued on next page

Table E.3 – continued from previous page

Gene Information				Mutant vs WT			INSeq					DE B
Gene ID	**Old ID**	**Dsg**	**Description**	**FC**	**m**	**WT**	**FL**	**Rh**	**R**	**NB**	**Bd**	**FC**
RL_RS28035	pRL100131		putative solute binding-protein of ABC transporter	85.50	76.99	0.90	NE	NE	NE	NE	NE	-1.49
RL_RS28045	pRL100134	aatK	Unclass orthologue of aatK (blcA) succinate semialdehyde dehydrogenase	13.64	51.18	3.75	NE	NE	NE	DE	DE	-0.49
RL_RS28050	pRL100135	aatL	dehydrogenase gene of A. tumefaciens orthologue of aatL (blcB) putative 4-hydroxybutyrate dehydrogenase	9.48	58.26	6.14	NE	NE	NE	NE	NE	-0.35
RL_RS28055	pRL100136	aatM	gene of A. tumefaciens orthologue of aatM (blcC) lactonase gene of A. tumefaciens	11.44	52.18	4.56	NE	NE	NE	NE	NE	0.13
Genes downregulated in pRL100133 mutant												
RL_RS05155	RL0991	matB	putative malonyl-CoA synthetase	0.33	3.81	11.46	NE	NE	NE	NE	NE	0.99
RL_RS11930	RL2307		conserved hypothetical protein	0.27	3.14	11.50	NE	NE	NE	DE	DE	0.46

Continued on next page

Table E.3 – continued from previous page

Gene Information				Mutant vs WT			INSeq					DE B
Gene ID	**Old ID**	**Dsg**	**Description**	**FC**	**m**	**WT**	**FL**	**Rh**	**R**	**NB**	**Bd**	**FC**
RL_RS15825	RL3065		conserved hypothetical exported protein	0.21	7.09	33.89	NE	NE	NE	NE	NE	2.22
RL_RS15830	RL3066		putative transmembrane protein	0.09	1.93	22.11	NE	NE	NE	NE	NE	4.63
RL_RS27625	pRL100041		hypothetical protein	0.32	2.97	9.16	NE	NE	NE	NE	NE	1.97
Genes downregulated in pRL100196 mutant												
RL_RS00090	RL0018		putative transmembrane protein	0.27	1600.42	5936.03	NE	NE	NE	NE	NE	1.24
RL_RS01215	RL0231		conserved hypothetical protein	0.31	548.14	1781.60						4.31
RL_RS01535	RL0293		putative methylated-DNA–protein-cysteine methyltransferase	0.32	63.42	198.28	NE	NE	NE	NE	NE	0.65
RL_RS02640	RL0505	fadD	putative long-chain-fatty-acid–CoA ligase	0.17	41.38	249.16	NE	NE	NE	NE	NE	2.14
RL_RS03100	RL0591		putative acyl-CoA dehydrogenase	0.30	62.70	211.38	NE	NE	NE	NE	NE	1.15
RL_RS04620	RL0883	groLch1	putative 60 kda chaperonin (protein cpn60)	0.31	547.02	1741.40	ES	ES	ES	ES	DE	0.10

Continued on next page

Table E.3 – continued from previous page

Gene Information				Mutant vs WT			INSeq					DE B
Gene ID	**Old ID**	**Dsg**	**Description**	**FC**	**m**	**WT**	**FL**	**Rh**	**R**	**NB**	**Bd**	**FC**
RL_RS05025	RL0965		putative acyl-CoA dehydrogenase	0.22	30.35	135.32	NE	NE	NE	NE	NE	2.52
RL_RS08365	RL1615		putative transmembrane protein	0.30	116.36	384.05	NE	NE	NE	NE	NE	0.68
RL_RS11470	RL2213	clpA	putative ATP-dependent Clp protease ATP-binding subunit	0.33	446.35	1372.80	NE	NE	NE	ES	NE	2.52
RL_RS15830	RL3066		putative transmembrane protein	0.15	8.24	55.33	NE	NE	NE	NE	NE	4.63
RL_RS17365	RL3374		conserved hypothetical protein	0.17	318.89	1854.38	NE	NE	NE	NE	NE	3.79
RL_RS18275				0.23	48.29	206.87						-0.26
RL_RS20550	RL3984		putative oxidoreductase	0.30	252.20	854.38	NE	NE	NE	NE	NE	3.07
RL_RS23435	RL4549		putative 30S ribosomal protein S16	0.33	112.47	338.02	DE	ES	ES	ES	ES	0.03
RL_RS37435	RL2028		hypothetical exported protein	0.29	173.71	600.57						2.48
RL_RS40200			hypothetical protein	0.30	50.36	169.24						-0.25
RL_RS40230			hypothetical protein	0.23	825.84	3555.57						2.66
RL_RS27890	pRL100097		conserved hypothetical protein	0.28	355.65	1268.21	NE	NE	NE	NE	NE	7.18

Continued on next page

Table E.3 – continued from previous page

Gene Information				Mutant vs WT			INSeq					DE B
Gene ID	**Old ID**	**Dsg**	**Description**	**FC**	**m**	**WT**	**FL**	**Rh**	**R**	**NB**	**Bd**	**FC**
RL_RS27895	pRL100098		conserved hypothetical protein	0.17	333.85	2018.04	NE	NE	NE	NE	NE	7.31
RL_RS27900	pRL100099		conserved hypothetical protein	0.15	747.80	5121.07	NE	NE	NE	NE	NE	8.38
RL_RS39925	pRL100100		pseudogene	0.01	42.66	3352.55						11.07
RL_RS39930			hypothetical protein	0.01	37.59	4172.05						15.19
RL_RS27910	pRL100101		pseudogene, acetolactate synthase	0.19	1473.19	7651.00						9.36
RL_RS27915	pRL100103		putative hydroxyvalerate/ alcohol dehydrogenase	0.25	902.96	3614.20	ES	NE	NE	NE	NE	11.26
RL_RS27920	pRL100104	phaE	putative phaE protein, subunit of polyhydroxyalkanoate (PHA) synthase	0.30	918.62	3077.61	NE	NE	NE	NE	NE	11.07
RL_RS27925	pRL100105	phaC2	putative polyhydroxyalkanoate (PHA) synthase subunit C	0.29	658.37	2278.02	NE	NE	NE	NE	NE	9.89
RL_RS27975	pRL100119		putative propionate CoA-transferase	0.06	61.52	1011.77	NE	NE	NE	NE	NE	10.88

Continued on next page

Table E.3 – continued from previous page

Gene Information				Mutant vs WT			INSeq					DE B
Gene ID	**Old ID**	**Dsg**	**Description**	**FC**	**m**	**WT**	**FL**	**Rh**	**R**	**NB**	**Bd**	**FC**
RL_RS27980	pRL100120		phasin-like protein, often associated with PHB or other polyhydroxyalkanoate (PHA)	0.08	68.65	881.14	NE	NE	NE	NE	NE	9.37
RL_RS27985	pRL100121	acsA3	putative acetyl-coA synthetase	0.11	60.57	566.16	NE	NE	NE	NE	NE	8.95
RL_RS27990	pRL100122	adk2	putative adenylate kinase	0.25	56.11	222.75	NE	NE	NE	NE	NE	7.68
RL_RS28150	pRL100154		hypothetical protein	0.06	7.70	125.85	NE	NE	NE	NE	NE	7.43
RL_RS38405	pRL100156	fdxB1	putative ferredoxin	0.03	15.59	492.90	NE	DE	NE	NE	NE	9.29
RL_RS28165	pRL100158	nifN	putative nitrogenase iron-molybdenum cofactor biosynthesis protein NifN	0.01	21.24	1546.31	NE	NE	NE	DE	DE	10.16
RL_RS28170	pRL100159	nifE	putative nitrogenase iron-molybdenum cofactor biosynthesis protein NifE	0.01	16.98	2010.34	NE	NE	NE	NE	DE	10.32
RL_RS28175	pRL100160	nifK	nitrogenase molybdenum-iron protein beta chain NifK	0.01	26.54	4932.71	NE	NE	NE	ES	DE	11.18

Continued on next page

Table E.3 – continued from previous page

Gene Information				Mutant vs WT			INSeq					DE B
Gene ID	**Old ID**	**Dsg**	**Description**	**FC**	**m**	**WT**	**FL**	**Rh**	**R**	**NB**	**Bd**	**FC**
RL_RS28180	pRL100161	nifD	nitrogenase molybdenum-iron protein alpha chain NifD	0.00	16.38	4373.58	NE	NE	NE	ES	DE	12.12
RL_RS28185	pRL100162	nifH	nitrogenase iron protein NifH	0.00	10.39	5013.73	NE	NE	NE	NE	DE	12.69
RL_RS28335	pRL100192		hypothetical protein	0.30	163.09	540.56	NE	NE	NE	NE	NE	9.02
RL_RS28340	pRL100193		conserved hypothetical protein	0.23	203.31	878.47	NE	NE	NE	NE	NE	8.40
RL_RS28350		nifT	putative nitrogen fixation protein NifT	0.16	282.24	1713.71						8.96
RL_RS38450			4Fe-4S binding protein	0.14	218.54	1513.79						8.92
RL_RS28355	pRL100195	nifB	FeMo cofactor biosynthesis protein NifB	0.16	212.87	1303.74	NE	NE	NE	NE	DE	7.82
RL_RS28365	pRL100197	fixX	ferredoxin-like protein FixX	0.03	50.66	1772.35	NE	NE	NE	NE	DE	9.05
RL_RS28370	pRL100198	fixC	nitrogen fixation protein FixC	0.02	39.10	2140.40	NE	NE	NE	NE	DE	9.75
RL_RS28375	pRL100199	fixB	electron transfer protein FixB	0.00	15.77	4280.14	NE	NE	NE	DE	DE	12.59
RL_RS28380	pRL100200	fixA	electron transfer protein FixA	0.00	14.90	4088.65	NE	NE	NE	DE	DE	14.57

Continued on next page

Table E.3 – continued from previous page

Gene Information				Mutant vs WT			INSeq					DE B
Gene ID	**Old ID**	**Dsg**	**Description**	**FC**	**m**	**WT**	**FL**	**Rh**	**R**	**NB**	**Bd**	**FC**
RL_RS28385	pRL100201		conserved hypothetical protein	0.02	26.02	1230.21	NE	AD	AD	NE	NE	9.78
RL_RS28390	pRL100202		conserved hypothetical protein	0.24	23.08	97.64	NE	NE	AD	NE	NE	7.91
Genes upregulated in pRL100212 mutant												
RL_RS00935	RL0177	phnG	putative phosphonate utilisation protein G	3.15	50.24	15.97	NE	NE	NE	AD	NE	0.67
RL_RS03775	RL0719	flaB	putative flagellin B	3.66	11.94	3.26	NE	NE	NE	NE	NE	-3.63
RL_RS05100	RL0980		putative metalloprotease M24 family	3.34	37.94	11.36	AD	NE	NE	NE	NE	5.39
RL_RS05110	RL0982		putative permease component of ABC transporter PAAT	3.14	49.84	15.89	AD	NE	NE	NE	NE	6.11
RL_RS05115	RL0983		putative solute binding-protein of ABC transporter PAAT	3.52	89.52	25.41	AD	NE	NE	NE	NE	6.55
RL_RS05120	RL0984		putative ATP-binding component of ABC transporter PAAT	3.01	28.15	9.34	AD	NE	NE	NE	NE	6.62

Continued on next page

Table E.3 – continued from previous page

Gene Information				Mutant vs WT			INSeq					DE B
Gene ID	**Old ID**	**Dsg**	**Description**	**FC**	**m**	**WT**	**FL**	**Rh**	**R**	**NB**	**Bd**	**FC**
RL_RS06030	RL1157		putative two-component sensor/regulator; transcriptional regulator	4.19	16.93	4.04	NE	NE	NE	NE	NE	2.92
RL_RS18280	RL3549	glnII	putative glutamine synthetase II	3.51	91.65	26.14	NE	NE	NE	DE	ES	-1.95
RL_RS23515	RL4565	glnK	putative nitrogen regulatory protein PII	3.26	38.48	11.81	NE	NE	NE	NE	NE	0.80
RL_RS27820	pRL100081	qatX6	putative solute binding-protein of ABC transporter QAT glycine betaine/L-proline transporter	3.51	7.87	2.24	NE	AD	AD	NE	NE	-2.40
RL_RS28255	pRL100175	nodO	putative nodulation protein NodO	4.09	43.08	10.54	NE	NE	NE	NE	NE	2.13
RL_RS28295	pRL100183	nodF	nodulation protein NodF acyl carrier protein (ACP) used in Nod factor synthesis	3.37	16.37	4.86	NE	NE	NE	DE	DE	1.69
RL_RS28305	pRL100185	nodA	nodulation protein NodA	4.45	17.50	3.93	NE	NE	NE	DE	DE	4.98

Continued on next page

Table E.3 – continued from previous page

Gene Information				Mutant vs WT			INSeq					DE B
Gene ID	**Old ID**	**Dsg**	**Description**	**FC**	**m**	**WT**	**FL**	**Rh**	**R**	**NB**	**Bd**	**FC**
RL_RS28310	pRL100187	nodC	N-acetylglucosaminyltransferase NodC	3.21	18.34	5.72	NE	NE	NE	DE	DE	3.48
RL_RS30055	pRL110059		conserved hypothetical protein	3.11	54.56	17.54	NE	NE	NE	NE	NE	0.04
RL_RS30805	pRL110205	hutH1	putative histidine ammonia-lyase	3.11	9.48	3.05	NE	NE	NE	NE	NE	0.71
RL_RS31815	pRL110407		conserved hypothetical protein	3.30	18.91	5.73	NE	NE	NE	NE	NE	-1.89
RL_RS31840	pRL110412	rhaT	putative ATP-binding:ATP-binding (ABC:ABC) component of ABC transporter CUT2 involved in competition for nodulation	3.91	7.51	1.92	NE	NE	NE	NE	NE	-0.21
RL_RS31845	pRL110413	rhaS	putative solute binding-protein of ABC transporter CUT2 involved in competition for nodulation	4.63	69.33	14.99	NE	NE	NE	NE	NE	0.12

Continued on next page

Table E.3 – continued from previous page

Gene Information				Mutant vs WT			INSeq					DE B
Gene ID	**Old ID**	**Dsg**	**Description**	**FC**	**m**	**WT**	**FL**	**Rh**	**R**	**NB**	**Bd**	**FC**
RL_RS31850	pRL110414	rhaR	putative DeoR family transcriptional regulator (repressor) (involved in competition for nodulation)	4.40	9.96	2.26	NE	NE	NE	NE	NE	0.45
RL_RS31855	pRL110415	rhaD	putative short-chain dehydroge-nase/oxidoreductase involved in competition for nodulation	6.70	22.53	3.36	NE	NE	NE	DE	DE	0.07
RL_RS37125			YqaE/Pmp3 family membrane protein	3.56	14.44	4.05						-0.32
RL_RS31865	pRL110416	rhaI	putative rhamnose isomerase involved in competition for nodulation	3.69	18.98	5.14	NE	NE	NE	NE	AD	0.42
Genes downregulated in pRL100212 mutant												
RL_RS03985	RL0760		conserved hypothetical protein	0.17	8.95	52.22	N/A	N/A	N/A	N/A	N/A	1.45
RL_RS05175	RL0995		putative tartrate dehydrogenase	0.27	1.90	7.07	NE	NE	NE	ES	DE	2.30

Continued on next page

Table E.3 – continued from previous page

Gene Information				Mutant vs WT			INSeq					DE B
Gene ID	**Old ID**	**Dsg**	**Description**	**FC**	**m**	**WT**	**FL**	**Rh**	**R**	**NB**	**Bd**	**FC**
RL_RS05750	RL1107		putative transmembrane protein YiaAB domain protein	0.32	7.79	24.31	NE	NE	NE	NE	NE	2.22
RL_RS06075	RL1166		putative ribonuclease-L-PSP family protein	0.32	2.52	7.96	NE	NE	NE	DE	NE	2.09
RL_RS09685	RL1860	phhA	putative phenylalanine-4-hydroxylase	0.18	48.51	267.18	NE	NE	NE	NE	NE	2.16
RL_RS15065	RL2912		conserved hypothetical exported protein	0.17	3.19	19.25	NE	NE	NE	NE	NE	3.04
RL_RS20080	RL3890		putative transmembrane protein	0.28	14.25	50.79	NE	NE	NE	NE	NE	2.01
RL_RS20085	RL3891		conserved hypothetical exported protein	0.27	49.66	186.58	NE	NE	NE	NE	NE	2.59
RL_RS22960	RL4454		putative solute binding-protein of ABC transporter POPT	0.16	2.29	14.78	NE	NE	NE	NE	NE	3.16
RL_RS22965	RL4455	speB2	putative agmatinase	0.10	3.51	34.54	NE	NE	NE	NE	NE	3.98

Continued on next page

Table E.3 – continued from previous page

Gene Information				Mutant vs WT			INSeq					DE B
Gene ID	**Old ID**	**Dsg**	**Description**	**FC**	**m**	**WT**	**FL**	**Rh**	**R**	**NB**	**Bd**	**FC**
RL_RS22970	RL4456		putative solute binding-protein of ABC transporter PAAT	0.13	1.86	14.27	NE	NE	NE	NE	NE	3.15
RL_RS22975	RL4457		putative permease component of ABC transporter PAAT	0.12	1.13	9.73	AD	NE	NE	NE	NE	3.82
RL_RS36140	pRL120647		putative transposon-related protein	0.33	7.98	23.93	NE	NE	NE	NE	NE	3.29
RL_RS28245	pRL100173	rhiD	putative outer-membrane immunogenic protein precursor RhiD	0.29	2.09	7.13	NE	NE	NE	NE	NE	2.55
RL_RS28900	pRL100300		putative short-chain dehydroge-nase/reductase	0.25	5.19	21.09	NE	NE	NE	NE	NE	-1.20
Genes upregulated in pRL120291 mutant												

Continued on next page

Table E.3 – continued from previous page

Gene Information				Mutant vs WT			INSeq					DE B
Gene ID	**Old ID**	**Dsg**	**Description**	**FC**	**m**	**WT**	**FL**	**Rh**	**R**	**NB**	**Bd**	**FC**
RL_RS34360	pRL120287		putative permease component of ABC transporter CUT2 (S. mel solute binding-protein homologue SMa0203 induced by galactose, L-arabinose, fucose)	40.45	10.93	0.27	NE	NE	NE	NE	NE	-0.17
RL_RS34365	pRL120288		putative hydrolase	56.34	59.74	1.06	NE	AD	AD	NE	NE	-0.28
RL_RS34370	pRL120289		putative choline dehydrogenase	674.95	249.84	0.37	NE	NE	NE	NE	NE	-1.09
RL_RS34375	pRL120290		conserved hypothetical exported protein	2781.81	1920.26	0.69	NE	NE	NE	NE	NE	-3.34
RL_RS36630	pRL120744		putative transmembrane protein	16.77	8.05	0.48	NE	NE	NE	NE	NE	-0.04
Genes downregulated in pRL120291 mutant												
RL_RS03985	RL0760		conserved hypothetical protein	0.33	17.33	52.22	N/A	N/A	N/A	N/A	N/A	1.45
Genes upregulated in pRL120397 mutant												
RL_RS02350	RL0451	cspA2	putative cold shock protein	3.14	136.53	43.49	NE	NE	NE	NE	NE	1.13

Continued on next page

Table E.3 – continued from previous page

Gene Information				Mutant vs WT			INSeq					DE B
Gene ID	**Old ID**	**Dsg**	**Description**	**FC**	**m**	**WT**	**FL**	**Rh**	**R**	**NB**	**Bd**	**FC**
RL_RS05100	RL0980		putative metalloprotease M24 family	4.14	47.03	11.36	AD	NE	NE	NE	NE	5.39
RL_RS05105	RL0981		putative permease component of ABC transporter PAAT	3.36	46.52	13.86	AD	NE	NE	NE	NE	6.51
RL_RS05110	RL0982		putative permease component of ABC transporter PAAT	3.50	55.65	15.89	AD	NE	NE	NE	NE	6.11
RL_RS05115	RL0983		putative solute binding-protein of ABC transporter PAAT	4.12	104.77	25.41	AD	NE	NE	NE	NE	6.55
RL_RS05120	RL0984		putative ATP-binding component of ABC transporter PAAT	3.40	31.79	9.34	AD	NE	NE	NE	NE	6.62
RL_RS23515	RL4565	glnK	putative nitrogen regulatory protein PII	3.22	38.02	11.81	NE	NE	NE	NE	NE	0.80
RL_RS34370	pRL120289		putative choline dehydrogenase	29.57	10.95	0.37	NE	NE	NE	NE	NE	-1.09
RL_RS34375	pRL120290		conserved hypothetical exported protein	103.18	71.23	0.69	NE	NE	NE	NE	NE	-3.34
RL_RS34815	pRL120380		conserved hypothetical protein	17.79	28.82	1.62	NE	NE	NE	NE	NE	-3.62

Continued on next page

Table E.3 – continued from previous page

Gene Information				Mutant vs WT			INSeq					DE B
Gene ID	**Old ID**	**Dsg**	**Description**	**FC**	**m**	**WT**	**FL**	**Rh**	**R**	**NB**	**Bd**	**FC**
RL_RS34820	pRL120381		putative ATP-binding component of ABC transporter CUT1	434.49	473.80	1.09	NE	NE	NE	NE	NE	-1.15
RL_RS34825	pRL120382		conserved hypothetical protein	609.52	743.93	1.22	NE	NE	NE	NE	NE	-0.97
RL_RS34830	pRL120383		putative permease component of ABC transporter CUT1	224.94	339.80	1.51	NE	NE	NE	NE	NE	0.20
RL_RS34835	pRL120384		putative permease component of ABC transporter CUT1	374.35	584.23	1.56	NE	NE	NE	NE	NE	0.41
RL_RS34840	pRL120385		putative solute binding-protein of ABC transporter CUT1	529.20	1842.39	3.48	NE	NE	NE	NE	NE	0.04
RL_RS34845	pRL120386	dhaLp12	putative dihydroxyacetone kinase	16.00	29.61	1.85	NE	NE	NE	NE	NE	-1.31
RL_RS34850	pRL120387	dhaKp12	putative dihydroxyacetone kinase	13.37	43.73	3.27	NE	NE	NE	NE	NE	-1.51
RL_RS34855	pRL120388		conserved hypothetical protein	71.15	81.86	1.15	NE	NE	NE	NE	NE	0.69

Continued on next page

Table E.3 – continued from previous page

Gene Information				Mutant vs WT			INSeq					DE B
Gene ID	**Old ID**	**Dsg**	**Description**	**FC**	**m**	**WT**	**FL**	**Rh**	**R**	**NB**	**Bd**	**FC**
RL_RS34860	pRL120389		putative GFO/IDH/MocA family oxidoreductase	76.56	156.26	2.04	NE	NE	NE	NE	NE	1.83
RL_RS34865	pRL120390		putative permease component of ABC transporter CUT2	97.25	98.27	1.01	NE	NE	NE	NE	NE	0.17
RL_RS34870	pRL120391		putative ATP-binding:ATP-binding (ABC:ABC) component of ABC transporter CUT2	157.29	151.07	0.96	NE	NE	NE	NE	NE	0.38
RL_RS34875	pRL120392		conserved hypothetical protein	116.69	225.31	1.93	NE	NE	NE	NE	NE	0.02
RL_RS34880	pRL120393		putative solute binding-protein of ABC transporter CUT2	340.04	455.85	1.34	NE	NE	NE	NE	NE	-0.10
RL_RS34885	pRL120394		conserved hypothetical protein	584.03	666.07	1.14	NE	NE	NE	NE	NE	-1.31
RL_RS34890	pRL120395		conserved hypothetical protein	217.27	113.03	0.52	NE	NE	NE	NE	NE	-0.54
RL_RS34895	pRL120396	cbbl1	putative ribulose bisphosphate carboxylase large chain 1	174.95	168.02	0.96	NE	NE	NE	NE	NE	-0.12

Continued on next page

Table E.3 – continued from previous page

Gene Information				Mutant vs WT			INSeq					DE B
Gene ID	**Old ID**	**Dsg**	**Description**	**FC**	**m**	**WT**	**FL**	**Rh**	**R**	**NB**	**Bd**	**FC**
Genes downregulated in pRL120397 mutant												
RL_RS03985	RL0760		conserved hypothetical protein	0.10	5.37	52.22	N/A	N/A	N/A	N/A	N/A	1.45
RL_RS03990	RL0761	aceA	putative isocitrate lyase	0.15	1.35	8.88	NE	NE	NE	NE	NE	1.13
RL_RS20075	RL3889		putative transmembrane protein	0.17	5.15	30.59	NE	AD	AD	NE	NE	1.03
RL_RS20080	RL3890		putative transmembrane protein	0.11	5.68	50.79	NE	NE	NE	NE	NE	2.01
RL_RS20085	RL3891		conserved hypothetical exported protein	0.10	17.73	186.58	NE	NE	NE	NE	NE	2.59
RL_RS22960	RL4454		putative solute binding-protein of ABC transporter POPT	0.14	2.11	14.78	NE	NE	NE	NE	NE	3.16
RL_RS22965	RL4455	speB2	putative agmatinase	0.10	3.56	34.54	NE	NE	NE	NE	NE	3.98
RL_RS22970	RL4456		putative solute binding-protein of ABC transporter PAAT	0.11	1.55	14.27	NE	NE	NE	NE	NE	3.15

Continued on next page

Table E.3 – continued from previous page

Gene Information				Mutant vs WT			INSeq					DE B
Gene ID	**Old ID**	**Dsg**	**Description**	**FC**	**m**	**WT**	**FL**	**Rh**	**R**	**NB**	**Bd**	**FC**
RL_RS22975	RL4457		putative permease component of ABC transporter PAAT	0.12	1.21	9.73	AD	NE	NE	NE	NE	3.82
RL_RS24230	RL4709		putative component of MFS transporter, homology with shikimate transporter	0.27	21.04	76.75	AD	NE	NE	NE	NE	0.39
Genes upregulated in pRL120694 mutant												
RL_RS36340	pRL120685		putative solute binding-protein of ABC transporter CUT1	113.76	319.79	2.81	NE	NE	NE	NE	NE	0.38
RL_RS36345	pRL120686		conserved hypothetical protein	95.55	95.59	1.00	NE	NE	NE	NE	NE	-1.05
RL_RS36350	pRL120687		putative unsaturated glucuronyl hydrolase catalyses the hydrolytic release of unsaturated glucuronic acids from oligosaccharides	131.60	96.11	0.73	NE	NE	NE	NE	NE	-1.20
RL_RS36355	pRL120688		putative non-secreted polygalacturonase (degrades pectin found in plant cell walls)	269.32	51.19	0.19	NE	NE	NE	NE	NE	-0.72

Continued on next page

Table E.3 – continued from previous page

Gene Information				Mutant vs WT			INSeq					DE B
Gene ID	**Old ID**	**Dsg**	**Description**	**FC**	**m**	**WT**	**FL**	**Rh**	**R**	**NB**	**Bd**	**FC**
RL_RS36360	pRL120689		putative ATP-binding component of ABC transporter CUT1	75.49	34.74	0.46	NE	NE	NE	NE	NE	-0.68
RL_RS36365	pRL120690		putative permease component of ABC transporter CUT1	28.67	21.22	0.74	AD	NE	NE	NE	NE	0.05
RL_RS36370	pRL120691		putative permease component of ABC transporter CUT1	33.17	34.84	1.05	AD	NE	NE	NE	NE	-0.39
Genes downregulated in pRL120694 mutant												
RL_RS06075	RL1166		putative ribonuclease-L-PSP family protein	0.27	2.13	7.96	NE	NE	NE	DE	NE	2.09
Genes upregulated in pRL120695 mutant												
RL_RS05155	RL0991	matB	putative malonyl-CoA synthetase	3.05	21.50	7.06	NE	NE	NE	NE	NE	0.99
RL_RS34365	pRL120288		putative hydrolase	18.39	26.50	1.44	NE	AD	AD	NE	NE	-0.28
RL_RS34370	pRL120289		putative choline dehydrogenase	438.38	140.32	0.32	NE	NE	NE	NE	NE	-1.09
RL_RS34375	pRL120290		conserved hypothetical exported protein	17841.75	1427.72	0.08	NE	NE	NE	NE	NE	-3.34
RL_RS34380	pRL120291		putative MarR family transcriptional regulator	7.85	76.56	9.75	NE	NE	NE	DE	DE	-1.36

Continued on next page

Table E.3 – continued from previous page

Gene Information				Mutant vs WT			INSeq					DE B
Gene ID	**Old ID**	**Dsg**	**Description**	**FC**	**m**	**WT**	**FL**	**Rh**	**R**	**NB**	**Bd**	**FC**
Genes downregulated in pRL120695 mutant												
RL_RS00805	RL0150		conserved hypothetical protein (homologous with SMc02277)	0.24	3.92	16.01	NE	NE	NE	NE	NE	-0.56
RL_RS00940	RL0178		putative GntR family transcriptional regulator	0.31	8.12	26.12	NE	NE	NE	AD	NE	0.03
RL_RS01685	RL0323		putative methyltransferase	0.33	4.44	13.37	NE	NE	NE	NE	NE	-0.98
RL_RS01845	RL0353		putative transmembrane transporter protein	0.26	2.09	8.10	NE	NE	NE	NE	NE	1.10
RL_RS01850	RL0354		putative GntR family transcriptional regulator	0.20	4.75	23.73	NE	NE	NE	NE	NE	1.96
RL_RS02155	RL0413		putative MutT/Nudix family ADP compounds hydrolase	0.30	2.58	8.51	NE	NE	NE	NE	NE	0.58
RL_RS02270	RL0436	dapE	putative succinyl-diaminopimelate desuccinylase	0.33	3.98	11.98	DE	ES	ES	ES	ES	-1.63

Continued on next page

Table E.3 – continued from previous page

Gene Information				Mutant vs WT			INSeq					DE B
Gene ID	**Old ID**	**Dsg**	**Description**	**FC**	**m**	**WT**	**FL**	**Rh**	**R**	**NB**	**Bd**	**FC**
RL_RS02275	RL0437	dapD	putative 2,3,4,5-tetrahydropyridine-2,6-dicarboxylate N-succinyltransferase	0.32	3.86	12.03	DE	ES	ES	ES	ES	-2.19
RL_RS03120	RL0595	cspA1	putative cold shock protein	0.22	195.17	872.26	NE	NE	NE	NE	NE	0.59
RL_RS03195	RL0610		hypothetical exported protein	0.31	3.26	10.56	NE	NE	NE	NE	NE	-0.17
RL_RS03340			hypothetical protein	0.19	2.12	11.35						0.06
RL_RS03535	RL0671		putative transmembrane protein	0.32	6.72	20.94	NE	NE	NE	NE	NE	1.59
RL_RS04035	RL0770		putative phasin, phasin-2 superfamily	0.33	31.85	97.80	AD	NE	NE	NE	NE	-2.07
RL_RS05745	RL1106	pspA	putative (phage shock protein A) PspA family regulator by protein-protein interactions	0.23	15.09	65.09	NE	NE	NE	NE	NE	3.07
RL_RS05750	RL1107		putative transmembrane protein YiaAB domain protein	0.18	7.20	39.12	NE	NE	NE	NE	NE	2.22

Continued on next page

Table E.3 – continued from previous page

Gene Information				Mutant vs WT			INSeq					DE B
Gene ID	**Old ID**	**Dsg**	**Description**	**FC**	**m**	**WT**	**FL**	**Rh**	**R**	**NB**	**Bd**	**FC**
RL_RS06015	RL1154		conserved hypothetical exported protein	0.16	1.80	11.40	NE	NE	NE	NE	NE	4.27
RL_RS06030	RL1157		putative two-component sensor/regulator; transcriptional regulator	0.26	3.52	13.42	NE	NE	NE	NE	NE	2.92
RL_RS06790	RL1302	osmC2	putative OsmC-like organic hydroperoxide resistance protein	0.28	4.86	17.24	NE	NE	NE	NE	NE	-0.03
RL_RS07065	RL1356		putative transmembrane protein	0.23	3.53	15.62	NE	NE	NE	NE	NE	1.32
RL_RS07285	RL1400		putative transmembrane fatty acid desaturase	0.31	9.99	32.08	NE	NE	NE	NE	NE	-0.28
RL_RS07530			hypothetical protein	0.30	87.83	292.26						2.07
RL_RS07550	RL1454		putative RND family efflux transporter (membrane fusion protein)	0.31	3.23	10.45	NE	NE	NE	NE	NE	1.28
RL_RS08110	RL1564	ksgA	putative dimethyladenosine transferase	0.31	2.39	7.77	NE	NE	NE	NE	NE	-1.31

Continued on next page

Table E.3 – continued from previous page

Gene Information				Mutant vs WT			INSeq					DE B
Gene ID	**Old ID**	**Dsg**	**Description**	**FC**	**m**	**WT**	**FL**	**Rh**	**R**	**NB**	**Bd**	**FC**
RL_RS09220	RL1769	rpsL	putative 30S ribosomal protein S12	0.26	23.62	90.54	ES	ES	ES	ES	ES	-0.92
RL_RS09225	RL1770	rpsG	putative 30S ribosomal protein S7	0.20	11.81	59.28	ES	ES	ES	ES	ES	-2.08
RL_RS09230	RL1771	fus	putative translation elongation factor G, EF-G	0.25	16.73	67.24	ES	ES	ES	ES	ES	-2.43
RL_RS09245	RL1774	rplC	putative 50S ribosomal protein L3	0.21	11.87	57.91	ES	ES	ES	ES	ES	-3.03
RL_RS09250	RL1775	rplD	putative 50S ribosomal protein L4	0.19	9.38	49.00	ES	ES	ES	ES	ES	-3.30
RL_RS09255	RL1776	rplW	putative 50S ribosomal protein L23	0.23	9.18	39.61	ES	ES	ES	ES	ES	-3.13
RL_RS09260	RL1777	rplB	putative 50S ribosomal protein L2	0.23	10.53	46.13	ES	ES	ES	ES	ES	-3.25
RL_RS09265	RL1778	rpsS	putative 30S ribosomal protein S19	0.23	12.07	52.94	ES	ES	ES	ES	ES	-3.13
RL_RS09270	RL1779	rplV	putative 50S ribosomal protein L22	0.24	18.49	77.32	ES	ES	ES	ES	ES	-3.20
RL_RS09275	RL1780	rpsC	putative 30S ribosomal protein S3	0.21	12.20	59.33	ES	ES	ES	ES	ES	-3.50
RL_RS09280	RL1781	rplP	putative 50S ribosomal protein L16	0.25	17.02	68.55	ES	ES	ES	ES	ES	-3.56
RL_RS09285	RL1782	rpmC	putative 50S ribosomal protein L29	0.29	25.19	88.13	ES	ES	ES	ES	ES	-3.41

Continued on next page

Table E.3 – continued from previous page

Gene Information				Mutant vs WT			INSeq					DE B
Gene ID	**Old ID**	**Dsg**	**Description**	**FC**	**m**	**WT**	**FL**	**Rh**	**R**	**NB**	**Bd**	**FC**
RL_RS09290	RL1783	rpsQ	putative 30S ribosomal protein S17	0.22	17.81	82.23	ES	NE	DE	ES	ES	-3.34
RL_RS09295	RL1784	rplN	putative 50S ribosomal protein L14	0.33	44.28	136.21	ES	ES	ES	ES	ES	-1.46
RL_RS09300	RL1785	rplX	putative 50S ribosomal protein L24	0.29	29.45	100.47	ES	ES	ES	ES	ES	-2.00
RL_RS09305	RL1786	rplE	putative 50S ribosomal protein L5	0.23	21.40	95.02	ES	ES	ES	ES	ES	-2.28
RL_RS09310	RL1787	rpsN	putative 30S ribosomal protein S14	0.27	21.77	81.72	ES	ES	ES	ES	ES	-2.99
RL_RS09330	RL1791	rpsE	putative 30S ribosomal protein S5	0.25	20.49	81.62	DE	ES	ES	ES	ES	-3.20
RL_RS09335	RL1792	rpmD	putative 50S ribosomal protein L30	0.20	15.47	79.24	DE	ES	ES	ES	ES	-3.07
RL_RS09340	RL1793	rplO	putative 50S ribosomal protein L15	0.20	17.22	88.02	DE	ES	ES	ES	ES	-3.21
RL_RS09360	RL1797	rpsK	putative 30S ribosomal protein S11	0.31	34.79	113.79	ES	ES	ES	DE	ES	-1.17
RL_RS10880	RL2097	hom1	putative homoserine dehydrogenase	0.33	4.54	13.79	NE	ES	ES	DE	ES	-1.79
RL_RS11335	RL2186		putative AAA-family ATPase protein	0.30	21.60	72.14	NE	NE	NE	NE	NE	0.93
RL_RS11810	RL2283		putative aminotransferase	0.28	2.27	8.05	NE	NE	NE	NE	NE	-1.30

Continued on next page

Table E.3 – continued from previous page

Gene Information				Mutant vs WT			INSeq					DE B
Gene ID	**Old ID**	**Dsg**	**Description**	**FC**	**m**	**WT**	**FL**	**Rh**	**R**	**NB**	**Bd**	**FC**
RL_RS12900	RL2504		putative acetyltransferase	0.21	3.95	19.03	NE	NE	NE	NE	NE	1.61
RL_RS12905	RL2505	deaD	putative cold-shock DEAD-box protein A (ATP-dependent RNA helicase DEAD)	0.26	10.56	41.10	NE	NE	NE	NE	NE	1.60
RL_RS12985	RL2521		putative IAA acetyltransferase	0.33	3.33	10.15	NE	NE	NE	NE	NE	-0.30
RL_RS13320	RL2584		conserved hypothetical protein	0.33	5.28	16.22	DE	ES	ES	ES	ES	-1.46
RL_RS15150	RL2929		hypothetical protein	0.33	3.64	10.92	NE	AD	AD	NE	NE	-0.77
RL_RS15825	RL3065		conserved hypothetical exported protein	0.30	18.98	62.78	NE	NE	NE	NE	NE	2.22
RL_RS15830	RL3066		putative transmembrane protein	0.21	11.72	55.33	NE	NE	NE	NE	NE	4.63
RL_RS16395	RL3181		conserved hypothetical protein	0.31	10.71	34.83	NE	NE	NE	NE	NE	1.26
RL_RS16520	RL3206		putative TetR family transcriptional regulator	0.27	2.62	9.80	NE	ES	ES	ES	DE	0.90

Continued on next page

Table E.3 – continued from previous page

Gene Information				Mutant vs WT			INSeq					DE B
Gene ID	**Old ID**	**Dsg**	**Description**	**FC**	**m**	**WT**	**FL**	**Rh**	**R**	**NB**	**Bd**	**FC**
RL_RS17635	RL3423	lpcC	putative lipopolysaccharide core biosynthesis protein	0.23	3.09	13.63	NE	ES	ES	ES	ES	-0.72
RL_RS17640	RL3424	dctA	putative C4-dicarboxylate transport protein	0.26	155.44	588.21	NE	NE	NE	NE	NE	1.43
RL_RS17805	RL3457		conserved hypothetical protein	0.33	5.53	16.91	NE	NE	NE	NE	NE	-0.13
RL_RS18315	RL3556		conserved hypothetical protein	0.20	1.74	8.82	NE	NE	NE	NE	NE	-0.70
RL_RS18530	RL3597		putative DEAD-box ATP-dependent RNA helicase protein	0.25	9.08	36.71	NE	DE	NE	DE	NE	0.24
RL_RS19045	RL3693		conserved hypothetical protein	0.24	2.87	11.86	NE	NE	NE	NE	NE	-2.86
RL_RS19055	RL3695		conserved hypothetical exported protein	0.17	2.31	13.83	NE	NE	NE	NE	NE	-3.14
RL_RS19060	RL3696	hmuP	hemin uptake protein	0.32	3.78	11.89	NE	NE	NE	NE	NE	-1.67

Continued on next page

Table E.3 – continued from previous page

Gene Information				Mutant vs WT			INSeq					DE B
Gene ID	**Old ID**	**Dsg**	**Description**	**FC**	**m**	**WT**	**FL**	**Rh**	**R**	**NB**	**Bd**	**FC**
RL_RS19110	RL3706		putative two-component sensor/regulator; histidine kinase, ECF15 gene organisation (histidine kinase, response regulator, asf, ecf)	0.31	2.28	7.26	DE	ES	ES	DE	ES	-1.33
RL_RS19360	RL3754		conserved hypothetical protein	0.19	1.38	7.10	DE	ES	NE	ES	ES	-3.85
RL_RS19510	RL3784		putative MFS family efflux transporter (membrane fusion protein)	0.24	9.80	40.71	AD	NE	NE	NE	NE	2.73
RL_RS19515	RL3785		putative MarR family transcriptional regulator of RND family efflux transporter	0.30	40.63	134.85	AD	NE	NE	NE	NE	5.09
RL_RS20470	RL3968		putative OmpA family peptidoglycan-associated lipoprotein precursor	0.33	13.46	41.26	ES	ES	ES	NE	NE	-2.83

Continued on next page

Table E.3 – continued from previous page

Gene Information				Mutant vs WT			INSeq					DE B
Gene ID	**Old ID**	**Dsg**	**Description**	**FC**	**m**	**WT**	**FL**	**Rh**	**R**	**NB**	**Bd**	**FC**
RL_RS21095	RL4088		putative transmembrane protein	0.32	6.82	21.46	NE	NE	NE	NE	NE	1.29
RL_RS21955	RL4260	nrdI	putative ribonucleotide reductase protein	0.29	3.10	10.81	NE	NE	NE	NE	NE	-0.56
RL_RS22360	RL4336		putative acetyltransferase	0.32	2.84	8.97	NE	NE	NE	ES	NE	-0.83
RL_RS22440	RL4351		putative CBS domain family transmembrane protein	0.29	2.42	8.46	NE	NE	NE	NE	NE	-0.75
RL_RS23605	RL4582	fbpC	putative ATP-binding component of ABC transporter Unclass ferric cations transporter	0.33	5.59	17.12	DE	ES	DE	DE	DE	-0.20
RL_RS23835	RL4628	ispA	putative geranyl-transtransferase	0.32	3.42	10.85	NE	NE	NE	NE	NE	-0.48
RlC26				0.26	1.93	7.43						1.26
RlC31				0.27	23.00	85.01						1.50
RlD01				0.33	2.33	7.10						1.68
RL_RS37090	RL1507		conserved hypothetical protein	0.28	282.32	1007.67	DE	DE	DE	NE	DE	3.29
RL_RS40190			hypothetical protein	0.31	85.14	276.65						0.52

Continued on next page

Table E.3 – continued from previous page

Gene Information				Mutant vs WT			INSeq					DE B
Gene ID	**Old ID**	**Dsg**	**Description**	**FC**	**m**	**WT**	**FL**	**Rh**	**R**	**NB**	**Bd**	**FC**
RL_RS36395	pRL120696		putative RND family efflux transporter (membrane fusion protein)	0.30	4.67	15.55	NE	NE	NE	DE	NE	1.37
RL_RS36480	pRL120713		putative solute binding-protein of ABC transporter FeCT	0.30	4.00	13.53	NE	NE	NE	NE	NE	-2.23
RL_RS26175	pRL90059	rmrA	putative MFS family efflux transporter (membrane fusion protein)	0.23	10.10	43.21	NE	NE	NE	NE	NE	3.81
RL_RS37120	pRL90066		putative transmembrane protein	0.21	2.32	10.96	NE	ES	NE	NE	NE	-1.30
RL_RS27810	pRL100079	qatV6	putative ATP-binding component of ABC transporter QAT glycine betaine/L-proline transporter	0.33	8.60	25.93	NE	AD	AD	NE	NE	-0.25
RL_RS40275			hypothetical protein	0.30	9.31	31.23						1.56
RL_RS28220	pRL100169	rhiA	rhizosphere-induced protein RhiA	0.12	6.59	54.07	NE	NE	NE	NE	NE	2.17

Continued on next page

Table E.3 – continued from previous page

Gene Information				Mutant vs WT			INSeq					DE B
Gene ID	**Old ID**	**Dsg**	**Description**	**FC**	**m**	**WT**	**FL**	**Rh**	**R**	**NB**	**Bd**	**FC**
RL_RS28225	pRL100170	rhiB	rhizosphere-induced protein RhiB	0.13	0.94	7.43	NE	NE	GD	NE	NE	2.68
RL_RS29425	pRL100404	mctP	putative transmembrane sodium-solute symporter	0.09	0.87	9.63	NE	NE	NE	NE	NE	-4.75
RL_RS29430	pRL100405		putative transmembrane protein	0.18	1.69	9.59	NE	NE	NE	NE	NE	-3.66
RL_RS24895	pRL70118		putative crown gall tumour-like protein	0.31	2.16	7.04	NE	AD	NE	NE	NE	-0.06
RL_RS25345	pRL80036		conserved hypothetical protein	0.19	5.47	29.39	NE	NE	NE	NE	NE	4.75
RL_RS25350	pRL80037		putative short-chain dehydrogenase	0.27	1.87	6.90	NE	NE	NE	NE	NE	3.03
RL_RS25365	pRL80040		putative LacI family transcriptional regulator (repressor)	0.32	2.17	6.87	NE	NE	NE	NE	NE	0.15
RL_RS25460	pRL80060		putative solute binding-protein of ABC transporter PAAT closest homol mimosine transporter Rhizobium sp. TAL1145	0.24	3.50	14.48	NE	AD	AD	NE	NE	4.02

Continued on next page

Table E.3 – continued from previous page

Gene Information				Mutant vs WT			INSeq					DE B
Gene ID	**Old ID**	**Dsg**	**Description**	**FC**	**m**	**WT**	**FL**	**Rh**	**R**	**NB**	**Bd**	**FC**
RL_RS25525	pRL80075		putative endoribonuclease L-PSP family protein	0.22	1.93	8.62	NE	NE	NE	NE	NE	2.46
RL_RS30170	pRL110082		putative TetR family transcriptional regulator	0.33	3.16	9.65	NE	NE	NE	NE	NE	-0.80
RL_RS31055	pRL110257A		conserved hypothetical protein	0.23	1.99	8.81	NE	NE	NE	NE	NE	-0.95
RL_RS31685			putative K(+)-transporting ATPase subunit F	0.32	5.04	15.66						-2.34
Genes upregulated in RL0149 mutant												
RL_RS05155	RL0991	matB	putative malonyl-CoA synthetase	3.18	22.44	7.06	NE	NE	NE	NE	NE	0.99
RL_RS29540	pRL100427		putative permease component of ABC transporter Unclass	3.35	6.23	1.86	NE	AD	AD	NE	NE	2.30
Genes downregulated in RL0149 mutant												
RL_RS28220	pRL100169	rhiA	rhizosphere-induced protein RhiA	0.19	10.41	54.07	NE	NE	NE	NE	NE	2.17
RL_RS28225	pRL100170	rhiB	rhizosphere-induced protein RhiB	0.13	0.94	7.43	NE	NE	GD	NE	NE	2.68

Continued on next page

Table E.3 – continued from previous page

Gene Information				Mutant vs WT			INSeq					DE B
Gene ID	**Old ID**	**Dsg**	**Description**	**FC**	**m**	**WT**	**FL**	**Rh**	**R**	**NB**	**Bd**	**FC**
RL_RS29425	pRL100404	mctP	putative transmembrane sodium-solute symporter	0.33	3.13	9.63	NE	NE	NE	NE	NE	-4.75
Genes upregulated in RL0652 mutant												
RL_RS03405	RL0646		putative aldolase	15.23	156.17	10.25	NE	NE	NE	NE	NE	-0.77
RL_RS03410	RL0647		putative GFO/IDH/MocA family oxidoreductase	16.77	119.27	7.11	NE	NE	NE	NE	NE	-0.26
RL_RS03415	RL0648		conserved hypothetical protein	16.35	109.55	6.70	NE	NE	NE	NE	NE	0.16
RL_RS03420	RL0649		putative PfkB family sugar kinase	16.24	47.43	2.92	NE	NE	NE	NE	NE	0.39
RL_RS03425	RL0650		putative sorbitol dehydrogenase	14.96	51.18	3.42	NE	NE	NE	NE	NE	0.19
Genes downregulated in RL0652 mutant												
RL_RS01850	RL0354		putative GntR family transcriptional regulator	0.21	4.99	23.73	NE	NE	NE	NE	NE	1.96
RL_RS02460	RL0469		putative exported surface protein	0.22	4.33	19.28	NE	NE	NE	NE	NE	1.47
RL_RS06685	RL1282		putative DeoR family transcriptional regulator (repressor)	0.31	3.52	11.23	NE	NE	NE	NE	NE	1.71

Continued on next page

Table E.3 – continued from previous page

Gene Information				Mutant vs WT			INSeq					DE B
Gene ID	**Old ID**	**Dsg**	**Description**	**FC**	**m**	**WT**	**FL**	**Rh**	**R**	**NB**	**Bd**	**FC**
RL_RS28220	pRL100169	rhiA	rhizosphere-induced protein RhiA	0.17	9.18	54.07	NE	NE	NE	NE	NE	2.17
RL_RS28225	pRL100170	rhiB	rhizosphere-induced protein RhiB	0.13	0.94	7.43	NE	NE	GD	NE	NE	2.68
RL_RS25365	pRL80040		putative LacI family transcriptional regulator (repressor)	0.33	2.26	6.87	NE	NE	NE	NE	NE	0.15
Genes upregulated in RL0727 mutant												
RL_RS03775	RL0719	flaB	putative flagellin B	3.63	11.83	3.26	NE	NE	NE	NE	NE	-3.63
RL_RS03820	RL0728	flgE1	putative flagellar hook protein FlgE	6.77	148.09	21.86	NE	AD	AD	DE	NE	-2.28
RL_RS03825	RL0729	flgK1	putative flagellar hook-associated protein FlgK	7.76	88.56	11.40	NE	NE	NE	NE	NE	-2.33
RL_RS03830	RL0730	flgL1	putative flagellar hook-associated protein FlgK	6.23	74.63	11.98	NE	NE	NE	NE	DE	-1.73
RL_RS03835	RL0731	flaF1	putative flagellar synthesis related protein FlaF	6.04	100.58	16.66	NE	NE	NE	DE	DE	-1.31

Continued on next page

Table E.3 – continued from previous page

Gene Information				Mutant vs WT			INSeq					DE B
Gene ID	**Old ID**	**Dsg**	**Description**	**FC**	**m**	**WT**	**FL**	**Rh**	**R**	**NB**	**Bd**	**FC**
RL_RS03840	RL0732	flbT	putative flagellar biosynthesis repressor protein, binds to and promotes degradation of flagellar mRNA (post-transcriptional regulation)	5.40	58.00	10.73	NE	NE	NE	DE	DE	-3.01
RL_RS03845	RL0733	flgD	putative flagellar hook formation protein FlgD	3.75	90.72	24.18	NE	NE	NE	DE	NE	-2.33
RL_RS07360	RL1416		putative myo-inositol degradation/rhizopine catabolism protein	3.12	23.69	7.59	NE	NE	NE	NE	NE	3.35
Genes downregulated in RL0727 mutant												
RL_RS22965	RL4455	speB2	putative agmatinase	0.26	8.96	34.54	NE	NE	NE	NE	NE	3.98
RL_RS22970	RL4456		putative solute binding-protein of ABC transporter PAAT	0.28	4.06	14.27	NE	NE	NE	NE	NE	3.15
Genes upregulated in RL2924 mutant												
RL_RS15130	RL2925	mutT	putative mutator MutT protein	5.69	105.78	18.59	NE	NE	NE	NE	NE	0.03
RL_RS15135	RL2926		putative adenylyl cyclase	49.08	227.78	4.64	NE	NE	NE	NE	NE	1.25

Continued on next page

Table E.3 – continued from previous page

Gene Information				Mutant vs WT			INSeq					DE B
Gene ID	**Old ID**	**Dsg**	**Description**	**FC**	**m**	**WT**	**FL**	**Rh**	**R**	**NB**	**Bd**	**FC**
RL_RS15140	RL2927	ohrB	putative organic hydroperoxide resistance protein	288.23	847.63	2.94	NE	NE	NE	NE	NE	-0.30
Genes downregulated in RL2924 mutant												
RL_RS21380	RL4146		putative MarR family transcriptional regulator	0.28	2.00	7.05	NE	NE	NE	NE	NE	-1.18
RL_RS28220	pRL100169	rhiA	rhizosphere-induced protein RhiA	0.20	10.78	54.07	NE	NE	NE	NE	NE	2.17
RL_RS28225	pRL100170	rhiB	rhizosphere-induced protein RhiB	0.13	0.94	7.43	NE	NE	GD	NE	NE	2.68
RL_RS29425	pRL100404	mctP	putative transmembrane sodium-solute symporter	0.29	2.82	9.63	NE	NE	NE	NE	NE	-4.75

List of the genes coexpressed differentially expressed in the studied mutants, compared to the WT bacteroid. We include those genes with a FC value higher than 3 (upregulated genes) or lower than 1/3 (downregulated genes) and that have an expression higher than the median expression in the mutant (upregulated genes) or in the WT (downregulated genes). "Dsg" indicates the gene name. The columns under the "Mutant vs WT" title indicate the expression values (TPM) of the gene in the mutant (m) and in the WT (WT), and the FC value (FC). The last five columns indicate the INSeq value for the genes in free-living conditions

(FL), the rhizosphere (Rh), the root (R), nodule bacteria (NB), and bacteroid (Bd), according to (Wheatley *et al.*, 2020); possible values for these columns are essential (ES), defective (DE), neutral (NE), and advantaged (AD).

Table E.4: Genes which are expressed in free-living conditions

Old Gene ID	Dsg	Description	W/O N	WN
Genes whose expression decreases in the bacteroid				
pRL100325	fhuA1	putative outer membrane siderophore receptor		X
pRL100327	fhuD	putative SBP of ABC transporter FeCT FhuD siderophore (vicibactin Fe3+ complex) transporter		X
pRL100404	mctP	putative transmembrane sodium-solute symporter	X	
pRL100405		putative transmembrane protein	X	
pRL110325		conserved hypothetical protein	X	
pRL110544	minE	putative cell division topological specificity factor	X	
pRL110545	minD	putative septum site-determining protein MinD	X	
pRL110546	minC	septum site-determining protein MinC		X
pRL110578		putative 4-oxalocrotonate tautomerase	X	
pRL110603		hypothetical protein	X	
pRL110605		conserved hypothetical protein	X	
pRL120091		putative transmembrane protein		X
pRL120313	vbsP	putative 4'-phosphopantetheinyl transferase	X	
pRL120314	vbsO	putative siderophore biosynthesis related protein	X	X
pRL120315	vbsS	putative non-ribosomal peptide synthetase	X	
pRL120316	vbsA	putative siderophore biosynthesis protein vicibactin	X	X
pRL120317	vbsD	putative MATE (Multidrug And Toxic compound Extrusion) family efflux transporter. Exports vicibactin?	X	X
pRL120318	vbsL	putative cystathionine beta-lyase	X	X
pRL120319	rpoI	putative RNA polymerase ECF sigma factor involved in iron uptake, ECF09 gene arrangment (no asf) induced in MA of iron limitation in Rlv3841		X
pRL120320	vbsC	putative acetylase	X	X

Continued on next page

Table E.4 – continued from previous page

Old Gene ID	Dsg	Description	W/O N	WN
pRL120322	fhuA2	outer membrane siderophore receptor precursor	X	X
pRL120713		putative SBP of ABC transporter FeCT	X	
pRL70019	hupA2	putative DNA-binding protein Hu-alpha	X	
pRL80010		hypothetical protein	X	
pRL90019	fixK	putative FNR/CRP family transcriptional regulator FixK	X	
pRL90025		putative FNR/CRP family transcriptional regulator	X	
pRL90069		conserved hypothetical exported protein	X	
pRL90303		putative outer membrane lipoprotein/ SBP of ABC transporter PAAT	X	
RL0036		putative two-component sensor/regulator; transcriptional regulator	X	
RL0076		conserved hypothetical protein	X	
RL0211	pilA	putative pilus subunit protein	X	
RL0264		conserved hypothetical protein	X	
RL0322		putative iron-regulated protein	X	
RL0334	dnaN	putative DNA polymerase III, beta chain	X	
RL0375	dnaA	putative chromosomal replication initiator protein	X	
RL0692	cheY2	putative chemotaxis protein	X	
RL0696	visN	putative LuxR/GerE family transcriptional regulator	X	
RL0718	flaA	putative flagellin A	X	
RL0719	flaB	putative flagellin B	X	
RL0727	rem	putative two-component sensor/regulator; transcriptional regulator	X	
RL0728	flgE1	putative flagellar hook protein FlgE	X	
RL0732	flbT	putative flagellar biosynthesis repressor protein, binds to and promotes degradation of flagellar mRNA (post-transcriptional regulation)	X	
RL0733	flgD	putative flagellar hook formation protein FlgD	X	
RL0745	aglE	putative SBP of ABC transporter CUT1 alpha-glucoside-transporter (S. mel SBP homologue SMc03061 (agl) induced by maltotriose, turanose)	X	

Continued on next page

Table E.4 – continued from previous page

Old Gene ID	Dsg	Description	W/O N	WN
RL0770		putative phasin, phasin-2 superfamily	X	
RL1016		putative D-alanyl-D-alanine carboxypeptidase (penicillin-binding protein)	X	
RL1260	cysN	putative sulfate adenylyltransferase subunit 1	X	
RL1312		conserved hypothetical protein	X	
RL1351		putative cation efflux system protein	X	
RL1371		putative transmembrane protein	X	
RL1552	rplI	putative 50S ribosomal protein L9	X	
RL1558	fabG1	putative 3-oxoacyl-[acyl-carrier-protein] reductase	X	
RL1589	ropB	putative outer membrane protein RopB	X	
RL1691	hupA1	putative DNA-binding protein HU	X	
RL1770	rpsG	putative 30S ribosomal protein S7	X	
RL1771	fus	putative translation elongation factor G, EF-G	X	
RL1773	rpsJ	putative 30S ribosomal protein S10	X	
RL1774	rplC	putative 50S ribosomal protein L3	X	
RL1775	rplD	putative 50S ribosomal protein L4	X	X
RL1776	rplW	putative 50S ribosomal protein L23	X	X
RL1777	rplB	putative 50S ribosomal protein L2	X	
RL1778	rpsS	putative 30S ribosomal protein S19	X	
RL1779	rplV	putative 50S ribosomal protein L22	X	
RL1780	rpsC	putative 30S ribosomal protein S3	X	
RL1781	rplP	putative 50S ribosomal protein L16	X	
RL1782	rpmC	putative 50S ribosomal protein L29	X	
RL1783	rpsQ	putative 30S ribosomal protein S17	X	
RL1786	rplE	putative 50S ribosomal protein L5	X	
RL1787	rpsN	putative 30S ribosomal protein S14	X	
RL1788	rpsH	putative 30S ribosomal protein S8	X	

Continued on next page

Table E.4 – continued from previous page

Old Gene ID	Dsg	Description	W/O N	WN
RL1789	rplF	putative 50S ribosomal protein L6	X	
RL1790	rplR	putative 50S ribosomal protein L18	X	
RL1791	rpsE	putative 30S ribosomal protein S5	X	
RL1792	rpmD	putative 50S ribosomal protein L30	X	X
RL1793	rplO	putative 50S ribosomal protein L15	X	
RL1799	rplQ	putative 50S ribosomal protein L17		X
RL1880		putative FNR/CRP family transcriptional regulator	X	
RL1948	gpt	putative xanthine-guanine phosphoribosyltransferase		X
RL2072		putative transmembrane protein	X	
RL2098		putative transmembrane protein	X	
RL2102	cspA5	putative cold shock protein	X	
RL2221	rpsB	putative 30S ribosomal protein S2	X	
RL2222	tsf	putative elongation factor Ts (EF-Ts)	X	
RL2432	cusF	putative periplasmic copper binding protein, CusF in E. coli involved in Cu and Ag resistance	X	
RL2433	copC	putative copper binding protein, CopC (homol with cytochrome C oxidase which contains Cu)	X	
RL2713		putative SBP of ABC transporter FeCT ferrisiderophore binding	X	
RL2775	ropA1	putative outer membrane porin protein RopA	X	
RL2776		putative AsnC family transcriptional regulator	X	
RL2916		putative siderophore interacting protein	X	
RL3024	plyB	putative polysaccharidase secreted via PrsDE Type I exporter, PlyB	X	
RL3085		conserved hypothetical exported protein	X	
RL3150		conserved hypothetical protein	X	
RL3169		putative hydrolase	X	
RL3205	ilvC	putative ketol-acid reductoisomerase	X	

Continued on next page

Table E.4 – continued from previous page

Old Gene ID	Dsg	Description	W/O N	WN
RL3234	lppE	putative lipoprotein LppE, ECF operon with ECF26 gene organisation (lpp, ecf, asf)	X	
RL3267		putative OmpA family outer membrane protein	X	
RL3298	ftsZ1	putative cell division protein FtsZ	X	
RL3299	ftsA	putative cell division protein FtsA	X	
RL3302	aqpZ	putative aquaporin Z (bacterial nodulin-like intrinsic protein)	X	
RL3305		putative UDP-N-acetylenolpyruvoylglucosamine reductase	X	
RL3331	metF	putative 5,10-methylenetetrahydrofolate reductase	X	
RL3332		putative methyltransferase transcriptional regulator protein, ArsR family	X	
RL3378	cinI	putative autoinducer synthesis protein, CinS is v small protein, translationally coupled	X	
RL3382		putative outer membrane protein related to bp26 antigen of Brucella	X	
RL3383		conserved hypothetical protein	X	
RL3410		putative GatB/Yqey family protein	X	
RL3522		putative cation transport protein		X
RL3663	pssO	putative polysaccharide transport outer membrane protein	X	
RL3691		conserved hypothetical protein	X	X
RL3693		conserved hypothetical protein	X	X
RL3694	tonB	putative TonB protein, possible frameshift near start or incorrect start site used in annotation		X
RL3695		conserved hypothetical exported protein	X	
RL3697	hmuS	putative haem iron transport component	X	X
RL3698	hmuT	putative SBP of ABC transporter FeCT hemin transporter	X	
RL3699	hmuU	putative permease component of ABC transporter FeCT hemin transporter		X

Continued on next page

Table E.4 – continued from previous page

Old Gene ID	Dsg	Description	W/O N	WN
RL3705	phyR	putative anti-anti-sigma factor; two-component sensor/regulator; transcriptional regulator, PhyR, ECF15 gene organisation (histidine kinase, response regulator, asf, ecf)	X	
RL3752	pssA	putative transferase involved in exopolysaccharide biosynthesis	X	
RL3754		conserved hypothetical protein	X	
RL3967		conserved hypothetical protein	X	
RL3968		putative OmpA family peptidoglycan-associated lipoprotein precursor	X	
RL4004		conserved hypothetical protein	X	
RL4005		conserved hypothetical protein	X	
RL4081	exoN	putative UTP–glucose-1-phosphate uridylyltransferase	X	
RL4297		putative foldase/peptidyl-prolyl cis-trans isomerase	X	
RL4407	atpD	putative ATP synthase beta chain		X
RL4562		conserved hypothetical exported protein	X	
RL4584	fbpA	putative SBP of ABC transporter Unclass ferric cations transporter	X	
RL4597		conserved hypothetical protein	X	
RL4621	phaA	putative acetyl-CoA acetyltransferase	X	X
RL4665		conserved hypothetical protein	X	
RL4707	leuB	putative 3-isopropylmalate dehydrogenase	X	
RL4715	asd	putative aspartate-semialdehyde dehydrogenase		X
RL4716		conserved hypothetical exported protein	X	
Genes whose expression increases in the bacteroid				
pRL100098		conserved hypothetical protein		X
pRL100105	phaC2	putative polyhydroxyalkanoate (PHA) synthase subunit C		X
pRL100111		hypothetical protein		X
pRL100112		putative dehalogenase/hydrolase		X

Continued on next page

Table E.4 – continued from previous page

Old Gene ID	Dsg	Description	W/O N	WN
pRL100120		phasin-like protein, often associated with PHB or other polyhydroxyalkanoate (PHA)		X
pRL100156	fdxB1	putative ferredoxin		X
pRL100207	fixP1	putative cytochrome oxidase subunit		X
pRL100210A	fixS1	putative component of cation pump FixS		X
pRL110422		conserved hypothetical protein, only found in Rhizobia and in a few soil organisms	X	
pRL90017	fixO2	putative transmembrane cytochrome c oxidase fixO protein		X
pRL90242		putative nitrilotriacetate monooxygenase		X
RL0012	gyrB1	putative DNA gyrase subunit B	X	X
RL0141	cycM	putative cytochrome c	X	
RL0151	dnaJ	putative chaperone protein	X	
RL0238		conserved hypothetical protein	X	
RL0384	hspH	putative small heat shock protein		X
RL0390	praR	putative XRE family transcriptional regulator, mutation leads to increased adhesion and competitveness	X	
RL0470		hypothetical exported protein	X	
RL0777	rirA	putative Rrf2 family transcriptional regulator (activator/repressor)	X	X
RL0858		putative Nudix family protein (phosphohydrolases)	X	
RL0902	kdtA	putative 3-deoxy-d-manno-octulosonic-acid transferase		X
RL0924	atpI	putative ATP synthase protein I	X	
RL0941		putative peptidase	X	X
RL0942		putative transmembrane protein	X	
RL1019	tldD	putative Zn-dependent peptidase U62 modulator of DNA gyrase activity	X	
RL1249		hypothetical protein	X	
RL1274		conserved hypothetical protein	X	

Continued on next page

Table E.4 – continued from previous page

Old Gene ID	Dsg	Description	W/O N	WN
RL1300	rpsU1	putative 30S ribosomal protein S21A	X	
RL1419		putative serine protease	X	X
RL1440	degP1	putative serine protease	X	
RL1741		putative ribonuclease	X	
RL1877		putative protease		X
RL1878		putative peptidoglycan binding protein		X
RL2046	tatA	putative Sec-independent protein translocase protein	X	
RL2047	tatB	putative Sec-independent protein translocase protein	X	X
RL2209	cysE1	putative serine acetyltransferase	X	
RL2212	clpS1	putative ATP-dependent Clp protease adaptor protein	X	
RL2213	clpA	putative ATP-dependent Clp protease ATP-binding subunit	X	
RL2234	gltA	putative citrate synthase	X	X
RL2530		putative nitrogen fixation protein		X
RL2531	folE	putative GTP cyclohydrolase I	X	X
RL2532	hisI	putative histidine biosynthesis bifunctional protein	X	
RL2577	sufE	putative protein involved in binding to SufS	X	
RL2578	sufS1	putative aminotransferase involved in iron-sulfur cluster biogenesis		X
RL2579	sufD	putative iron-sulfur cluster assembly protein	X	
RL2580	sufC	putative ATP-binding component of ABC transporter Unclass	X	X
RL2581		hypothetical protein	X	
RL2582	sufB	putative iron-sulfur cluster formation subunit protein	X	
RL2815	fabF2	putative 3-oxoacyl-[acyl-carrier-protein] synthase II	X	
RL2878		conserved hypothetical protein	X	
RL3048		conserved hypothetical protein		X
RL3214		putative transmembrane NADH dehydrogenase		X
RL3253	hflC	putative transmembrane serine protease	X	

Continued on next page

Table E.4 – continued from previous page

Old Gene ID	Dsg	Description	W/O N	WN
RL3254	hflK	putative transmembrane serine protease	X	
RL3374		conserved hypothetical protein	X	
RL3393		conserved hypothetical protein	X	
RL3449	cspA4	putative cold shock protein	X	
RL3484	petC	putative cytochrome c1 precursor	X	
RL3485	petB	putative cytochrome b	X	
RL3486	petA	putative ubiquinol-cytochrome c reductase iron-sulfur subunit	X	
RL3490		putative ATP-binding:permease (ABC:IMP) component of ABC transporter Export		X
RL3495		conserved hypothetical protein	X	
RL3516		conserved hypothetical protein		X
RL3530		putative transmembrane protein	X	X
RL3984		putative oxidoreductase	X	
RL4089	ibpA	putative heat shock protein A hspA	X	
RL4100		putative transmembrane protein		X
RL4130	rpsU2	putative 30S ribosomal protein S21	X	
RL4148		putative cobalamin synthesis protein	X	
RL4279	clpB	putative chaperone protein ClpB (heat-shock protein)	X	
RL4429	lpdA	putative dihydrolipoamide dehydrogenase	X	
RL4431		conserved hypothetical exported protein		X
RL4432		conserved hypothetical protein	X	X
RL4433	sucB	putative dihydrolipoyllysine-residue succinyltransferase component of 2-oxoglutarate dehydrogenase	X	X
RL4434		conserved hypothetical protein	X	
RL4435	sucA	putative 2-oxoglutarate dehydrogenase E1 component	X	X
RL4436	sucD	putative succinyl-CoA synthetase alpha chain	X	X

Continued on next page

Table E.4 – continued from previous page

Old Gene ID	Dsg	Description	W/O N	WN
RL4437		conserved hypothetical protein	X	X
RL4438	sucC	putative succinyl-CoA synthetase beta chain	X	X
RL4439	mdh	putative malate dehydrogenase	X	X
RL4442		hypothetical protein		X
RL4443	sdhB	putative succinate dehydrogenase iron-sulfur protein	X	X
RL4444	sdhA	putative succinate dehydrogenase flavoprotein subunit	X	X
RL4445	sdhD	putative succinate dehydrogenase hydrophobic membrane anchor protein	X	X
RL4446	sdhC	putative succinate dehydrogenase cytochrome b556 subunit	X	
RL4530		putative ATP-binding component of ABC transporter possibly involved in export of lipids from inner to outer membrane Export	X	
RL4532		putative BAX Inhibitor (BI-1), may interact with FtsH (from E. coli) 57% id with pRL110005	X	
RL4536	acnA	putative aconitate hydratase		X
Genes whose expression remains constant in the bacteroid				
pRL100451	rapA2	putative autoaggregation protein	X	
pRL110304	treX	putative glycosylhydrolase		X
pRL110375	ropA3	putative outer membrane porin protein RopA	X	
pRL120001	repAp12	putative replication protein RepA		X
pRL120087		conserved hypothetical protein	X	
pRL120556		putative SBP of ABC transporter CUT1 (sucrose?)	X	
pRL70042		putative insertion sequence protein	X	
pRL90212	msrA	putative peptide methionine sulfoxide reductase MsrA	X	
RL0008		conserved hypothetical protein	X	X
RL0009	mltA	putative membrane-bound lytic murein transglycosylase A precursor		X
RL0040	coaA	putative pantothenate kinase	X	
RL0049	hslV	putative heat-shock ATP-dependent protease	X	

Continued on next page

Table E.4 – continued from previous page

Old Gene ID	Dsg	Description	W/O N	WN
RL0050		putative acetyltransferase	X	
RL0051	hslU	putative ATP-dependent heat shock protease component	X	
RL0106	rpsA	putative 30S ribosomal protein S1		X
RL0120	pnp	putative polynribonucleotide nucleotidyltransferase (polynucleotide phosphorylase) (PNPase)	X	
RL0121	rpsO	putative 30S ribosomal protein S15		X
RL0125	infB	putative translation initiation factor IF-2	X	
RL0127	nusA	putative transcription elongation protein (N utilization substance protein A) (L factor)	X	
RL0250		conserved hypothetical protein	X	
RL0266	infC	putative translation initiation factor		X
RL0389	metK	putative S-adenosylmethionine synthetase	X	
RL0392		putative hemolysin-like protein	X	
RL0416		putative transmembrane protein	X	
RL0417	ihfB	putative integration host factor beta subunit	X	
RL0424		putative sigma-54 modulation protein (has ribosome-associated inhibitory protein (raiA) domain)	X	
RL0546	phoU	putative phosphate uptake regulator PhoU, unknown mechanism to regulate expression of high-affinity ABC systems		X
RL0548	gcrA	cell cycle regulatory protein	X	
RL0550	argF	putative ornithine carbamoyltransferase	X	
RL0778	dppA	putative SBP of ABC transporter PepT dipeptide transporter (S. mel SBP homologue SMc00786 induced by CAS, hydroxyproline)	X	
RL0868		lipid A oxidase	X	
RL0883	groLch1	putative 60 kda chaperonin (protein cpn60)	X	X
RL0884	groSch1	putative 10 kda chaperonin (protein cpn10)	X	X

Continued on next page

Table E.4 – continued from previous page

Old Gene ID	Dsg	Description	W/O N	WN
RL0890		conserved hypothetical protein	X	
RL0892		putative ribosomal large subunit pseudouridine synthase B	X	
RL0927	atpG2	putative ATP synthase B' chain (subunit II)		X
RL1020		putative invasion associated protein	X	
RL1022	ctaD1	putative cytochrome c oxidase polypeptide I (cytochrome aa3 subunit 1)	X	
RL1023	ctaB	putative protoheme IX farnesyltransferase (heme o synthase)		X
RL1026	ctaE1	putative cytochrome c oxidase polypeptide III (cytochrome aa3 subunit 3)	X	
RL1083		putative disulfide bond formation protein D	X	
RL1366	uspF	putative universal stress protein	X	
RL1379	rosR	putative MucR/RosR family transcriptional regulator involved in nodulation competitiveness RosR	X	
RL1479		conserved hypothetical protein	X	
RL1504		conserved hypothetical protein	X	
RL1527		conserved hypothetical protein	X	
RL1555	rpsF	putative 30S ribosomal protein S6	X	X
RL1566		conserved hypothetical protein	X	
RL1697		hypothetical protein	X	
RL1700	nuoA	putative transmembrane NADH-quinone oxidoreductase subunit A	X	
RL1701	nuoB	putative NADH-quinone oxidoreductase subunit B	X	
RL1707	nuoG	putative NADH-quinone oxidoreductase subunit G	X	
RL1762	rplA	putative 50S ribosomal protein L1	X	X
RL1979	cysB	putative cysteine synthase	X	
RL2054	yajC	putative Sec translocase associated protein		X
RL2236	lexA	putative LexA repressor (SOS regulatory protein)	X	
RL2284	hfq	putative host factor protein	X	

Continued on next page

Table E.4 – continued from previous page

Old Gene ID	Dsg	Description	W/O N	WN
RL2289		conserved hypothetical protein		X
RL2290	cysI	putative sulfite reductase	X	
RL2424		putative acetyltransferase		X
RL2491		conserved hypothetical exported protein	X	
RL2512	secG	putative transmembrane protein SecG		X
RL2631	icd	putative isocitrate dehydrogenase [NADP]	X	
RL2688	nadE1	putative glutamine-dependent NAD(+) synthetase	X	
RL2697		conserved hypothetical exported protein	X	
RL2700		conserved hypothetical exported protein	X	
RL3256	thyA	putative thymidylate synthase	X	
RL3278		hypothetical protein	X	
RL3422	greA	putative transcription elongation factor	X	
RL3437		conserved hypothetical protein	X	
RL3446		putative quinoprotein glucose dehydrogenase-B	X	
RL3464		conserved hypothetical protein		X
RL3471		putative 50S ribosomal protein L25	X	
RL3478		conserved hypothetical protein	X	
RL3499		conserved hypothetical protein	X	
RL3513	leuA1	putative 2-isopropylmalate synthase	X	
RL3515	clpP3	putative ATP-dependent Clp protease	X	
RL3639		putative XRE family transcriptional regulator	X	
RL3704	asfZ	putative anti-sigma factor (no membrane-spanning domains), NepR-like AsfZ, ECF15 gene organisation (histidine kinase, response regulator, asf, ecf)		X

Continued on next page

Table E.4 – continued from previous page

Old Gene ID	Dsg	Description	W/O N	WN
RL3705	phyR	putative anti-anti-sigma factor; two-component sensor/regulator; transcriptional regulator, PhyR, ECF15 gene organisation (histidine kinase, response regulator, asf, ecf)		X
RL3706		putative two-component sensor/regulator; histidine kinase, ECF15 gene organisation (histidine kinase, response regulator, asf, ecf)		X
RL3775		putative RND family efflux transporter (membrane fusion protein)	X	
RL3776		putative lipase	X	
RL4007	gap	putative glyceraldehyde-3-phosphate dehydrogenase	X	X
RL4012		putative fructose-bisphosphate aldolase	X	
RL4093	cpdR	putative two-component sensor/regulator; transcriptional regulator	X	
RL4142		putative SBP of TRAP transporter	X	
RL4226	osmC	putative osmotically inducible protein C OsmC	X	
RL4288		putative amidohydrolase	X	
RL4289	grxC	putative glutaredoxin	X	
RL4298	secA	putative preprotein translocase subunit SecA	X	
RL4324		conserved hypothetical exported protein	X	
RL4350		putative DnaJ family chaperone	X	
RL4405	atpC	putative ATP synthase epsilon chain	X	
RL4407	atpD	putative ATP synthase beta chain		X
RL4408	atpG	putative ATP synthase gamma chain		X
RL4441		putative outer membrane lipoprotein omp19	X	
RL4529		putative lipase/arylesterase	X	
RL4549		putative 30S ribosomal protein S16	X	
RL4603		putative aminotransferase	X	
RL4614	rpoH2	putative RNA polymerase sigma factor (heat shock)	X	
RL4619		putative transmembrane protein	X	

Continued on next page

Table E.4 – continued from previous page

Old Gene ID	Dsg	Description	W/O N	WN
RL4622	phaR	putative carbon flux regulator, phaR in P. denitificans which regulates PHB synthesis		X
RL4638		putative pyruvate carboxylase	X	
RL4689		conserved hypothetical protein	X	
RL4692	ctpA	putative carboxy-terminal processing protease precursor	X	
RL4704		putative glyoxalase	X	
RL4725		putative heat-shock peptidase	X	

List of the genes analysed in Chapter 6. The genes are grouped depending on their expression pattern: genes whose expression decreases in the bacteroid, genes whose expression increases in the bacteroid, and genes whose expression remains constant in the bacteroid, compared to free-living conditions. "Dsg" indicates the gene name. The last two columns indicate whether the gene was retrieved without using the network ("W/O N") or using the network ("WN").