PLOS ONE

RESEARCH ARTICLE

Reconstruction and inference of the *Lactococcus lactis* MG1363 gene co-expression network

Jimmy Omony[1,2], **Anne de Jong**[1,2], **Jan Kok**[1,2]*, **Sacha A. F. T. van Hijum**[1,3]

1 Top Institute Food and Nutrition (TIFN), Wageningen, The Netherlands, **2** Department of Molecular Genetics, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Groningen, The Netherlands, **3** NIZO, Ede, The Netherlands

* Jan.Kok@rug.nl

Citation: Omony J, de Jong A, Kok J, van Hijum SAFT (2019) Reconstruction and inference of the *Lactococcus lactis* MG1363 gene co-expression network. PLoS ONE 14(5): e0214868. https://doi.org/10.1371/journal.pone.0214868

Editor: Arun K. Bhunia, Purdue University, UNITED STATES

Received: November 15, 2018

Accepted: March 21, 2019

Published: May 22, 2019

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: The research was funded by TI Food and Nutrition, a public–private partnership on precompetitive research in food and nutrition. The funding organization had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Lactic acid bacteria are Gram-positive bacteria used throughout the world in many industrial applications for their acidification, flavor and texture formation attributes. One of the species, *Lactococcus lactis*, is employed for the production of fermented milk products like cheese, buttermilk and quark. It ferments lactose to lactic acid and, thus, helps improve the shelf life of the products. Many physiological and transcriptome studies have been performed in *L. lactis* in order to comprehend and improve its biotechnological assets. Using large amounts of transcriptome data to understand and predict the behavior of biological processes in bacterial or other cell types is a complex task. Gene networks enable predicting gene behavior and function in the context of transcriptionally linked processes. We reconstruct and present the gene co-expression network (GCN) for the most widely studied *L. lactis* strain, MG1363, using publicly available transcriptome data. Several methods exist to generate and judge the quality of GCNs. Different reconstruction methods lead to networks with varying structural properties, consequently altering gene clusters. We compared the structural properties of the MG1363 GCNs generated by five methods, namely Pearson correlation, Spearman correlation, GeneNet, Weighted Gene Co-expression Network Analysis (WGCNA), and Sparse PArtial Correlation Estimation (SPACE). Using SPACE, we generated an *L. lactis* MG1363 GCN and assessed its quality using modularity and structural and biological criteria. The *L. lactis* MG1363 GCN has structural properties similar to those of the gold-standard networks of *Escherichia coli* K-12 and *Bacillus subtilis* 168. We showcase that the network can be used to mine for genes with similar expression profiles that are also generally linked to the same biological process.

Introduction

Lactococcus lactis MG1363 is a worldwide studied plasmid-free derivative of the dairy starter strain NCDO712 [1]. Several genomes of *L. lactis* strains, including MG1363, have been sequenced to completion [2–4] and many regulons of *L. lactis* MG1363 are well characterized

[5,6]. Still, the functions of many genes in its genome remain poorly understood. Reliable prediction and assignment of gene function remains a challenge deeply rooted in computational biological methods such as gene annotation and comparative genomics. Another option for gene prediction and function assignment is to construct gene co-expression networks (GCNs) [7–9]. A GCN is a graphical structure consisting of genes (depicted as nodes) and co-expression relationships, depicted as edges. The most connected nodes are the hubs, which generally correspond to genes encoding transcription factors (TFs) that drive the expression of the genes to which they are connected. Co-expression networks are used to characterize gene neighborhood relationships (commonly referred to as guilt-by-association) [10], which can be used to identify genes/proteins with similar functions and/or physical interactions [11]. A biologically meaningful network should be highly structurally organized, with clusters of genes (or modules) and genes connecting those clusters [12–15].

For reconstructing a GCN, Pearson or Spearman correlation coefficients are the most widely used measures of association to quantify gene co-expression [16,17]. Reconstruction of co-expression networks involves determining associations between genes based on their expression profiles. Studies on uncovering directional regulatory effects often focus on small-sized networks (with less than 200 genes). Several methods exist to infer activation and repression mechanisms in networks, but this is not the focus of our work here [11,18]. Inter- and intra-modular connections within a network complicate determining module boundaries [19]. Inter-modular connections are edges that connect genes belonging to different modules and intra-connections refer to edges that link genes within the same module. The presence of more connections within, rather than between, modules enables reliable module detection due to increased modularity (Q) of a network [20]. In addition to the Pearson or Spearman correlation approaches to reconstruct co-expression networks, other popular methods are GeneNet [21], SPACE [22], WGCNA [23] and ARACNE [24]. The choice of which of these methods to use for network reconstruction can be influenced by various factors, such as data size or whether one needs to infer regulatory directions between genes. For instance, Allen et al. [25] found that, for small networks consisting of less than 100 genes, GeneNet and SPACE out-perform the WGCNA and ARACNE approaches. Each network reconstruction method has its strengths and weaknesses [26,27]. Bayesian Network-based approaches like BNArray [28], B-course [29], Bayesian Network Toolbox [30] and Werhli's BN implementation [31] perform relatively poorly for large networks [25]. Data quality and dimension, network size and robustness of the used reconstruction method all affect the quality of the network [32], while lowly expressed genes are known to introduce bias and reduce network accuracy [33].

Here, we present the *L. lactis* MG1363 gene co-expression network based on data from the NCBI Gene Expression Omnibus (GEO) database [34] and discuss its structural properties in comparison to two gold-standard bacterial networks, namely those of *Bacillus subtilis* 168 and *Escherichia coli* K-12. We expect this *L. lactis* MG1363 GCN to provide an excellent basis for data mining and a template for designing further experiments. Such experiments would particularly be focused on sub-networks or on the functional analyses of specific genes of interest.

