npj | systems biology and applications

Article

Published in partnership with the Systems Biology Institute

https://doi.org/10.1038/s41540-025-00522-0

Patient-specific gene co-expression networks reveal novel subtypes and predictive biomarkers in lung adenocarcinoma

Patricio López-Sánchez[1], Federico Ávila-Moreno[2,3], Enrique Hernández-Lemus[1], Marieke L. Kuijjer[4] & Jesús Espinal-Enríquez[1]

Lung adenocarcinoma (LUAD) is a highly heterogenous and aggressive form of non-small cell lung cancer (NSCLC). The use of genome-wide gene co-expression networks (GCNs) has been paramount to describe changes in the transcriptional regulatory programs found between diseased and healthy states of LUAD. Recently, studies have shown that multiple cancerous phenotypes share a distinct GCN architecture, suggesting that network topology holds promise for understanding disease pathology. However, conventional GCN inference methods struggle to capture the inherent context-specificity within a patient population, thus flattening its heterogeneity. To address this issue, the use of single-sample network (SSN) modelling has emerged as a promising solution into studying heterogeneous traits of cancer through network-based approaches. Here, we reconstructed patient-specific GCNs (n=334) using the LIONESS equation and mutual information as the network inference method. Unsupervised analysis revealed six novel LUAD subtypes based on inter-patient network similarity, each with distinct network motifs reflecting unique biological programs. Supervised analysis, employing regularized Cox regression, identified 12 genes (CHRDL2, SPP2, VAC14, IRF5, GUCY1B1, NCS1, RRM2B, EIF5A2, CCDC62, CTCFL, XG, and TP53INP2) whose weighted degree in SSNs is predictive of patient survival in LUAD. These findings suggest that topological features of SSNs offer valuable insights into the context-specific nature of LUAD malignancy, highlighting the potential of SSN-based approaches for further research.

Lung cancer remains the most fatal and second most frequently diagnosed cancer globally. In 2020, it constituted 18% of cancer-related deaths and was responsible for 11.4% of new cancer cases[1].

Lung cancer can be classified into two main sub-types: non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC), with NSCLC accounting for 85% of all newly diagnosed cases. Adenocarcinomas and squamous-cell carcinomas are the two most common subtypes of NSCLC, constributing to 50% and 30% of all NSCLC cases, respectively[2]. Among NSCLCs, squamous cell carcinomas often arise in the central bronchial tubes and express molecular markers of squamous cell differentiation. In contrast, adenocarcinomas typically develop in the lungs' periphery and are frequently associated with the formation of glandular bodies and high mucin production. Both cancer subtypes exhibit significant differences in clinical associations such as gender, age, smoking history, TNM clinical stages, 5-year survival rates post-diagnosis, and treatment options, among others[3].

Lung cancer cells exhibit aberrant regulatory programs at multiple physiological scales, ranging from their mechanisms of genetic and epigenetic regulation to the interaction between the tumor tissue and its surrounding environment[4]. For instance, genetic alterations in a set of proto-oncogenes such as KRAS, EGFR, BRAF, PI3K, MEK, and HER2 have been identified to severely dysregulate various signaling pathways crucial for the

[1]Computational Genomics Division, National Institute of Genomic Medicine, Mexico City, Mexico. [2]Facultad de Estudios Superiores-Iztacala (FES-Iztacala), Universidad Nacional Autónoma de México (UNAM), Mexico State, Mexico. [3]Subdirección de Investigación Básica, Instituto Nacional de Cancerología (INCAN), Mexico City, Mexico. [4]Centre for Molecular Medicine Norway (NCMM), University of Oslo, Oslo, Norway. e-mail: jespinal@inmegen.gob.mx

control of cell proliferation, apoptosis, and other cellular functions[5]. The disruption of these mechanisms directly influences the phenotypic transition from healthy to cancerous tissue. Therefore, there is a growing motivation to study the molecular biology of lung cancer, employing a wide variety of multi- and interdisciplinary approaches with the aim of developing new strategies for prevention, early diagnosis, and treatment.

Within lung cancer, each subtype produces a large number of distinct pathological signatures, characterized by the differential interactions between the components in the system. To study these complex diseases, systems biology has sought to model these interactions through network-theoretical approaches. An example of this, and the main focus of this work, is gene co-expression networks (GCNs). GCNs are undirected graphs where nodes represent genes, and edges represent a statistically significant dependence between their expression levels[6]. The importance of GCN analysis in biomedicine lies in its ability to model interactions among genes expressed in a tissue, thus illustrating the landscape of their transcriptional program in healthy and diseased states.

Recent studies have revealed clear underlying topological differences between networks constructed from samples of cancerous tumors and those from adjacent healthy tissue. For example, a clear rupture of the largest connected component into smaller communities significantly enriched with genes physically close in the genome has been observed in various malignant tumors, including Luminal A, Luminal B, HER2-positive, and basal breast cancer[7,8], clear cell renal carcinomas[9], NSCLCs[10], and more recently in hematological cancers[11], thus suggesting that this is a common trait of cancer. Furthermore, comparative analyses of these networks have shown that differential gene expression alone is not capable of explaining the structural rewiring observed in the co-expression landscape of tumor derived networks[9]. Together, these results highlight that genome-wide GCN topology is intrinsically associated with cancer development. However, understanding how global or local network structural traits can be linked to the underlying disease's biology remains a topic of wide discussion.

Given that GCNs capture statistical dependence between genes, network inference algorithms require multiple samples to construct a robust aggregate network that accurately represents a sample population or cohort. One of the major limitations of these methods is that they can only represent characteristics shared by the chosen set of background samples, therefore losing relevant information about the specific context and environment of each sample in the process. Thus, while extremely useful at mining shared traits in a biological phenomenon, aggregate networks are unable to capture the underlying heterogeneity that exists within the population of interest.

To better understand how phenotypic complexity varies across individual samples from a network-driven perspective, the LIONESS method (Linear Interpolation to Obtain Network Estimates for Single Samples) for estimating sample-specific networks (SSNs) was developed[12]. This method is based on the assumption that aggregate networks are the result of a linear combination of individual sample networks. As a first step, the LIONESS equation contrasts two input networks: one built on data from all samples and another excluding data from a single sample. This step essentially calculates the contribution of the removed sample to the aggregate network. The estimated contribution is then normalized to account for sample size and added to the aggregate network without the sample of interest, effectively estimating that sample's individual network. This process can then be repeated iteratively to infer a unique network for each sample within the population.

While it has been shown that LIONESS-based networks can effectively stratify patients based on regulatory differences beyond differential expression[13], previous work has primarily focused on transcription factor-to-gene bipartite regulatory networks. Notably, large-scale single-sample co-expression networks, particularly those based on mutual information (MI), have not been explored in the literature. This gap is due to both the extreme size of co-expression networks and the computational complexity of MI calculation compared to linear correlation methods. Thus, by enabling analysis at the individual sample level, LIONESS-based single-sample GCNs could offer a deeper understanding of how variations in gene-correlation patterns might influence disease progression and patient outcomes.

Here, by using LIONESS, we have inferred patient-specific GCNs of the LUAD-TCGA cohort using MI as the network reconstruction algorithm, and subsequently developed two parallel frameworks for their analyses. The first approach consisted of a graph-clustering strategy whereby focusing on communities of patients based of their on SSN-similarity, we explored the significant relationship between conserved network motifs and LUAD clinical, molecular and cellular contexts. The second one, is focused in node weighted degrees; we developed a regularized regression scheme to extract genes whose structural importance in a SSN is predictive of LUAD overall survival. Finally, we used random resampling methods in order to ascertain the robustness of our findings and compared our results with those obtained using gene expression data alone.

Results

Most patient clusters based on single-sample network similarity are not associated with clinical phenotypes

As a first exploration of the extent to which SSN topology is capable of defining heterogeneous LUAD phenotypes, we investigated whether patients sharing a substantial number of their most important edges also shared similar clinical outcomes. Figure 1 shows the presence of the community structure inside the SSN-similarity graph. Notably, Cluster 1 displays the subset of SSNs that have the highest resemblance to each other. On the other hand, Cluster 4 follows the opposite trend, where networks exhibit low similarity not only with those outside their community but also among themselves.

To ensure that the observed pattern of community structure was not dependent on the selected threshold of 10,000 edges, we performed the same clustering strategy on sets of SSNs filtered to their top 50,000 and 100,000 edges. Furthermore, we verified that the resulting network-based patient communities were distinct from those obtained from standard gene expression clustering using k-means. Supplementary Fig. 1 shows the overlap between the detected communities of SSNs at different network densities, as well as the alignment of clusters obtained using gene expression data alone. As expected, clusters that were highly conserved such as clusters 1, 2, 3, 5, and 6 were concordant with communities detected at varying SSN densities. Contrarily, cluster 4 was prone to be sub-clustered when considering additional interactions. On the other hand, clusters derived from gene expression data showed a poor alignment with the rest of the SSN-similarity groups, showcasing that clustering strategies using patient-specific networks convey new information about the active regulatory programs.

Next, we conducted overrepresentation testing within patient clusters to determine if clinical outcomes could be discerned from network similarity alone. Factors considered included gender, disease stage, tissue of origin, tumor size and extent (T), lymph node involvement (N), and presence of distant metastasis (M). Interestingly, Cluster 1 and Cluster 5 showed a significant enrichment (Fisher's exact test, adjusted p-value < 0.05) of T1 tumors, while Cluster 5 showed an overrepresentation of female patients. In addition, Cluster 2 was enriched with non-metastatic M0 samples (red marques, Fig. 1 & Supplementary Fig. 2-A).

We conducted the same over-representation test on gene-expression-derived clusters. We found that Cluster 9 was also enriched with T1 and Stage I neoplasms of female patients (Supplementary Fig. 2B). As shown in Supplementary Fig. 1, this group is split mainly into Clusters 1 and 5 in the SSN-similarity communities. This suggests that co-expression differences captured in these communities are capable of further stratifying patients with similar clinical and expression profiles.

Furthermore, we compared the survival distributions between clusters to investigate if the resulting patient communities could be associated with survival. No significant differences (log-rank test p-value < 0.05) were observed (Supplementary Fig. 2C). Similarly, no significant differences were observed within gene expression-derived clusters (Supplementary Fig. 2D).

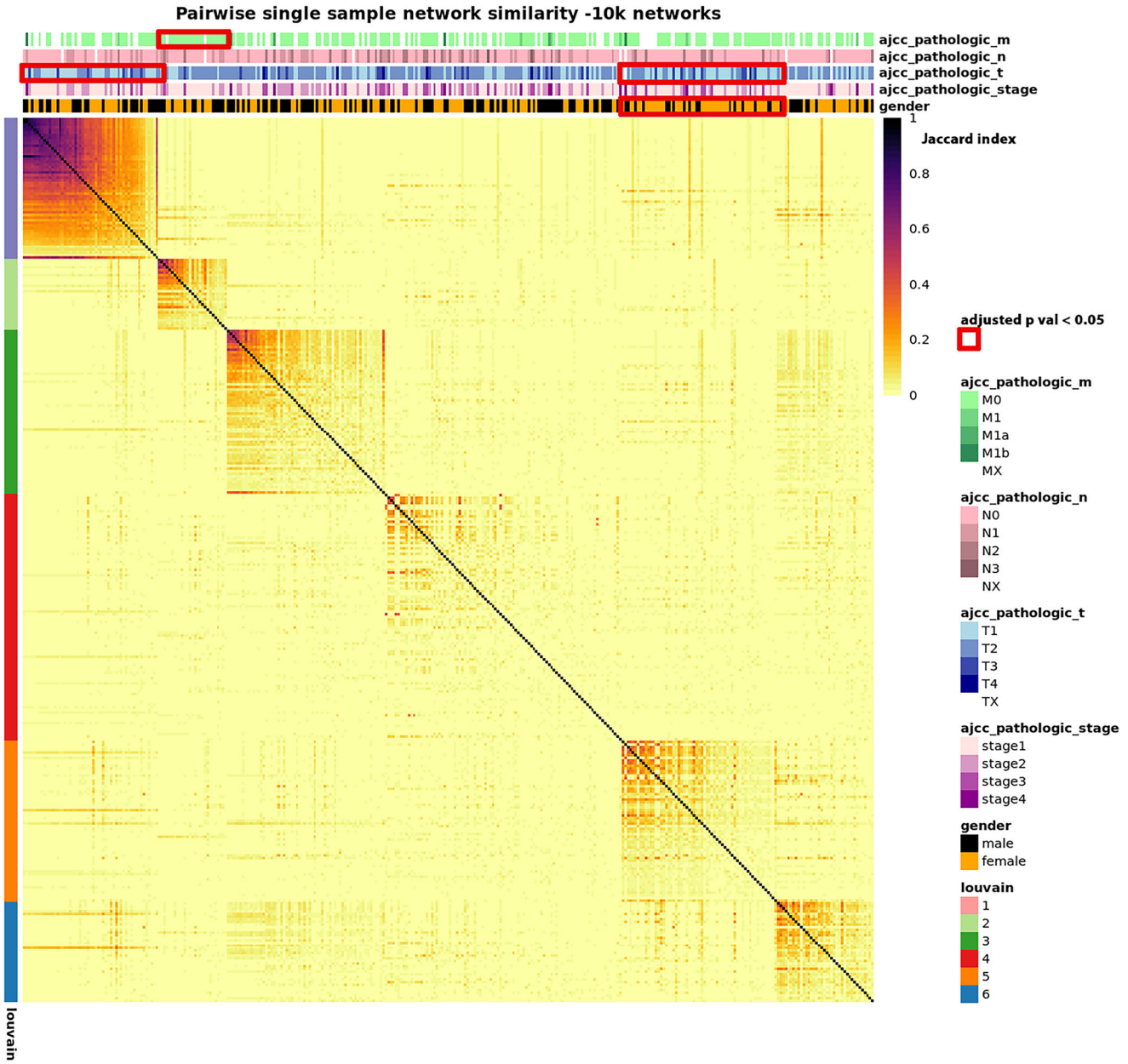

Fig. 1 | