The structure of a gene co-expression network reveals biological functions underlying eQTLs

Nathalie Villa-Vialaneix[1,*], Laurence Liaubet[2], Thibault Laurent[3], Pierre Cherel[4], Adrien Gamot[2], Magali SanCristobal[2]

1 INRA, UR875, Unité de Biométrie et Intelligence Artificielle (UBIA), F-31326 Castanet Tolosan cedex, France

2 INRA, UMR444 Laboratoire de Génétique Cellulaire, F-31326 Castanet Tolosan cedex, France

3 Toulouse School of Economics, Université Toulouse 1, F-31000 Toulouse, France

4 Hendrix Genetics RTC, F-45808 St Jean en Braye Cedex, France

∗ **E-mail: nathalie.villa@toulouse.inra.fr**

Abstract

What are the commonalities between genes, whose expression level is partially controlled by eQTL, especially with regard to biological functions? Moreover, how are these genes related to a phenotype of interest? These issues are particularly difficult to address when the genome annotation is incomplete, as is the case for mammalian species. Moreover, the direct link between gene expression and a phenotype of interest may be weak, and thus difficult to handle. In this framework, the use of a co-expression network has proven useful: it is a robust approach for modeling a complex system of genetic regulations, and to infer knowledge for yet unknown genes.

In this article, a case study was conducted with a mammalian species. It showed that the use of a co-expression network based on partial correlation, combined with a relevant clustering of nodes, leads to an enrichment of biological functions of around 83%. Moreover, the use of a spatial statistics approach allowed us to superimpose additional information related to a phenotype; this lead to highlighting specific genes or gene clusters that are related to the network structure and the phenotype.

Three main results are worth noting: first, key genes were highlighted as a potential focus for forthcoming biological experiments; second, a set of biological functions, which support a list of genes under partial eQTL control, was set up by an overview of the global structure of the gene expression network; third, pH was found correlated with gene clusters, and then with related biological functions, as a result of a spatial analysis of the network topology.[1]

Introduction

Integrative and systems biology is a very promising tool for deciphering the biological and genetic mechanisms underlying complex traits. In this context, gene networks are used to model interactions between genes of interest: gene networks have been increasingly applied to understand the basis of complex biological phenomena [1, 2].

A gene network can be variously defined. For instance, some are based on bibliographic knowledge obtained by literature mining with software like Ingenuity Pathway Analysis (IPA), Pathway Studio or Cytoscape (compared in [3, 4]). Others combine experimental and computational approaches to define Protein-Protein Interaction - PPI - networks [5] or known biochemical and physiologic data to define metabolic networks [6]. Although biological knowledge networks are useful tools, they have some limitations due to a major lack of annotation of the genomes, and the fact that most associated literature is devoted mainly to only a few mammalian species (e.g., humans, mice and rats in IPA).

[1] Authors' contributions: NVV lead the statistical network analyses, LL the biological interpretation, AG, TL, PC, MSC participated in the analyses. All authors read and approved the manuscript.

An alternative approach is to infer the network directly from gene expression data, leading to the definition of a so-called "gene co-expression network" [7]. Inferring a co-expression network directly from gene expression data aims at focusing on direct co-expressions between genes by calculating, for instance, partial correlations [8]. Unlike in ontological enrichment analysis or bibliographic networks, information available on both functionally known and unknown genes is used for the network definition.

Once the network is given, a full analysis of its structure could be performed, from either the point of view of the network [9–11], or in correlation with a variable of interest [12]. Such analyses search for key genes, or for functional modules, or also for an understanding of the relations between the network structure and additional information (e.g., a phenotype of interest). However, regardless of the increasing number of papers focusing on networks only a few present a full analysis, starting from raw expression data, then inferring and mining the network to end up with an understanding of its relation with an external variable. For instance, [13] demonstrates the usefulness of network inference and mining for the analysis of microarray data: in the present article, the process is pushed further, allowing ones to integrate information pertaining to a phenotype. Similarly, [14] integrates expression data and PPI bibliographic network to identify candidate genes associated with a given phenotype but they do not rely on a network directly based on expression data.

In the present article, a thorough analysis is conducted. In a previous study [15], gene expressions regulated by eQTLs had been identified. 272 genes have been outlined and their biological relevance studied for those that were already annotated. Indeed the limited annotation prevented the performance of functional annotation for each cluster of eQTL. Moreover, the possible interactive links between these 272 genes, whose expression are partially regulated by eQTL, has not yet been investigated. These links can be an insight on the biological processes and can lead to the extraction of particularly important genes that are good candidates for further biological experiments.

Moreover the eQTL analysis has been done without preselecting genes to be related to a phenotype of interest. Therefore the present analysis of the gene co-expression network was made in relation to a complex phenotype, e.g., muscle pH. The muscle pH has a major industrial interest, as it is well known to be related to meat quality [16]. As the expression of the 272 genes regulated by an eQTL is only weakly correlated to muscle pH (these genes were not selected to be differentially expressed), individual analysis of gene correlation with pH is not relevant in our case. Nevertheless, our proposal is to focus on gene clusters rather than on individual relations, because clusters are more robust (i.e., less prone to be modified by noisy measurements) than each individual relation [17]. We also used an approach based on spatial statistics in order to highlight important genes that are related to the muscle pH and also to the network structure.

Focusing on this dataset, the purpose of the present paper is to gain biological knowledge from expression data for a set of genes that are partially controlled by eQTL, by proposing an adequate statistical pipeline. This pipeline is aimed at being a general tool for dissecting biological functions and interactions. The context of this work is a mammalian species with medium to low genome annotation and a gene list that does not result from a differential analysis. The proposed statistical pipeline will be briefly presented, as well as the main results, in the first Section. The Section "Materials and Methods" will then describe it in details.

Results

The raw data, which consisted of the expression of 272 genes partially controlled by eQTL, were measured *post mortem* in a muscle on 56 half sibs [15]. The statistical pipeline that was used to gain knowledge from our list of genes is summarized in Figure 1. In the remaining of this section, all results obtained from the statistical analysis are described. The following section discusses these results and a final Section "Materials and Methods" provides further details on the dataset, on the statisical methods and on their validation.

[Figure 1 about here.]

A co-expression network is first built from the 272 gene expressions, and the structure of this network is highlighted, in terms of nodes of particular importance (hubs for instance), and in terms of decomposition into "communities" or "modules". The biological meaning of each gene or of each set of genes is systematically investigated in order to validate the statistical tools. Finally, the way a quantitative trait is related to the structure of the network, is analyzed.

Network definition

A co-expression network between the 272 genes was built on partial correlations using the Gaussian Graphical Model (GGM) approach described in [8]. In this model, the network nodes are the 272 genes and edges between two nodes, which model significant correlations between the expressions of the corresponding genes. To measure the strength of the link between gene expressions, partial correlations were estimated: they are defined as the correlations between the expression of two genes *knowing the expressions of all the other genes*. As pointed out by [13], because networks focus only on the most significant links between genes, they are far less subject to noisy data; as such, they are a more robust approach than conventional analyses based on raw expression data to extract key genes and find groups of highly co-expressed genes. Moreover, the use of a partial correlation based network was compared to a more classical network based on simple correlations (i.e., "relevance network" [18]). According to the result of a node clustering combined with biological validation, the structure of the network based on partial correlations was found to be more consistent to prior biological knowledge than the one based on simple correlations (see section"Materials and Methods" for further details on this comparison). This can be explained by the fact that partial correlations focus on direct correlations only, discarding indirect links due to a common strong correlation with a third gene.

A bootstrap approach was used to estimate partial correlations. In a previous simulation study (not shown), the robustness of this approach was assessed: simulated data were generated with a given correlation design corresponding to a GGM. The estimation of the partial correlations from the bootstrap approach was compared to the real model, and about thirty observations were needed to obtain stable and accurate estimations. Thereby fifty-six observations were considered as a consistent dataset and the resulting network was indeed reliable. Finally, once the partial correlations were estimated, a Bayesian significance test was performed to discard non-significant links, i.e., edges that correspond to partial correlations, which are too small, as described in [8].