Methods

Transcriptome and regulon data sources

Transcriptome data used for the *L. lactis* MG1363 GCN reconstruction was obtained from the NCBI GEO database (http://www.ncbi.nlm.nih.gov/), Table A in S1 File. The GEO accession numbers of the 64 experiments used are given in the header of this table and give access to the detailed descriptions of the experiments. The raw data was Lowess-normalized and scaled as described in [35]. The resulting normalized signals were used for the network reconstruction.

To ensure a fair comparison between samples, experiments were grouped by (i) the growth medium used (GM17 or G-CDM, a rich and a chemically defined medium, respectively, containing 0.5% glucose), and (ii) the growth phase from which the samples were taken, namely early-, mid-, late-exponential or stationary phase, or based on ranges in culture optical density (OD). The processed data encompassed 64 conditions, after computing the median expression values per replicate and excluding datasets with genes with noisy expression. Downstream analysis of the data was performed using T-REx [36].

Biological network reconstruction

The *L. lactis* MG1363 networks were generated in R language v3.0.2. Network structural properties were analyzed using R's igraph package (version 1.0.0) and visualized in Cytoscape v3.2.0 [37]. Network density, modularity, average path length, diameter and number of detected modules were calculated for five methods, namely Pearson correlation, Spearman correlation, GeneNet, SPACE and WGCNA. We compared the networks thus generated and ranked them for performance. Networks generated using the Pearson or Spearman correlation coefficients were assessed by comparing the results of the degree distribution of the networks resulting from these two approaches to those of the power-law distribution [38,39]. To generate networks with the other three methods, the association parameters were varied. Only networks generated using specific regions of threshold parameters were considered for further analyses; hence, GCNs with (i) very high connectivity, (ii) low modularity, and (iii) very sparse connectivity (only a few hundred genes) were discarded. To examine the structural robustness of the *L. lactis* MG1363 network, a probabilistic random edge addition was performed using the approach described in [40]. A topological overlap matrix showing the degree to which directly linked nodes are connected was created to perform this analysis. In the WGCNA approach, we used a soft threshold approach on the adjacency matrix [23], which is a derivative of the topological matrix. Since the performance and reliability of network module detection methods are known to vary [25,41], we used at least four approaches to partition the networks, namely Walk-trap [42], Fast-Greedy [43], Infomap community [44] and label propagation [45].

Data for regulatory network reconstruction of *E. coli* K-12 were obtained from regulonDB (http://regulondb.ccg.unam.mx/menu/download/datasets/index.jsp) [46], those for the reconstruction of the *B. subtilis* 168 regulatory network from the SubtiWiki database (http://subtiwiki.uni-goettingen.de/) [47]. Gene-set enrichment analysis (GSEA) was performed using the Genome2D web-server (http://genome2d.molgenrug.nl/). The summarized workflow is presented in Fig 1.

Analysis of enriched network motifs

The detected network modules were subjected to DNA sequence motif enrichment analysis. We used MEME version 5.0.3 (http://meme-suite.org/) [48,49] for all network modules with at least four genes. Upstream regions of all genes within each module were extracted and used for the motif enrichment analysis (http://genome2d.molgenrug.nl/). The MEME search for motifs with a length between 6 and 16 bp was done on the upstream intergenic region, which are of variable length. Only the best motif of each cluster is reported—excluding the RBS motifs. Subsequently, the selected motifs we screened against the prokaryote TFBS database of PRODORIC Release 8.9 (using the TomTom Motif Comparison Tool (Version 5.0.4) of the MEME suite with default setting).

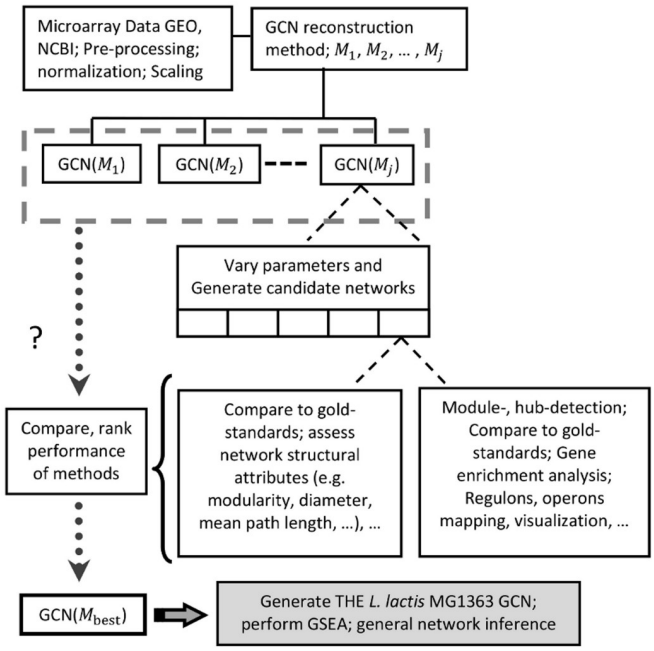

Fig 1. Workflow of gene co-expression network (GCN) reconstruction using different methods.

https://doi.org/10.1371/journal.pone.0214868.g001

Results and discussion

Construction of *L. lactis* MG1363 gene co-expression networks (GCNs)

Publicly available DNA microarray data on *L. lactis* MG1363 was used as input for gene network reconstruction. Prior to this, data exploratory analysis was performed to compare the distributions of the normalized mean and median expressed signals (Fig A, panel A in S1 File) and the density distributions of the Pearson correlation coefficients and Spearman correlation coefficients (Fig A panel B in S1 File). These two plots have similar density distributions except for a shift in the center of the measure of central tendency, namely the mean and median values. Overall, they provide an overview of the distribution of the correlation coefficients and indicate the quantity of high, medium and low correlation values. A summary of parameters resulting from the comparison of the structural properties of the *L. lactis* MG1363 co-expression networks to those of the gold-standard networks of *E. coli* K-12 [46] and *B. subtilis* 168 [47] is shown in Table B in S1 File.

We iteratively evaluated the performance of five network reconstruction methods, namely Pearson correlation, Spearman correlation, GeneNet, SPACE and WGCNA. For computing the adjacency matrix, we used a soft-power threshold value of 5. This value was determined based on the lowest power for which the scale-free model fits the data. Network structural properties such as the number of edges and the module sizes were compared to those of the gold-standard networks. For non-randomly connected biological networks, high modularity is a key indicator of high structural robustness (Fig B in S1 File) [50]. Modular GCNs have hubs in each module, which represent genes for TFs that are crucial for regulation of the genes in the network (Fig B in S1 File). Using the five network reconstruction methods, we searched for modular GCNs with a ratio between the number of edges to the number of genes (n_e/n_g) approximating those of the *E. coli* K-12 and *B. subtilis* 168 networks. Using this ratio criterion, we generated an *L. lactis* MG1363 GCN for each of the five methods. The number of lowly connected genes in the *L. lactis* MG1363 networks was marginally higher than those in the

gold standards. To obtain a high modularity (Q ≈ 0.80) in the networks, a stringent parameter threshold of $\boldsymbol{r} \approx 0.80$ was used for the Pearson correlation or Spearman correlation and the WGCNA. A lower threshold parameter ($\rho \approx 0.70$) was required for SPACE (Fig B panel C in S1 File, see also equation A3 in S2 File) and WGCNA (Fig B panel D in S1 File) to prevent a significant reduction of genes and edges in the network, which would result in a very sparse network. More on modularity and community structures in networks can be found in the work of Newman [51]. A further analysis shows that SPACE and WGCNA yield less dense and less modular networks than those generated by Pearson or Spearman correlation (Table C in S1 File). This is deduced from the ratio $\boldsymbol{n_e/n_g}$ and from the network modularity Q [41,52]. SPACE yielded networks with modules of various sizes and on the lower bound the networks had on average a value of $\boldsymbol{n_e/n_g}$ of about 6.5 (Table C in S1 File). This is a decent value since many connected genes in a network do not have a regulatory function and most TFs only regulate a few genes [53]. By considering networks corresponding to the plots in Fig B in S1 File and using only networks of approximately the same size (about the same number of genes and edges), the resultant GCNs from using the Pearson correlation coefficients or Spearman correlation coefficients were more densely connected and less modular than those obtained from the other three methods, especially for max(Q) ≈ 0.50 (Fig B in S1 File). Previous studies have shown that using different network reconstruction methods on the same dataset may yield varying network structures [21,22]. In our case, using the Pearson correlation coefficient or Spearman correlation coefficient of 0.90 leads to a near scale-free behavior of the obtained networks (Fig B panel C in S1 File Fig).

The structural connectivity of the *L. lactis* MG1363 GCN generated using SPACE (Fig 2) was fitted with the power law distribution model (Supplementary Material, S2 File). This model did not support the GCNs obtained using Pearson correlation or Spearman correlation (Figs C and D in S1 File). Using a less stringent threshold parameter results in a large and densely connected network Fig E in S1 File. The term ρ in Fig F in S1 File enables pruning of the adjacency matrix to remove spurious weak and non-significant edges between genes [54]. Smaller ρ values correspond to increased numbers of enriched gene classes, which is indicated by the total number of significantly enriched Gene Ontology (GO) terms (Fig F panels C and D in S1 File). In these plots, we observed a near-linear relationship with a curve that is similar to that observed between the values of ρ and ψ_1, where ψ_1 is the average number of GO terms per module with at least one significantly enriched GO term (Supplementary Material,

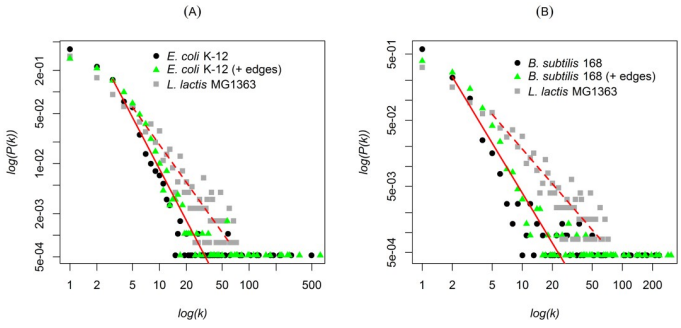

Fig 2. Bench-marking *L. lactis* MG1363 SPACE network to gold-standards. A: Degree distribution plot for the *E. coli* K-12 network (black circles). *E. coli* K-12 (+ edges) represents degree distributions of the network with random edge addition (green triangles). The *x*-axis shows the log-degree distribution (*k*); *y*-axis shows the log-probability of the degree distributions. B: The same as in A for the *B. subtilis* 168 and *B. subtilis* 168 (+ edges) plots. The criterion for edge addition is described in S2 File. The degree distribution of the *L. lactis* MG1363 network is plotted as grey squares in panels A and B. The red dotted lines show the power-law fit to the degree distributions of the *L. lactis* MG1363 network.

https://doi.org/10.1371/journal.pone.0214868.g002

S2 File). Overall, for the network inference, we used the five methods mentioned above to generate networks and subsequently compared and ranked their performances (Table D in S1 File). The results show that SPACE and the WGCNA performed best in the network reconstruction while GeneNet generated the least GO-enriched networks.

Network module detection

Modules were detected for all networks generated using the five network reconstruction methods. The plots for ρ versus ψ_2 (Fig F panels C and D in S1 File) show that larger ψ_2 values correspond to the region of the parameter $0.68 \leq \rho \leq 0.8$. Here $\psi_2$is the proportion of the total number of significant Fisher's exact tests (FETs) to the total number of modules with at least one significant GO term—irrespective of the significance of the *p*-value for the FET. The decision of which region in the plots corresponds to a good network is based on how large the ψ_1 and ψ_2 values on the vertical axis are and also on the total number of significantly enriched GO terms. This implies that only a specific choice of parameter values results in optimal enrichment of the gene sets in the modules (Fig G in S1 File) of an *L. lactis* MG1363 network. Therefore, we selected a range of parameter values and assessed them with respect to ability to yield good quality networks (shaded regions Fig B panels A to E in S1 File; Fig F panels B, D and F in S1 File). Unlike Walk-trap [42], Fast-Greedy [43] and the Infomap community [44] module detection methods, label propagation [45] shows a dip at $\rho \approx 0.7$ (Fig F panel C and D in S1 File)–which is indicative of a portioned network with only a few lowly enriched modules (low ψ_2 values) and is attributed to this particular method. Label propagation was relatively slow in partitioning the networks and did not yield modules with the most enriched gene sets. The networks with enriched modules that have the best partitioning were generated using the Walk-trap approach, which was our method of preference after the comparisons. We used it to detect modules in the *L. lactis* MG1363 GCNs because it was computationally faster and cheaper and yielded better results (Fig F panels C and D and Fig G in S1 File).

Structural properties: *L. lactis* MG1363 and gold-standard networks

To explore the structural differences between the *L. lactis* MG1363 networks and the gold-standard networks, random edges were simulated and added to the *E. coli* K-12 and *B. subtilis* 168 networks without altering their structural properties (Fig 2). We used the probabilistic random edge addition approach for the edge simulations [55] (S2 File). The *E. coli* K-12 and *B. subtilis* 168 networks were generated on the basis of literature-validated directed regulatory effects (TFs and their targets). These directed networks were represented as co-expression networks by ignoring the directional regulatory effects and only maintaining edges between genes. The addition of random edges to the gold-standard networks was aimed at explaining any differences in the degree distributions of the *E. coli* K-12 and *B. subtilis* 168 networks to that of the finally selected *L. lactis* MG1363 network, that obtained using SPACE. Fig 2 shows a comparison of the networks of all three organisms. Overall, both the *E. coli* K-12 and *B. subtilis* 168 networks are less densely connected than that of *L. lactis*, even after the addition of random edges (Fig 2A and 2B). The degree distribution plots for the *E. coli* K-12 (+ edges) and *B. subtilis* 168 (+ edges) networks both shift to the right towards the degree distribution line of the *L. lactis* MG1363 network, indicating that differences exist in certain regulatory mechanisms in the organisms. Both the *E. coli* and *B. subtilis* 168 networks show long-tailed distributions, revealing the presence of TFs such as sigma factors that regulate many targets (typically over 100 genes) [46]. A long tail was absent in the *L. lactis* MG1363 network (compare Fig 2A, 2B and Fig B in S1 File); the large sub-networks (regulons) of the pleiotropic regulators CodY and CcpA of *L. lactis* [6,56] do reside in the short tail.

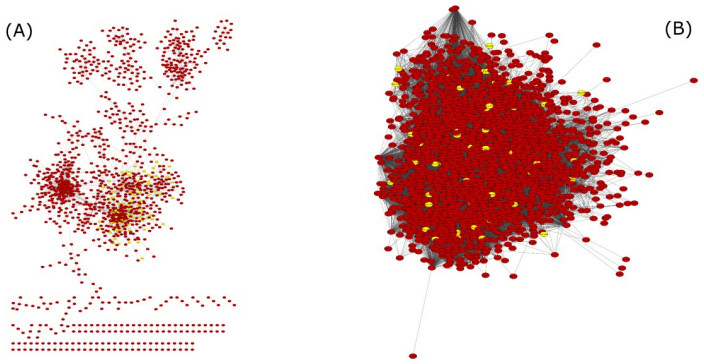

Fig 3. *L. lactis* MG1363 GCN visualized in Cytoscape v3.2.0. A: GCN generated using SPACE (ρ = 0.68). Projection of genes (shown in yellow) associated to significantly enriched GO groups in "module 0", other genes are colored red. The network consists of 1262 genes and 4112 edges. Only genes that satisfied the association threshold levels for inclusion in the adjacency matrix are shown in the network. B: Example network of *L. lactis* MG1363 generated using GeneNet (ω = 0.90; 2235 genes and 70386 edges). For instance, the GCN obtained using SPACE has enriched gene sets in "module 0", which are clustered together in the network (enriched gene sets in yellow), while the same genes are spread out in the GeneNet network.

https://doi.org/10.1371/journal.pone.0214868.g003

We chose the network reconstructed using SPACE and $\rho = 0.68$ as the most enriched and informative network for further analysis. Some of the network modules contained hubs, which were defined as genes connected to at least 5 other genes [19,57] (Figs H and I in S1 File). The value $\rho = 0.68$ is stringent but still not all the genes in the regulons that have been studied to date mapped in the GCN (Fig J in S1 File). Genes in the network were assigned to groups of the same ontology (biological processes, cellular components or molecular functions). Our final *L. lactis* MG1363 network generated using SPACE comprised of 94 modules, 16 of which contained significantly enriched gene sets (for various GO terms, Table E in S1 File). Only modules that had significantly enriched gene sets these were explored further. The 16 modules contained a varying number of genes with the smallest ones having only two genes and the largest 248 genes. The network is more modular than the one generated using GeneNet (Fig 3A and 3B; see also Fig F panel B in S1 File, which yielded the least modular networks from all the methods used for the network reconstruction).

Gene-set enrichment analysis also shows that SPACE generates the best *L. lactis* MG1363 network. All five network reconstruction methods were scrutinized for the gene-set enrichment in the network modules they generate (Fig F panels A to F in S1 File) in order to generate the *L. lactis* MG1363 GCN of choice. The selection was based on: (i) how closely the resulting network structure matched those of the gold-standards, and (ii) having biologically relevant (enriched) modules. The analyses probe whether modularity and scale-free behavior positively correlate to the biological enrichment of the gene sets in the modules of the GCNs. The number of enriched gene sets was compared for GCNs obtained using thresholds of different correlation parameter values and ρ. The results of the GSEA for GO terms on the *L. lactis* MG1363 network modules are provided in Table E in S1 File. Module detection in the GCNs obtained using Spearman correlation or WGCNA was performed using the Walk-trap method. Only a few modules were significantly enriched in the Spearman correlation network (low ψ_2 values in Fig F panels A and B in S1 File compared to those in Fig F panels C to F in S1 File). This indicates a trade-off in the relationship between network connectivity (densely, moderately and lowly connected) and enrichment for about the same number of genes. Densely connected networks further complicate GSEA since the boundaries between modules in such networks are fuzzy and difficult to detect, e.g. the low modularity (low Q values) for

the densely connected network resulting from GeneNet (Fig B panel E and Fig C panel B in S1 File). This can also be seen for the network in Fig F panel B in S1 File, which was generated using GeneNet. Low values of ψ_1 and ψ_2 indicate less enrichment of GO terms in the modules of the networks acquired with WGCNA and Spearman correlation than those obtained using SPACE (Fig F in S1 File). Densely connected GCNs with a low Q may have many enriched GO terms; however, the FETs shows that only a specific range of parameter values for the Spearman correlation coefficient r_S and ρ yield a good representation of significantly enriched networks (Fig F in S1 File). These results show that SPACE generates the most biologically enriched and structurally best network (Fig F panel A in S1 File).

We integrated and mapped known operons and regulons from literature onto the *L. lactis* MG1363 network reconstructed using SPACE. Thus, genes from 22 regulons were projected on the *L. lactis* MG1363 GCN to assess their distribution over the different modules. The results show that genes from the same operon and small regulons (e.g. PurR, HrcA and PyrR) often belong to the same GCN modules (Fig 4). Genes from larger regulons such as CcpA and CodY (Fig K in S1 File) were more broadly distributed over the network. A biological reason might be that genes in the same regulon might not always be co-expressed.

Enriched network motifs

Nineteen network modules showed evidence of overrepresented motifs (Table G in S1 File). Some genes in a module may be under the control of more than one regulator while a certain regulator may also control the activity of genes in multiple modules (eg, CodY, Fur and LuxR). Additionally, genes can be regulated by multiple other factors, e.g. small RNAs, RNA processing or via co-factor-riboswitch interaction, which could scatter the regulon over multiple

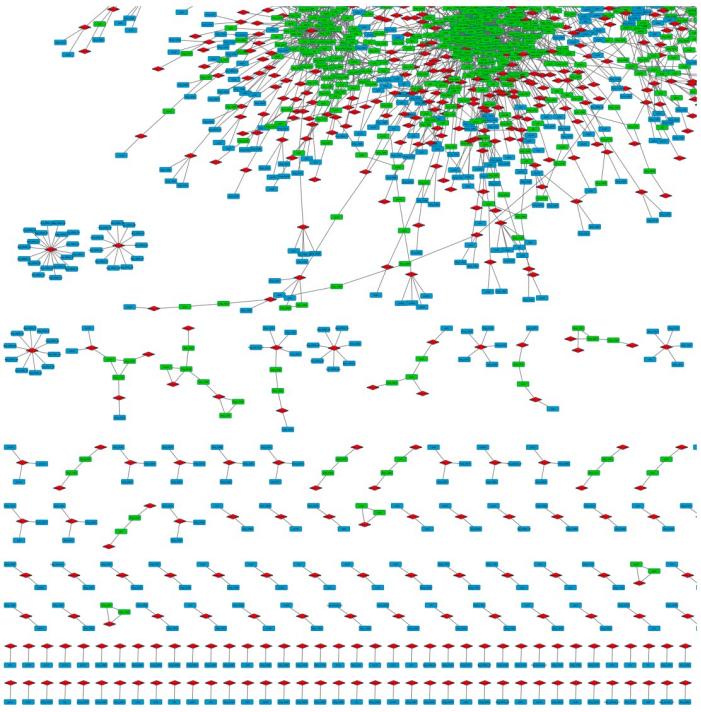

Fig 4. *L. lactis* MG1363 GCN integrated with literature-predicted operons (visualized in Cytoscape v3.2.0). The operon IDs are indicated in red, genes predicted to belong to operons are in green, and genes belonging to specific operons based on literature information (http://genome2d.molgenrug.nl) are shown in blue.

https://doi.org/10.1371/journal.pone.0214868.g004

Table 1. Enrichment of the most representative biological processes in the modules of the *L. lactis* MG1363 GCN.

Module	Members	Over represented function
Module 15	231	Transmembrane transport
Module 0	134	Regulation of transcription
Module 1	93	Carbohydrate metabolic process
Module 9	64	Amino acid transport
Module 2	57	Transmembrane transport
Module 7	25	Phosphoribosyltransferase-like
Module 13	23	General stress proteins
Module 27	14	Acyl-CoA N-acyltransferases
Module 33	11	DNA-binding HTH domain, TetR-type
Module 26	10	Universal stress proteins
Module 25	7	Universal stress proteins

https://doi.org/10.1371/journal.pone.0214868.t001

modules. Most TFs control the activity of one operon and conserved motifs can only be uncovered by searching the genomes of other organisms for the presence of orthologous DNA patterns. In addition to motifs of the global regulators CodY and CcpA those for more specific regulators such as CtrA, PerR and ArgR were also observed (Table G in S1 File).

Validation and use of the *L. lactis* MG1363 GCN generated using SPACE

To validate the biological relevance of the network modules detected in the *L. lactis* MG1363 GCN obtained with SPACE ($\rho = 0.68$), 19 network modules (Tables E and F in S1 File) with at least 5 genes per module were used as input for GSEA. Table 1 contains a summary of these modules and the corresponding overrepresented biological processes within each module. Module 0 and Module 1 are relatively large and predicted to fulfill the general functions transcription regulation and carbohydrate metabolism, respectively. We could associate hypothetical proteins to certain modules and predict their involvement in biological processes. For example, the InterPro IPR017853 protein domain (Glycoside hydrolase, super-family) is represented by 3 genes in Module 1. Two of the genes encode beta-glucosidases while one gene (*llmg_0186*) has no predicted function (Table F in S1 File) but has, apparently, the same expression behavior in many experiments. Indeed, the NCBI link for *llmg_0186* shows that this gene is likely in an operon with the gene for CelB (phosphotransferase system cellobiose-specific component IIC) and is probably involved in sugar (cellobiose) metabolism.

Conclusions

We have reconstructed and benchmarked the *L. lactis* MG1363 GCN using in-house and literature-derived transcriptome data. By analyzing the performance of five network reconstruction methods, namely Pearson correlation, Spearman correlation, WGCNA, GeneNet and SPACE, the latter was shown to yield the best network for *L. lactis* MG1363, both by looking at the structure of the network and at the biological content of the modules. The differences in network structure and corresponding parameters are attributed to the methods for computing the network adjacency matrices. Functional analyses demonstrated that the obtained network modules have biological relevance. Examination of the *L. lactis* MG1363 GCN shows that some regulons are not members of the same module, an indication that genes in such regulons are regulated by multiple transcription factors also in this organism. A list of differentially expressed genes obtained by DNA microarraying or RNA sequencing, or proteins acquired

through proteomics experiments, can be projected on the *L. lactis* MG1363 GCN in order to uncover gene/protein function.

Supporting information

S1 File. Fig A. Comparison of density distributions. Fig B. Comparison of network properties for different methods. Fig C. Model fit to network degree distribution of *L. lactis* MG1363 and the gold-standards. Fig D. Model fit to GCN degree distribution. Fig E. Correlation coefficients and network size. Fig F. Comparing gene-set enrichment of various *L. lactis* MG1363 networks. Fig G. Modules in *L. lactis* MG1363 network visualized in Cytoscape v3.2.0. Fig H. Hubs in the *L. lactis* MG1363 network. Fig I. Summary statistics of hubs in the *L. lactis* MG1363 network. Fig J. Annotated genes in major regulons in *L. lactis*. Fig K. Distribution of genes in regulons over the *L. lactis* MG1363 network. Table A. Curated data used in the network reconstruction. Table B. Comparison of *L. lactis* MG1363 GCN structural properties to gold-standards. Table C. Overview of performance measures of GCN reconstruction approaches. Table D. Ranking of performance of methods used to reconstruct *L. lactis* MG1363 GCN. Table E. Gene set enrichment of the *L. lactis* MG1363 network modules. Table F. Summary of the top-hit results from the GSEA of two large modules in the *L. lactis* MG1363 GCN. Table G. Overrepresented DNA sequence motifs in network modules.
(DOCX)

S2 File. Supporting methods and supporting results.
(DOCX)

Author Contributions

Conceptualization: Jimmy Omony, Anne de Jong, Jan Kok, Sacha A. F. T. van Hijum.

Data curation: Jimmy Omony, Sacha A. F. T. van Hijum.

Formal analysis: Jimmy Omony.

Funding acquisition: Jan Kok.

Investigation: Jimmy Omony, Anne de Jong, Sacha A. F. T. van Hijum.

Methodology: Jimmy Omony, Anne de Jong, Sacha A. F. T. van Hijum.

Project administration: Jan Kok, Sacha A. F. T. van Hijum.

Resources: Jan Kok, Sacha A. F. T. van Hijum.

Software: Jimmy Omony, Anne de Jong.

Supervision: Jan Kok, Sacha A. F. T. van Hijum.

Validation: Jimmy Omony, Anne de Jong, Sacha A. F. T. van Hijum.

Visualization: Jimmy Omony.

Writing – original draft: Jimmy Omony, Sacha A. F. T. van Hijum.

Writing – review & editing: Anne de Jong, Jan Kok, Sacha A. F. T. van Hijum.

References

1. Gasson MJ. Plasmid complements of Streptococcus lactis NCDO712 and other lactic streptococci after protoplast-induced curing. J Bacteriol. 1983; 154: 1–9. PMID: 6403500

2. Bolotin A, Wincker P, Mauger S, Jaillon O, Malarme K, Weissenbach J, et al. The complete genome sequence of the lactic acid bacterium Lactococcus lactis ssp. lactis IL1403. Genome Res. 2001; 11: 731–753. https://doi.org/10.1101/gr.169701 PMID: 11337471
3. Wegmann U, O'Connell-Motherway M, Zomer A, Buist G, Shearman C, Canchaya C, et al. Complete genome sequence of the prototype lactic acid bacterium Lactococcus lactis subsp. cremoris MG1363. J Bacteriol. 2007; 189: 3256–3270. https://doi.org/10.1128/JB.01768-06 PMID: 17307855
4. Siezen RJ, Bayjanov J, Renckens B, Wels M, van Hijum SA, Molenaar D, et al. Complete genome sequence of Lactococcus lactis subsp. lactis KF147, a plant-associated lactic acid bacterium. J Bacteriol. 2010; 192: 2649–2650. https://doi.org/10.1128/JB.00276-10 PMID: 20348266
5. Larsen R, van Hijum SA, Martinussen J, Kuipers OP, Kok J. Transcriptome analysis of the Lactococcus lactis ArgR and AhrC regulons. Appl Environ Microbiol. 2008; 74: 4768–4771. https://doi.org/10.1128/AEM.00117-08 PMID: 18539789
6. Zomer AL, Buist G, Larsen R, Kok J, Kuipers OP. Time-resolved determination of the CcpA regulon of *Lactococcus lactis* subsp. cremoris MG1363. J Bacteriol. 2007; 189: 1366–1381. https://doi.org/10.1128/JB.01013-06 PMID: 17028270
7. Jiang J, Sun X, Wu W, Li L, Wu H, Zhang L, et al. Construction and application of a co-expression network in Mycobacterium tuberculosis. Sci Rep. 2016; 6: 28422. https://doi.org/10.1038/srep28422 PMID: 27328747
8. Mao L, Van Hemert JL, Dash S, Dickerson JA. Arabidopsis gene co-expression network and its functional modules. BMC Bioinformatics. 2009; 10: 346-2105-10-346.
9. Shaik R, Ramakrishna W. Genes and co-expression modules common to drought and bacterial stress responses in Arabidopsis and rice. PLoS One. 2013; 8: e77261. https://doi.org/10.1371/journal.pone.0077261 PMID: 24130868
10. Aoki K, Ogata Y, Shibata D. Approaches for extracting practical information from gene co-expression networks in plant biology. Plant Cell Physiol. 2007; 48: 381–390. https://doi.org/10.1093/pcp/pcm013 PMID: 17251202
11. Spirin V, Mirny LA. Protein complexes and functional modules in molecular networks. Proc Natl Acad Sci U S A. 2003; 100: 12123–12128. https://doi.org/10.1073/pnas.2032324100 PMID: 14517352
12. Gutierrez-Rios RM, Rosenblueth DA, Loza JA, Huerta AM, Glasner JD, Blattner FR, et al. Regulatory network of Escherichia coli: consistency between literature knowledge and microarray profiles. Genome Res. 2003; 13: 2435–2443. https://doi.org/10.1101/gr.1387003 PMID: 14597655
13. Ma HW, Buer J, Zeng AP. Hierarchical structure and modules in the Escherichia coli transcriptional regulatory network revealed by a new top-down approach. BMC Bioinformatics. 2004; 5: 199. https://doi.org/10.1186/1471-2105-5-199 PMID: 15603590
14. Fu Y, Jarboe LR, Dickerson JA. Reconstructing genome-wide regulatory network of E. coli using transcriptome data and predicted transcription factor activities. BMC Bioinformatics. 2011; 12: 233-2105-12-233.
15. Omony J, de Jong A, Krawczyk AO, Eijlander RT, Kuipers OP. Dynamic sporulation gene co-expression networks for Bacillus subtilis 168 and the food-borne isolate Bacillus amyloliquefaciens: a transcriptomic model. Microb Genom. 2018.
16. Michalopoulos I, Pavlopoulos GA, Malatras A, Karelas A, Kostadima MA, Schneider R, et al. Human gene correlation analysis (HGCA): a tool for the identification of transcriptionally co-expressed genes. BMC Res Notes. 2012; 5: 265-0500-5-265.
17. Butte AJ, Tamayo P, Slonim D, Golub TR, Kohane IS. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. Proc Natl Acad Sci U S A. 2000; 97: 12182–12186. https://doi.org/10.1073/pnas.220392197 PMID: 11027309
18. Girvan M, Newman ME. Community structure in social and biological networks. Proc Natl Acad Sci U S A. 2002; 99: 7821–7826. https://doi.org/10.1073/pnas.122653799 PMID: 12060727
19. Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, et al. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature. 2004; 430: 88–93. https://doi.org/10.1038/nature02555 PMID: 15190252
20. Chen J, Yuan B. Detecting functional modules in the yeast protein-protein interaction network. Bioinformatics. 2006; 22: 2283–2290. https://doi.org/10.1093/bioinformatics/btl370 PMID: 16837529
21. Schafer J, Strimmer K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat Appl Genet Mol Biol. 2005; 4: Article32.
22. Peng J, Wang P, Zhou N, Zhu J. Partial Correlation Estimation by Joint Sparse Regression Models. J Am Stat Assoc. 2009; 104: 735–746. https://doi.org/10.1198/jasa.2009.0126 PMID: 19881892
23. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008; 9: 559. https://doi.org/10.1186/1471-2105-9-559 PMID: 19114008

24. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics. 2006; 7 Suppl 1: S7.
25. Allen JD, Xie Y, Chen M, Girard L, Xiao G. Comparing statistical methods for constructing large scale gene networks. PLoS One. 2012; 7: e29348. https://doi.org/10.1371/journal.pone.0029348 PMID: 22272232
26. Rice JJ, Tu Y, Stolovitzky G. Reconstructing biological networks using conditional correlation analysis. Bioinformatics. 2005; 21: 765–773. https://doi.org/10.1093/bioinformatics/bti064 PMID: 15486043
27. Wildenhain J, Crampin EJ. Reconstructing gene regulatory networks: from random to scale-free connectivity. Syst Biol (Stevenage). 2006; 153: 247–256.
28. Chen X, Chen M, Ning K. BNArray: an R package for constructing gene regulatory networks from microarray data by using Bayesian network. Bioinformatics. 2006; 22: 2952–2954. https://doi.org/10.1093/bioinformatics/btl491 PMID: 17005537
29. Myllymaki P, Silander T, Tirri H, Uronen P. B-course: A web-based tool for bayesian and causal data analysis. International Journal on Artificial Intelligence Tools. 2002; 11: 369–388.
30. Murphy K. The bayes net toolbox for matlab. Computing science and statistics. 2001; 33: 1024–1034.
31. Werhli AV, Grzegorczyk M, Husmeier D. Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. Bioinformatics. 2006; 22: 2523–2531. https://doi.org/10.1093/bioinformatics/btl391 PMID: 16844710
32. Dong J, Horvath S. Understanding network concepts in modules. BMC Syst Biol. 2007; 1: 24. https://doi.org/10.1186/1752-0509-1-24 PMID: 17547772
33. Yang EW, Girke T, Jiang T. Differential gene expression analysis using coexpression and RNA-Seq data. Bioinformatics. 2013; 29: 2153–2161. https://doi.org/10.1093/bioinformatics/btt363 PMID: 23793751
34. Barrett T, Edgar R. Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. Methods Enzymol. 2006; 411: 352–369. https://doi.org/10.1016/S0076-6879(06)11019-8 PMID: 16939800
35. van Hijum SA, de Jong A, Baerends RJ, Karsens HA, Kramer NE, Larsen R, et al. A generally applicable validation scheme for the assessment of factors involved in reproducibility and quality of DNA-microarray data. BMC Genomics. 2005; 6: 77. https://doi.org/10.1186/1471-2164-6-77 PMID: 15907200
36. de Jong A, van der Meulen S, Kuipers OP, Kok J. T-REx: Transcriptome analysis webserver for RNA-seq Expression data. BMC Genomics. 2015; 16: 663-015-1834-4.
37. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003; 13: 2498–2504. https://doi.org/10.1101/gr.1239303 PMID: 14597658
38. Stuart JM, Segal E, Koller D, Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. Science. 2003; 302: 249–255. https://doi.org/10.1126/science.1087447 PMID: 12934013
39. Albert R. Scale-free networks in cell biology. J Cell Sci. 2005; 118: 4947–4957. https://doi.org/10.1242/jcs.02714 PMID: 16254242
40. Deijfen M, Lindholm M. Growing networks with preferential addition and deletion of edges. Physics and Society. 2015;Physica A 388: 4297–4303.
41. Newman ME, Girvan M. Finding and evaluating community structure in networks. Phys Rev E Stat Nonlin Soft Matter Phys. 2004; 69: 026113. https://doi.org/10.1103/PhysRevE.69.026113 PMID: 14995526
42. Pons P, Matthieu Latapy M. Computing communities in large networks using random walks. Journal of Graph Algorithms and Applications. 2006; 10: 191–218.
43. Newman ME. Fast algorithm for detecting community structure in networks. Phys Rev E Stat Nonlin Soft Matter Phys. 2004; 69: 066133. https://doi.org/10.1103/PhysRevE.69.066133 PMID: 15244693
44. Rosvall M, Bergstrom CT. Maps of random walks on complex networks reveal community structure. Proc Natl Acad Sci U S A. 2008; 105: 1118–1123. https://doi.org/10.1073/pnas.0706851105 PMID: 18216267
45. Raghavan UN, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks. Phys Rev E Stat Nonlin Soft Matter Phys. 2007; 76: 036106. https://doi.org/10.1103/PhysRevE.76.036106 PMID: 17930305
46. Salgado H, Peralta-Gil M, Gama-Castro S, Santos-Zavaleta A, Muniz-Rascado L, Garcia-Sotelo JS, et al. RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. Nucleic Acids Res. 2013; 41: D203–13. https://doi.org/10.1093/nar/gks1201 PMID: 23203884

47. Michna RH, Commichau FM, Todter D, Zschiedrich CP, Stulke J. SubtiWiki-a database for the model organism Bacillus subtilis that links pathway, interaction and expression information. Nucleic Acids Res. 2014; 42: D692–8. https://doi.org/10.1093/nar/gkt1002 PMID: 24178028

48. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol. 1994; 2: 28–36. PMID: 7584402

49. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME Suite. Nucleic Acids Res. 2015; 43: W39–49. https://doi.org/10.1093/nar/gkv416 PMID: 25953851

50. Tran TD, Kwon YK. The relationship between modularity and robustness in signalling networks. J R Soc Interface. 2013; 10: 20130771. https://doi.org/10.1098/rsif.2013.0771 PMID: 24047877

51. Newman ME. Modularity and community structure in networks. Proc Natl Acad Sci U S A. 2006; 103: 8577–8582. https://doi.org/10.1073/pnas.0601602103 PMID: 16723398

52. Newman ME. Finding community structure in networks using the eigenvectors of matrices. Phys Rev E Stat Nonlin Soft Matter Phys. 2006; 74: 036104. https://doi.org/10.1103/PhysRevE.74.036104 PMID: 17025705

53. Sabatti C, James GM. Bayesian sparse hidden components analysis for transcription regulation networks. Bioinformatics. 2006; 22: 739–746. https://doi.org/10.1093/bioinformatics/btk017 PMID: 16368767

54. Gillis J, Pavlidis P. "Guilt by association" is the exception rather than the rule in gene networks. PLoS Comput Biol. 2012; 8: e1002444. https://doi.org/10.1371/journal.pcbi.1002444 PMID: 22479173

55. Barabasi AL, Albert R. Emergence of scaling in random networks. Science. 1999; 286: 509–512. PMID: 10521342

56. den Hengst CD, van Hijum SA, Geurts JM, Nauta A, Kok J, Kuipers OP. The *Lactococcus lactis* CodY regulon: identification of a conserved *cis*-regulatory element. J Biol Chem. 2005; 280: 34332–42. https://doi.org/10.1074/jbc.M502349200 PMID: 16040604

57. Lu X, Jain VV, Finn PW, Perkins DL. Hubs in biological interaction networks exhibit low changes in expression in experimental asthma. Mol Syst Biol. 2007; 3: 98. https://doi.org/10.1038/msb4100138 PMID: 17437023