The obtained network contained 272 nodes (the genes) and 4,690 edges between significantly co-expressed genes. The network density that corresponds to the number of edges, divided by the number of node pairs was equal to 6.4%. The network was completely connected; any node in the network could be reached from any other node by a path passing along the edges.

Important nodes

The network properties are useful for highlighting some key nodes/genes. "Hubs" are often viewed as important nodes in a network: they are nodes with the largest degrees, i.e., nodes that share the largest number of connections with the other nodes. The network contained 21 hubs having a degree larger than 26; three of them had a degree equal to 29, three to 28, five to 27 and ten to 26. Additionally, the node betweenness was also calculated: it is the number of shortest paths between two nodes that pass through the node under examination. Hence, twenty-five nodes with a high betweenness (here greater than 350) were those, which connect the network: if removed, the network is more likely to be disconnected. Finally, nine genes were found to be both hubs and nodes with a high betweenness. Among these nine genes, eight were annotated and found to be connected together by the ubiquitin and the huntingtin proteins: they might correspond to genes with a connecting role between metabolic and/or signaling pathways (see

Section "Discussion" for further details). Hubs and high betweenness genes are listed in Supplemental Table 1 and are emphasized on the network in Figures 2 and 3.

[Figure 2 about here.]

[Figure 3 about here.]

Figure 2 shows that the densest part of the network contained most of the hubs (14/21) and conversely, half of its genes (14/28) are hubs. Figure 3 emphasizes the twenty-five genes that had a high betweenness. Hubs and genes with high betweenness did not provide enrichment for any given molecular function. On the contrary, the hubs have various functions such as growth factor, enzyme, transporter, component of the cytoskeleton.... Nevertheless, nine genes were hubs with a high betweenness, out of which eight were annotated. Biological enrichment was tested with IPA software for these genes, which are important for connecting the other genes together. One bibliographic network was obtained, including twenty-five out of the twenty-seven genes (score 68: this score is a quality score given by IPA; see "Biological validation" in section "Materials and Methods" for further details about this score) involved in the regulation between several signaling pathways, metabolism and cell cycle/apoptosis. This network is given in Figure 4

[Figure 4 about here.]

Network clustering

Node clustering was performed using several approaches: modularity optimization, kernel k-means and kernel SOM (see Section "Materials and Methods" for further details and references on these methods). The obtained gene clusters were systematically tested for their enrichment of Gene Ontology categories with WebGestalt. This first step lead to select the network based on partial correlations instead of simple correlations and the clustering based on modularity optimization. The clustering obtained from the modularity optimization [19] was the most consistent with biological knowledge, and was thus the one retained for further analysis. It was also the one with the highest modularity, equal to 0.4.

Seven clusters were identified that contained from 28 to 58 genes. Figure 5 provides a simplified representation of the network divided into the seven clusters.

[Figure 5 about here.]

Most hubs (14/21) belonged to cluster 6, contrary to the genes with a high betweenness that were almost equally allocated between the seven clusters. Only cluster 3 contained a larger number of genes with a high betweenness (six while the other clusters contained two to four genes with a high betweenness). The biological relevance of each cluster, as a subset of genes, was first explored in terms of Gene Ontology as explained before. Only 45% of the 272 genes had an ontological annotation, so the biological relevance was verified using Ingenuity Pathways Analysis (IPA) to construct bibliographic networks. Up to 67% of the 272 genes were eligible for network analysis by IPA. The correspondence between the clusters and the networks from IPA is given in Table 1. The relevance of the list of genes for all clusters was high, with about 83% of the eligible genes belonging to a single IPA network (at least 71%, and up to 94%). This means that the sets of genes obtained by an automatic clustering of the co-expression network have a strong consistency with the literature: they are most probably reliable for inferring the biological function of yet unknown genes according to the cluster to which these genes belong.

[Table 1 about here.]

Each cluster extracted from this analysis is fully described in Figures 2 to 7 of the supplemental material, with more information about the possible functions supported by each sub-network. The legend of all these figures is given in Figure 1 of the supplemental material.

Relations between the co-expression network and a phenotype of interest

In order to assess if a correlation existed between the network topology (the clusters) and a phenotype of interest (muscle pH), the partial correlations between pH and gene expressions were calculated. The pH values of muscle tissue after slaughtering are related to meat quality. Only the ultimate pH value (measured 24h after slaughtering) is available but it is known to be not accurate enough to discriminate the metabolic processes underlying the way pH declines [20]. The purpose of the present section is thus to understand the relation between our set of genes (that are under eQTL control), their functions and this phenotype.

First, a Moran's permutation test was performed to assess the correlation between the network structure and the partial correlation values. This test aims at answering the following question: "Do nodes that are linked in the co-expression network have a tendency to be similarly correlated with pH?" To that aim, Moran's I [21] was calculated: Moran's I is a weighted correlation coefficient used to detect departures from spatial randomness. A statistical test, based on random permutations, as described in [22], was performed to assess the significance of its value: it was proven that Moran's I was significantly larger in our network than if the partial correlations were distributed among the nodes independently from the network structure. This means that nodes linked in the network have similar correlations with pH.

Moving down to the cluster level, it was then possible to show that genes in cluster 4 had a significantly higher partial correlation with the pH than the genes in the other clusters (Figure 6), according to a t-test. Note that the values of the partial correlation should not be compared to the values of the correlations: a strong correlation between two genes results in a correlation coefficient close to one or to minus one but a similar behavior is not to be expected from the partial correlations: these quantities are conditional correlations and are thus much smaller than the direct correlations. To check that the other clusters had no correlation with the pH, the absolute values of the partial correlation with the genes expressions in cluster 4 were also calculated. This confirmed that cluster 4 is significantly more connected to the variation of muscle pH than the genes in the other clusters, according to a t-test. With a bibliographic network IPA analysis, cluster 4 was found to be related to cell death and cell cycle, with three genes (*GPI*, *B2M* and *XIAP*) essentially regulating cell death. Further discussion is provided in Section "Discussion".

[Figure 6 about here.]

Finally, the gene level was also studied by using Moran's plot to detect influential genes [23]. Moran's plot displays the average values for partial correlation with pH in the neighborhood of a node as a function of the partial correlation with pH for this node (Figure 7). In this plot, the way a gene is linked to pH is analyzed together with its neighboring genes in the network. For instance, *GPI* is an influential gene in the quadrant "H-H": this means that, not only is its expression highly correlated to the pH value but its neighboring genes also have an expression that is highly correlated to the pH value. Indeed *GPI* has an expected function (glycolysis) involved in the regulation of pH. A more complete discussion about *GPI* is provided in Section "Discussion".

[Figure 7 about here.]

Thereby, influential nodes for pH [24] were extracted from Moran's plot; most of them belonged to cluster 4 (Figure 8).

[Figure 8 about here.]

Twenty genes were detected as influential in Moran's plot and eleven of them were in cluster 4 (out of twenty-eight genes classified in cluster 4). From these twenty genes, ten genes were eligible by IPA and were all included in the same biological network (Figure 9).

[Figure 9 about here.]

Supplemental Table 1 contains the gene description (accession number, gene name, gene description, heritability, number of eQTL, putative cis-eQTL, genomic localization), along with the results of our analysis (degree, hub, betweenness, cluster, differentially expressed for pH, influent for partial correlation with pH, influent for absolute value of partial correlation with pH).

Discussion

The overall methodology described in this paper is a pipeline of statistical methods to gain knowledge from raw data on a selected (here genetically regulated by eQTL) set of genes. This pipeline includes three steps.

- The definition of a co-expression network to give a simplified and significant view of the interaction structure between those genes. This network can be used to identify key genes.
- A clustering of the nodes based on this network, built only from significant relations between genes. It helps to identify relevant groups of genes with a common function.
- External information, related to a trait, has been integrated into this network. In our case, the network structure was proven to be correlated with the value of the correlation between the gene expression and the pH. Moreover, the correlation between gene expression and pH, used together with the network structure, helped to identify important genes related to pH. Most of the genes that were identified as related to the pH, were also involved in a same cluster with other genes sharing biological functions (cluster 4). Moreover, all the annotated genes correlated to the pH were also involved in one biological network (Figure 8).

A relevant strategy to model a gene network

Inferring a co-expression network directly from gene expression data can be achieved with a large number of statistical approaches: among them, the most studied are probably Gaussian Graphical Model (GGM) [25], Bayesian networks [26, 27] or mutual information networks [28]. As network inference is a topic of much interest, several packages have also been developed for the free statistical software R: for example, **GeneNet** [8] is a Graphical Gaussian method including a Bayesian significance test; **GGMselect** is a sparse Graphical Gaussian approach (see Baraud et al. 2009: `http://fr.arxiv.org/abs/0907.0619`); **minet** [28] is an R/bioconductor package using mutual information; **bnlearn** [29] is based on Bayesian network learning.

In the present article, a GGM was used, as implemented in the R package **GeneNet** to infer the network from a bootstrap approach and a Bayesian test. GGM is based on the estimation of partial correlations. As mentioned in the review of [30], the use of partial correlations instead of simple correlations is more appropriate to measure the dependence between variables. The correlation has to be preferred when the independence between variables is the targeted problem. Hence, the method combines the availability of a dedicated R package, with good performances, compared to several other alternatives [31].

Extracting putative key genes from the network

The analysis of the network had two main purposes: first was to highlight key genes for co-expression, and second was to gain knowledge on unknown genes. Key genes were found by a direct analysis of the structure of the network, or by superimposing information (related to a phenotype of interest) to the network.

Several characteristics related to the network structure can be calculated to provide insights about key genes [9]. For instance, hubs are genes with the highest degree and have been proven to organize

the proteome by connecting biological processes [10] or to be implicated in cancer [32]. The betweenness centrality measure [33] is well known in social network analysis but less standard in biological network analysis. Betweenness is an interesting criterion as nodes with a high betweenness form a strong network connection and hence have a strong impact on the network structure. Therefore the modification of these genes might have a large impact on underlying biological functions. This fact has already been described in medicine [11], in a study on network evolution [34], and in protein-protein interaction networks [35].

A few examples of extracted genes are provided thereafter. Their possible relevance in the way a muscle functions, or their possible involvement in pH values, is emphasized when existing studies have previously described that point. These examples aim at illustrating that some genes, which were automatically extracted thanks to the co-expression network model, showed a strong relevance for the understanding of the considered biological process. In our study, nine genes were both hubs and nodes with a high betweenness (*TRIAP1*, *SUZ12*, *PRDX4*, *GPI*, *SSR4*, *FTH1*, *MGP*, *SLC39A14* and *BX921641*). The eight that were annotated, were connected together by the ubiquitin and the huntingtin proteins (see Figure 4). A hypothesis is that these nodes could correspond to genes with a connecting role between metabolic and/or signaling pathways. These two proteins (ubiquitin and huntingtin) are ubiquitous and involved in several pathways. As explained by [36] in a review dedicated to the function of the huntingtin protein, huntingtin may interfere with transcriptional mechanisms common to many genes including markers of terminal muscle differentiation, metabolic enzymes (as GPI in cluster 4), signal transduction molecules, and fast myofibrillar fibers (as troponin 1 present in the cluster 2). Some mRNAs (e.g., ubiquitin-conjugating enzymes) concurrently increased in muscle, implying a cellular stress response.

Globally, extracted genes were either:

- annotated genes known to be involved in muscle physiology or even in meat quality. For example, GPI (glucose-6-phosphate isomerase) was a hub, a gene with a high betweenness, and an influential node for the correlation with pH. GPI protein is known to be involved in energy pathways, glycolysis and gluconeogenesis and these pathways are well known to be related to meat quality. Moreover GPI is localized on chromosome 6 at the position of several QTLs (Quantitative Trait Locus) affecting ultimate pH in loin muscle. Hence proposing GPI as a positional and functional candidate gene makes sense [37].

- annotated genes never cited for being involved in muscle physiology and even less in meat quality. For example, *MGP* (Matrix gla protein) was the hub with the highest degree and also had a high betweenness. [38] proved that it is involved in the inhibition of the switch from vascular smooth muscle cell in osteoblast-like cells and also calcification of arteries. To our knowledge, nothing has been described for *MGP* in skeletal muscle except in our first study [15] in which we identified a putative *cis*-eQTL for this gene.

From the relevance of the previous conclusions, it seems therefore interesting to focus on:

- genes which are un-annotated and whose function is therefore unknown. For instance, *BX921641* is a hub and also the gene with the highest betweenness. In further studies, it would be interesting to investigate the function of this gene in muscle tissue.

The main biological finding of this study, compared to a bibliographic gene network study (like IPA), lies in the fact that the combination of statistical methods is able to be used in the same analysis for all the genes of interest, either functionally known or not. Among the 56 genes highlighted as being "important" (hubs, high betweenness, or high influence for their correlation with a trait), only 67% are functionally known, and the others would have been discarded with solely a standard analysis, based on bibliographic knowledge.

Gaining knowledge from the gene clustering

The second step of the proposed pipeline was to elucidate the biological meaning of the gene network. The complete network with 272 genes was difficult to read except for the densest part of the network, as it is usual for networks with more than a hundred nodes. Indeed, as explained in [39] the standard way to display networks, i.e., by the use of force directed placement algorithms such as the algorithm described in [40] is not enough to identify a structure inside the network. Indeed, groups of genes (also often called "modules") that are the most densely connected (and comparatively less connected to the other nodes) can often not be identified visually. The general structure of the network, decomposed into sub-graphs, can be revealed using node clustering. [41] provided a very complete review of methods used to cluster the nodes of a network and [42] compared several popular methods to cluster protein-protein interaction networks. This promising approach aims at revealing the biological structure behind the statistical one: it is a well-known fact that biological functions are carried out by modules in interaction networks [43]. Moreover, as pointed out by [17], network inference is more robust when dealing with modules than with individual interactions. Here, this approach was proven to be highly powerful to cluster together genes with common biological functions. Several methods to cluster genes were tested and biologically compared to each other with systematic measurements of ontological enrichment with the WebGestalt software [44] (see Section "Materials and Methods" for further details). The best clustering was also submitted to IPA. Nearly all the genes eligible to be submitted to IPA (about 80%) were included in a same bibliographic network: one cluster corresponded to one IPA bibliographic network (Table 1). It thus gave clues to the biological role of a group of genes, including the unknown genes.

A more complete discussion is proposed for each cluster in the supplemental material, where each cluster is displayed in Figures 2 to 7.

Integrating additional information related to a phenotype of interest

It is of major interest to add phenotypic information to a co-expression network in an integrative strategy to combine different levels of information. Moreover in our context, the 272 genes have been identified, so as to have their expression genetically regulated by eQTL, and moreover, without being selected to be differentially expressed according to a phenotype [15]. An important biological result of this work was to be able to merge co-expression information with the correlation to a trait of interest (muscle pH). In addition to the structure of the gene network, [1] showed that the relation with a phenotype or a trait of interest can be helpful to decipher the molecular interactions underlying a complex trait. When the phenotype is discrete, such as healthy/cancer, methods have been proposed such as COSINE [12]: differentially expressed genes (DEG) and differential correlation between groups are combined in this method. When the genes under study are not DEG, the relation between the selected genes and the phenotype may be weak, and such an approach is not usable anymore. In our study, this issue was addressed by using a labeled network, i.e., a network whose nodes are labeled by additional information, because this approach combines interactions between genes and correlations to a phenotype of interest, in the same model. It does not rely on individual tests for each gene and it is thus better suited for understanding the process in its totality and to extract groups of genes related to the phenotype.

Finally, the pH reflects the acid-base homeostasis of a living system (muscle tissue). From the 20 genes found to be important for the partial correlation with the pH, only *GPI* has an expected function (glycolysis) involved in the regulation of pH: accelerated postmortem glycolysis affects a rapid pH fall. Moreover, *GPI* gene is included in cluster 4, which has the highest correlation with the variation in pH. This gene seemed to be the most important gene in this cluster. *GPI* is altogether a hub with a high betweenness, and a gene highly correlated to pH variation. In the postmortem muscle of pigs, the energy metabolism shifts from an aerobic metabolism of lipids to anaerobic metabolism of muscle glycogen. Unfortunately, the way the ultimate pH decreases is rather difficult to control. Nevertheless, ultimate pH is often measured as a consequential factor [45]. With the identification of *GPI*, which was

central in a network related to pH, geneticists are offered interesting proposals for further experiments. *GPI* gene is a functional and positional gene candidate to explain effects observed at a QTL position on chromosome 6 on muscle pH values [37]. The expression of *GPI* is genetically regulated by two *trans*-eQTL on chromosomes 5 and 8 in our context, and no *cis*-eQTL was identified on chromosome 6 [15]. With a bibliographic network from IPA analysis, cluster 4 was found to be related to cell death and cell cycle: this IPA network included 78% of the annotated genes classified in cluster 4. Intracellular pH has an important role in the maintenance of normal cell function, and cellular modifications leading to pH changes have been implicated in both cell proliferation and cell death [46]. In our study, three genes essentially regulate cell death, (*GPI*, *B2M* and *XIAP*) suggesting a relation between pH regulation, which is a metabolic process, and cell death, which is the cell biological consequence of the failure of metabolism.

Conclusion

An adequate combination of statistical methods, namely network inference using partial correlation under a Graphical Gaussian model, followed by node clustering, can lead to a significant improvement of our biological knowledge in the underlying biological functions of a set of genes. This approach is particularly useful in the context partial bibliographic knowledge where only half of the genes a given genome are still unknown. Moreover, this approach allows one to link the structure of the network to a phenotype, and then to identify key genes.

Materials and Methods

Data description: eQTL data

56 half sib pigs were produced from an F2 cross between two production sire lines (France Hybrides SA, St. Jean de Braye, France). Procedures and facilities were approved by the French Veterinary Services. Longissimus dorsi muscle RNA was extracted as described by [47]. The normalized data were submitted to NCBI (GEO accession number GSE26924). The eQTL analysis identified 335 eQTLs affecting the expression of 272 transcripts with an average heritability of 0.45 $\pm$0.25 [15].

Network definition

In the Gaussian graphical model framework, gene expressions are modeled by a Gaussian variable $(X_j)_{j=1,\dots,p}$, where p is the number of genes under study, with a covariance matrix Σ. It can be proved that the partial correlations

$$\mathbb{C}\text{or}\left(X_j, X_{j'} | (X_i)_{i \neq j,j'}\right)$$

are obtained directly from Σ^{-1} [25]. Many articles focus on the estimation of this inverse in the context of ill-posed problems: typically, the number of genes is much larger than the number of available observations, and directly inverting the empirical correlation matrix leads to numerical instability and bad estimations. One solution is the bootstrap estimation described in [8].

This approach was used, combined with the estimation functions implemented in the R package **GeneNet**. In this package, a shrinkage of the empirical covariance matrix Σ is performed prior to its inversion in order to limit numerical instability. This method simply consists in adding a small positive number to the diagonal of Σ. A bootstrap approach was then performed to obtain more robust estimates and made it possible to construct a co-expression network with the 272 genes. 4,000 bootstrap samples (size 20) were enough to obtain a stabilization of the estimation procedure. Then, the Bayesian test of significance, described in [8], and implemented in the R package **GeneNet**, was used to discard the smallest partial correlations. Finally, the network was displayed using the Fruchterman and Reingold algorithm [40] as implemented in the R package **igraph** [48].

Network clustering

Node clustering aims at finding densely connected groups of genes, called *clusters* or *modules*, in the network. As many methods to cluster the nodes in a network exist [41], three were chosen and compared. The first one consisted in optimizing the modularity: the modularity is a quality criterion for node clustering introduced by [19]. For a network with nodes $\{1, \ldots, n\}$ and edges weighted by W_{ij} (where $W_{ij} = W_{ji}$ are either positive or null when there is no edge between x_i and x_j) and for a partition $\mathcal{C}_1, \ldots, \mathcal{C}_K$ of the nodes, the modularity is equal to:

$$\sum_{k=1}^{K} \sum_{i,j \in \mathcal{C}_k} (W_{ij} - P_{ij}) \,,$$

with $P_{ij} = \frac{d_i d_j}{2m}$, where d_i is the degree of x_i and m is the number of edges in the network. Its aim is to compare the actual weights of the edges to a null model where the edges depend only on the nodes degrees and not on their cluster. Hence, the higher the modularity, the more the edges are concentrated inside the clusters. In the case of unweighted networks (as in our study), W_{ij} are either 1 (when there is an edge between x_i and x_j) or 0. The modularity measure has already been used by [49] to recover functional modules in protein interaction networks with an optimization based on the original approach of [19]. Following the ideas of [50], the modularity was optimized by a simulated annealing algorithm, which is a more efficient approach for optimizing the criterion than the one proposed in the original article [19]. The annealing parameter of simulated annealing was chosen in an exponential search grid (varying from 10 to 105).

Modularity optimization was compared to alternative approaches that were based on *kernels* (see, among others, [51]) and all relate to spectral clustering [52]. More precisely, kernel k-means and batch kernel SOM [53] were processed as implemented in the R package **yasomi** (development version available at `https://r-forge.r-project.org/projects/yasomi`).

To compare the three different methods (modularity optimization by simulated annealing, kernel k-means and batch kernel SOM), the following methodology was used:

- for each of the three methods, several parameters were used to provide different results: the number of (initial) clusters of the algorithm varied from 4 to 12 (or, for kernel SOM, the data were projected on a 2-dimensional grid whose dimension varied from 2 to 4) and the heat kernel [54] and the Commute Time kernel [55] were tested;
- for each of the three methods, only one of these results was selected: the selection was made according to the modularity value (hence the modularity was also used as a quality measure for selecting the "best" clustering among the cluterings obtained with different tuning parameters);
- the resulting three clustering were subjected to a biological validation (as described in the next section).

Also, the same methods were also used to cluster a network based on simple correlations ("relevance network", see [56]) in order to assess the relevance of the use of partial correlations compared to simple correlations.

Finally, as explained in the next section, biological validation leads to select the clustering based on the partial correlation network and on modularity optimization by simulated annealing.

Biological validation

In a first step, the WebGestalt software [44] provided a statistical enrichment analysis of the Gene Ontology Terms. The results were illustrated with an acyclic network of the ontology terms. Biological

information given by GO enrichment is only based on a low number of genes (47%). However WebGestalt produced results faster than IPA and was more useful for the comparisons between different networks (defined from partial correlation or direct correlation) and different clustering algorithms with various parameters. A systematic comparison was performed to assess the biological relevance of the clusters obtained from these different methods: only the most relevant clustering was then analyzed, i.e., that which was obtained from the modularity optimization of the network built with partial correlations. A unique network (the one based on partial correlations) and a unique clustering (the one based on modularity optimization) was then kept, because they had the best agreement with biological knowledge, as computed by using the WebGestalt software.

In a second step, Ingenuity Pathways Analysis (IPA, `https://analysis.ingenuity.com/pa`) was used to explore and confirm the biological relevance of the identified clusters. IPA software contains a large bibliographic database with various types of links already identified between two genes (protein-protein interaction, ligand-receptor regulation, enzymatic modification, transcriptional expression regulation, etc.). IPA software was used to build biological networks, which correspond to the best possible arrangement of the eligible genes. 67% of the 272 genes (Table 1) are genes that have already been studied elsewhere and are annotated and referenced in IPA. For each IPA network a score is used to rank networks according to their degree of relevance to the "Network Eligible Molecules" (the input gene list) in the submitted dataset. The score is derived from a p-value (based on the hypergeometric distribution and calculated with the right-tailed Fisher's exact test) and indicates the likelihood of the submitted genes to be found together in the same network due to random chance; for instance, a score of 2 indicates that there is 1 in 100 chance that the submitted genes are together in a network due to random chance.

Using spatial statisics to analyze the link with a phenotype of interest

A final analysis focused on the relation between the network structure and a phenotype of interest (muscle pH). This analysis was performed by first calculating partial correlations between gene expressions and pH, using the same method that was described in the section "Network definition". Then, tools coming from spatial statistics were used to extract influential genes. This approach is the one described in [22]. Briefly, it consisted first in calculating the Moran's I statistics to measure the correlation between the network structure and the phenotype of interest and to perform a permutation test to assess its significance and, then in finding the genes that had the strongest effects in the correlation between the value of the variable for a given node and the average value of this variable for its neighbors.

Additionally, the significance of a higher correlation with pH in one particular cluster compared to the others, was assessed by means of a t-test with level 1%, testing the difference in average between the absolute value of the partial correlation with pH in the considered cluster and the absolute value of the partial correlation with pH in the other 6 clusters.

Acknowledgments

The authors thank the reviewers for their useful comments and suggestions that helped improve the quality of this paper. The authors are also grateful to Nicholas Szczesniak for the English revision of the manuscript. in

References

1. Schadt E (2009) Molecular networks as sensors and drivers of common human diseases. Nature 461: 218-223.

2. Barabási A, Gulbahcel N, Loscalzo J (2011) Network medecine: a network-based approach to human disease. Nature Reviews Genetics 12: 56-68.

3. Hedegaard J, Arce C, Bicciato S, Bonnet A, Buitenhuis B, et al. (2009) Methods for interpreting lists of affected genes obtained in a DNA microarray experiment. BMC Proceedings 3: S5.

4. Bonnet A, Lagarrigue S, Liaubet L, Robert-Granié C, SanCristobal M, et al. (2009) Pathway results from the chicken data set using GOTM, Pathway Studio and Ingenuity softwares. BMC Proceedings 3: S11.

5. von Mering C, Krause R, Snel B, Cornell M, Olivier S, et al. (2003) Comparing assessment of large-scale data sets of protein-protein interaction data? Journal of Molecular Biology 327: 919-923.

6. Bordbar A, Palsson B (2012) Using the reconstructed genome-scale human metabolic network to study physiology and pathology. Journal of Internal Medecine 271: 131-141.

7. Zhang B, Horvath S (2005) A general framework for weighted gene co-expression network analysis. Statistical Applications in Genetics and Molecular Biology 4: 17.

8. Schäfer J, Strimmer K (2005) An empirical bayes approach to inferring large-scale gene association networks. Bioinformatics 21: 754-764.

9. Dong J, Horvath S (2007) Understanding network concepts in modules. BMC Systems Biology 1: 24.

10. Han J, Bertin N, Hao T, Goldberg D, Berriz G, et al. (2004) 17. evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature 430: 88-93.

11. Lord L, Allen P, Expert P, Howes O, Lambiotte R, et al. (2011) Characterization of the anterior cingulate's role in the at-risk mental state using graph theory. Neuroimage 56: 1531-1539.

12. Ma H, Schadt E, Kaplan L, Zhao H (2011) COSINE: condition-specific sub-network identification using a global optimization method. Bioinformatics 27: 1290-1298.

13. Freeman T, Goldovsky L, Brosch M, van Dongen S, Mazière P, et al. (2007) Construction, visualisation, and clustering of transcription networks from microarray expression data. PLoS Computional Biology 3: e206.

14. Ma X, Lee H, Wang L, Sun F (2007) CGI: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data. Bioinformatics 23: 215-221.

15. Liaubet L, Lobjois V, Faraut T, Tircazes A, Benne F, et al. (2011) Genetic variability or transcript abundance in pig peri-mortem skeletal muscle: eQTL localized genes involved in stress response, cell death, muscle disorders and metabolism. BMC Genomics 12: 548.

16. Le S, Joo S, Ryu Y (2010) Skeletal muscle fiber type and myofibrillar proteins in relation to meat quality. Meat Science 86: 166-170.

17. De Smet R, Marchal K (2010) Advantages and limitations of current network inference methods. Nature Reviews Microbiology 8: 717-729.

18. Butte A, Kohane I (1999) Unsupervised knowledge discovery in medical databases using relevance networks. In: Proceedings of the AMIA Symposium. pp. 711-715.

19. Newman M, Girvan M (2004) Finding and evaluating community structure in networks. Physical Review, E 69: 026113.

20. Lengerken G, Maak S, Wicke M (2002) Muscle metabolism and meat quality of pigs and poultry. Veterinarija Ir Zootechnika 20: 82-86.

21. Moran P (1950) Notes on continuous stochastic phenomena. Biometrika 37: 17-23.

22. Laurent T, Villa-Vialaneix N (2011) Using spatial indexes for labeled network analysis. Information, Interaction, Intelligence (i3) 11.

23. Anselin L (1995) Local indicators of spatial association-lisa. Geographical Analysis 27: 93-115.

24. Cook R, Weisberg S (1982) Residuals and Influence in Regression. London: Chapman and Hall.

25. Edwards D (1995) Introduction to Graphical Modelling. New York: Springer.

26. Pearl J (1998) Probabilistic reasoning in intelligent systems: networks of plausible inference. San Francisco, California, USA: Morgan Kaufmann.

27. Pearl J, Russel S (2002) Bayesian Networks. Cambridge, Massachussets, USA: Bradford Books (MIT Press).

28. Meyer P, Lafitte F, Bontempi G (2008) minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information. BMC Bioinformatics 9.

29. Scutari M (2010) Learning Bayesian networks with the bnlearn R package. Journal of Statistical Software 35: 1-22.

30. Markowetz F, Spang R (2007) Inferring cellular networks - a review. BMC Bioinformatics 8: S5.

31. Villers F, Schaeffer B, Bertin C, Huet S (2008) Assessing the validity domains of graphical Gaussian models in order to infer relationships among components of complex biological systems. Statistical Applications in Genetics and Molecular Biology 7: 14.

32. Carter S, Brechbühler C, Griffin M, Bond A (2004) Gene co-expression network topology provides a framework for molecular characterization of cellular state. Bioinformatics 20: 2242-2250.

33. Freeman L (1979) 15. centrality in social networks I: conceptual clarification. Social Networks 1: 215-239.

34. Jordan I, Katzl L, Denver D, Streelman J (2008) Natural selection governs local, but not global, evolutionary gene coexpression networks in Caenorhabditis elegans. BMC Systems Biology 13: 2-96.

35. Hwang W, Chol Y, Zhang A, Ramanathan M (2006) A novel functional module detection algorithm for protein-protein interaction networks. Algorithms for Molecular Biology 1: 24.

36. Sassone J, Colciago C, Cislaghi G, Silani V, Ciammol A (2009) Huntington's disease: The current state of research with peripheral tissues. Experimental Neurology 219: 385-397.

37. Li H, Lund M, Christensen O, Gregersen V, Henckel P, et al. (2010) Quantitative trait loci analysis of swine meat quality traits. Journal of Animal Science 88: 2904-2912.

38. Verhave G, Siegert C (2010) Role of vitamin D in cardiovascular disease. Netherlands Journal of Medicine 68: 113-118.

39. Noack A (2007) Energy models for graph clustering. Journal of Graph Algorithms and Applications 11: 453-480.

40. Fruchterman T, Reingold B (1991) Graph drawing by force-directed placement. Software, Practice and Experience 21: 1129-1164.

41. Fortunato S (2010) Community detection in graphs. Physics Reports 486: 75-174.

42. Brohée S, van Helden J (2006) Evaluation of clustering algorithms for protein-protein interaction networks. BMC Bioinformatics 7.

43. Hartwell L, Hopfield J, Leibler S, Murray A (1999) From molecular cell biology. Nature 402: C47-C54.

44. Zhang B, Kirov S, Snoddy J (2005) WebGestalt: an integrated system for exploring gene sets in various biological contexts. Nucleic Acids Research 1: W741-W748.

45. Ngapo T, Gariepy C (2008) Factors affecting the eating quality of pork. Critical Reviews in Food Science and Nutrition 48: 599-633.

46. Lagadic-Gossmann D, Huc L, Lecureur V (2004) Alterations of intracellular pH homeostasis in apoptosis: origins and roles. Cell Death and Differentiation 11: 953-961.

47. Lobjois V, Liaubet L, SanCristobal M, Glenisson J, Feve K, et al. (2008) A muscle transcriptome analysis identifies positional candidate genes for a complex trait in pig. Animal Genetics 39: 147-162.

48. Csardi G, Nepusz T (2006) The igraph software package for complex network research. InterJournal Complex Systems.

49. Dunn R, Dudbridge F, Sanderson C (2005) The use of edge-betweenness clustering to investigate biological function in protein interaction networks. BMC Bioinformatics 6: 39.

50. Reichardt J, Bornholdt S (2006) Statistical mechanics of community detection. Physical Review, E 74.

51. Schölkopf B, Tsuda K, Vert J (2004) Kernel methods in computational biology. London: MIT Press.

52. Inoue K, Weijiang L, Kurata H (2010) Diffusion model based spectral clustering for protein-protein interaction networks. PloS One 5: e12623.

53. Boulet R, Jouve B, Rossi F, Villa N (2008) Batch kernel SOM and related laplacian methods for social network analysis. Neurocomputing 71: 1257-1273.

54. Kondor R, Lafferty J (2002) Diffusion kernels on graphs and other discrete structures. In: Proceedings of the 19*th* International Conference on Machine Learning. pp. 315-322.

55. Fouss F, Pirotte A, Renders J, Saerens M (2007) Random-walk computation of similarities between nodes of a graph, with application to collaborative recommendation. IEEE Transactions on Knowledge and Data Engineering 19: 355-369.

56. Butte A, Kohane I (2000) Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In: Proceedings of the Pacific Symposium on Biocomputing. pp. 418-429.

List of Figures

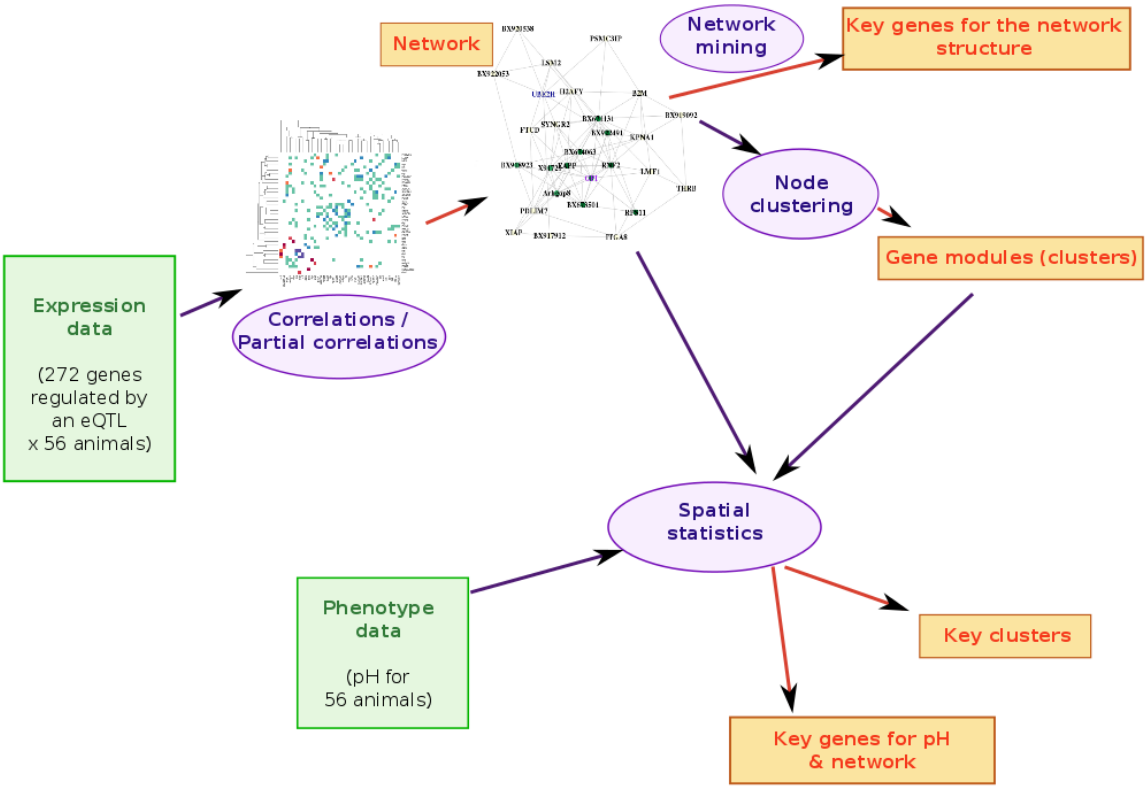

Figure 1. Summary of the statistical pipeline. Data are represented in green (expression data and pH), statistical methods are represented in purple, results are represented in red.

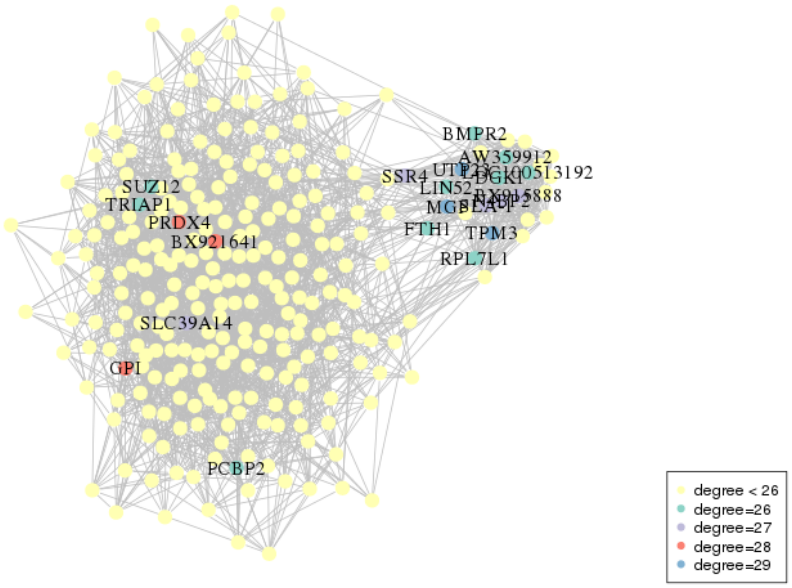

Figure 2. The co-expression network where hubs are highlighted. The names are also given. The list of hubs is available in Supplemental Table 1.

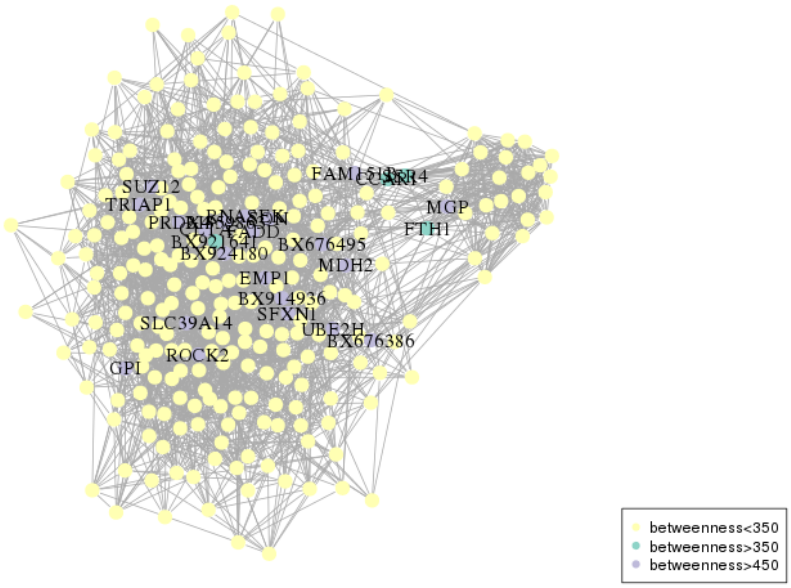

Figure 3. The co-expression network where genes with high betweenness are highlighted. The names are also given. The list of genes with high betweenness is available in Supplemental Table 1.

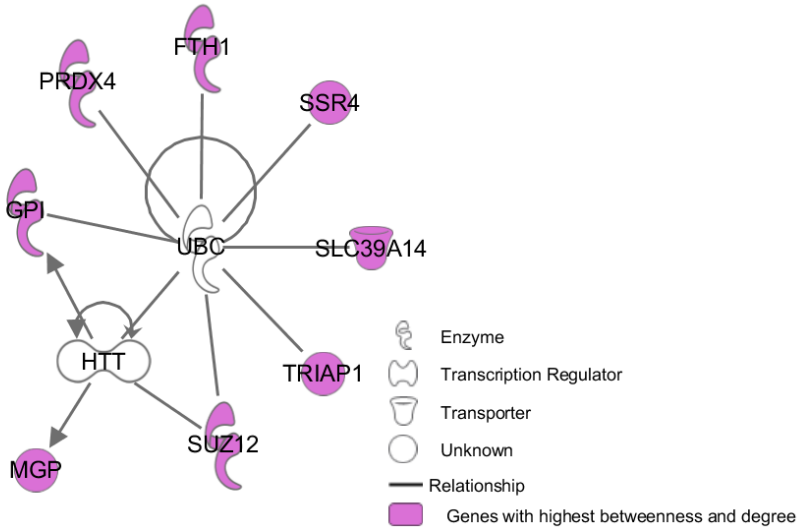

Figure 4. Bibliographic network obtained with the 8 annotated genes out of the 9 having the highest degree and betweenness. This network (score 68: this score is a quality score given by IPA; see "Biological validation" in section "Materials and Methods" for further details about this score) is related to regulation between several signaling pathways, metabolism and cell cycle apoptosis.

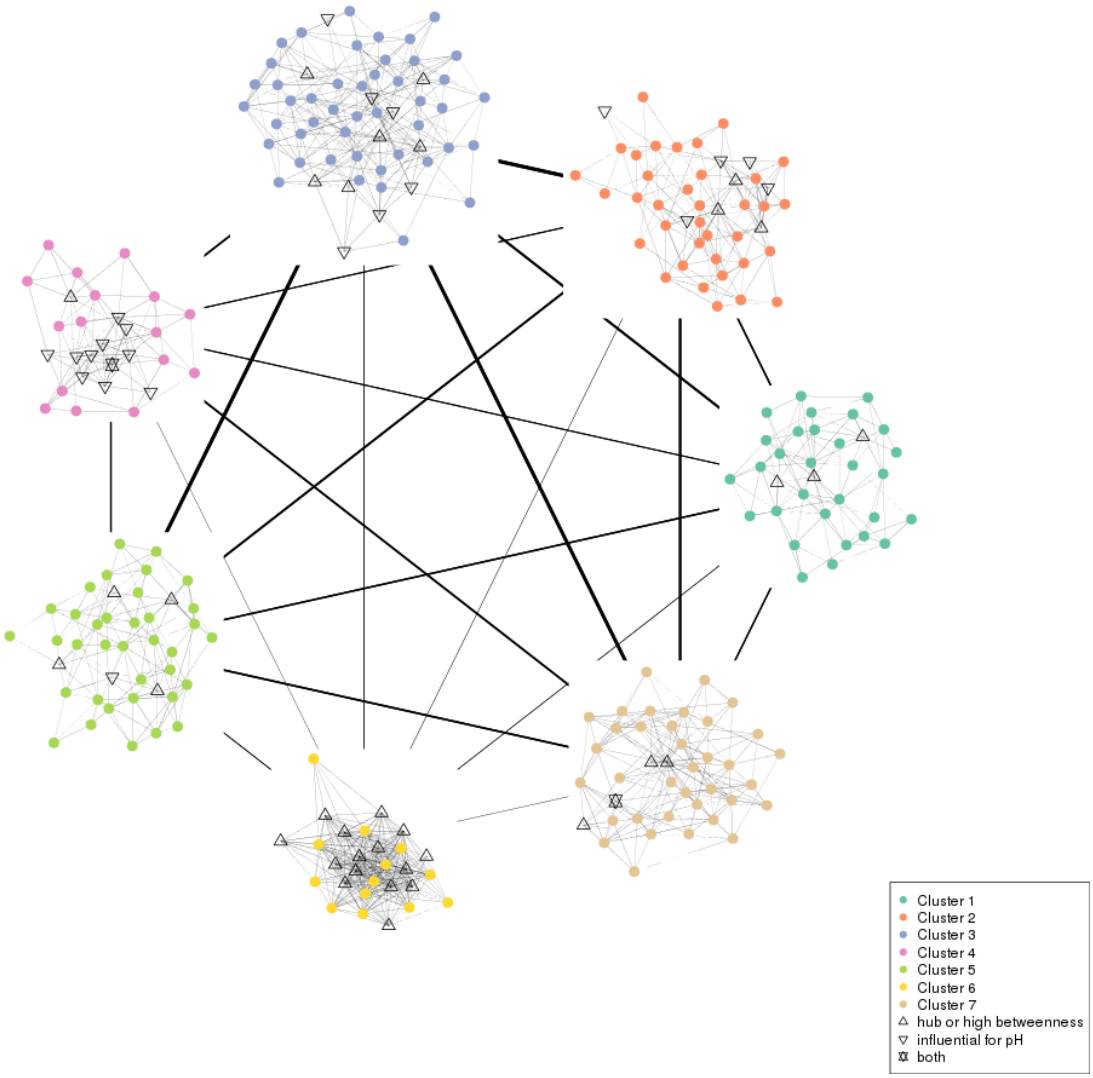

Figure 5. Simplified representation of the network. Special nodes are highlighted according to their level of degree or betweenness, and/or their partial correlation to a phenotype related to meat quality (pH 24h after slaughtering). The line width between clusters is proportional to the number of links between the nodes of the corresponding clusters.

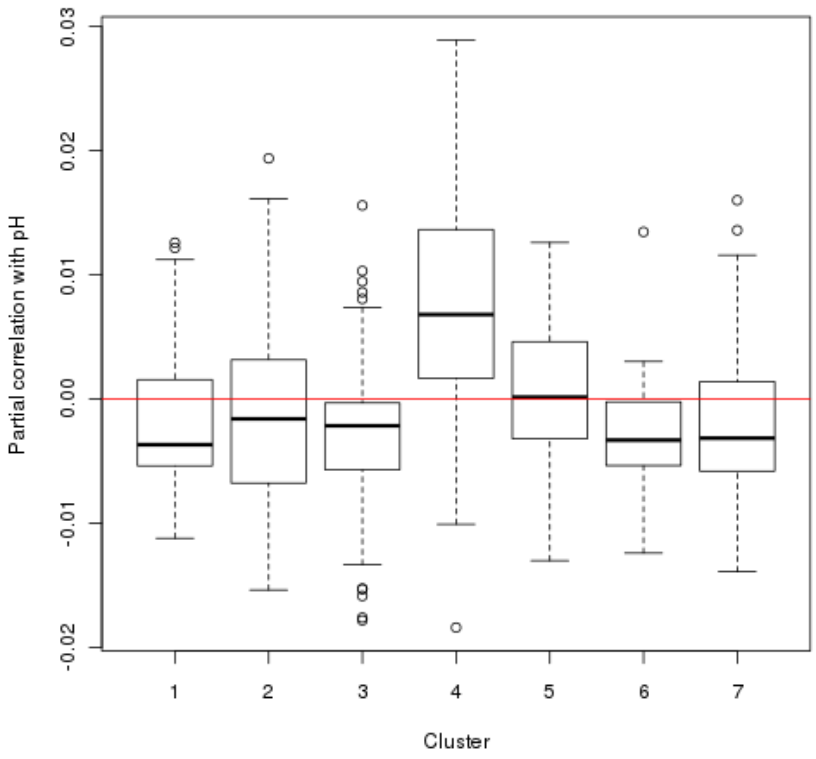

Figure 6. Boxplots of the partial correlations between the gene expressions and the pH for each cluster. Cluster 4 is significantly correlated with the pH phenotype (p-value is equal to 0.001).

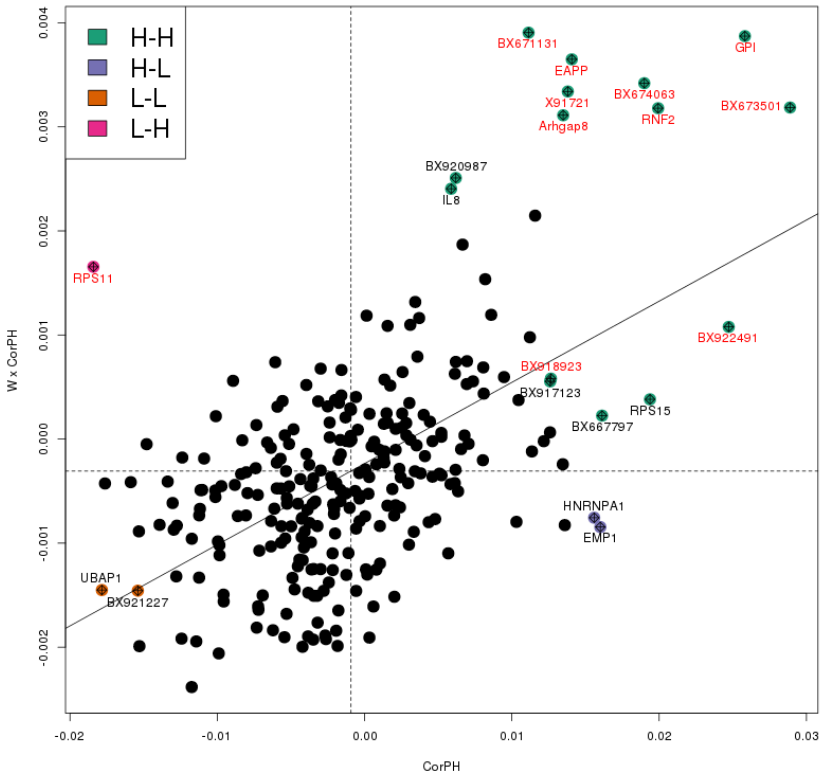

Figure 7. Moran's plot of the partial correlation between pH and expression levels in the co-expression network. Influential nodes are displayed in color and their names are given. Influential genes labeled "H-H" have a strong positive correlation with pH (above the mean) and are linked to genes having a strong positive correlation with pH (above the mean); influential genes labeled "H-L" have a strong positive correlation with pH (above the mean) and are linked to genes having a strong negative correlation with pH (below the mean); influential genes labeled "L-H" have a strong negative correlation with pH (below the mean) and are linked to genes having a strong positive correlation with pH (above the mean); influential genes labeled "L-L" have a strong negative correlation with pH (below the mean) and are linked to genes having a strong negative correlation with pH (below the mean). Genes in red are in cluster 4, the cluster that is the most correlated to pH.

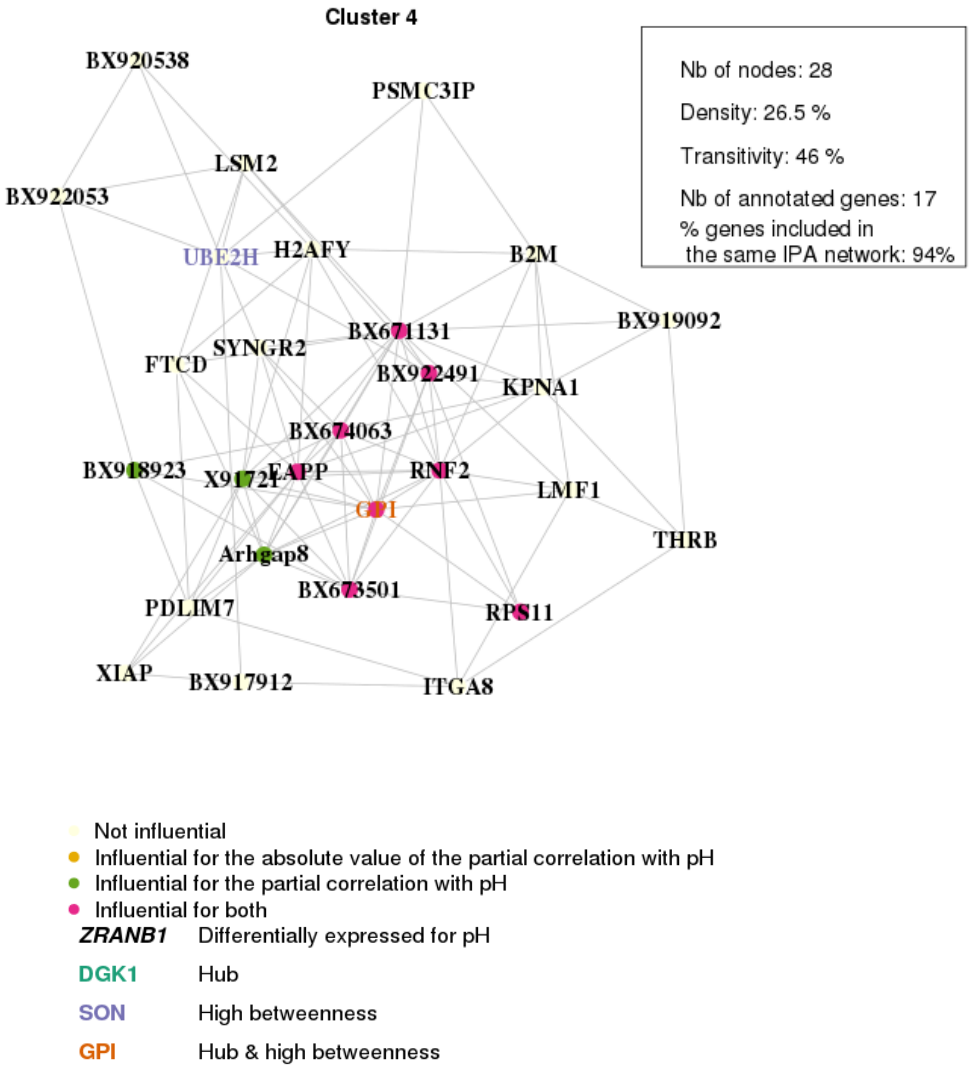

Figure 8. Detailed display of cluster 4. Nodes that are influential for the partial correlation with pH, as well as nodes that are important for the structure of the graph (hubs, high betweenness), are highlighted. The other clusters are displayed similarly in Supplemental Material.

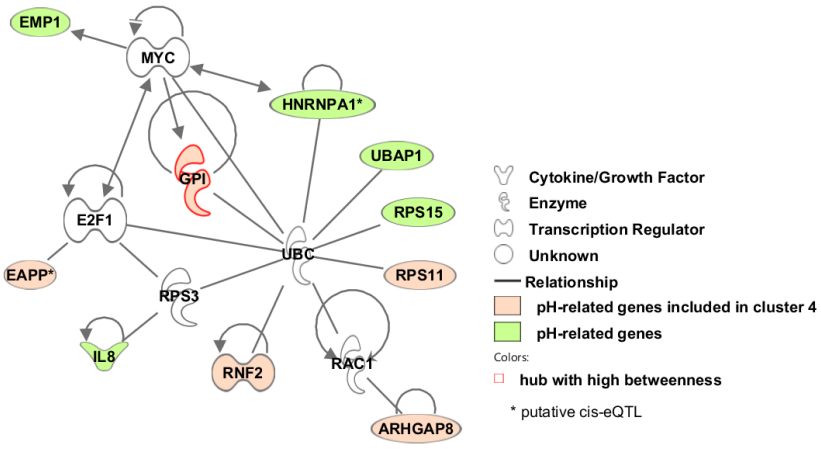

Figure 9. Bibliographic network obtained with 10 pH-related genes. Pink nodes are the genes included in cluster 4; the other nodes are green. Finally, white nodes are the genes included by IPA to define the network but not shown to be regulated by an eQTL in our previous study.

List of Tables

Table 1. Correspondence between clusters found by node clustering and bibliographic network

Cluster	**Nb of genes in the cluster**	**Nb of genes called eligible**	**% of the eligible genes involved in the same biological network**	**Score**	**Main biological functions associated with the network**
1	33	24	71	49	Development, cell death
2	44	28	93	70	Folding of protein, neuromuscular disease
3	58	38	71	65	Stress response, muscle development protein synthesis
4	28	17	94	44	Cell cycle and cell death
5	41	30	80	61	Gene expression cellular maintenance
6	28	19	84	40	Muscle and connective tissue development regulation of RNA expression
7	40	26	88	59	Cell death
Total	272	182 (67%)	mean is equal to 83%		

The list of genes for each cluster was submitted to IPA software and only one biological network was obtained. The eligible genes are those with a gene name accepted by IPA for having biological functions. An average of 83% of the eligible genes were included in the same network. IPA gives also the top biological functions associated with each cluster.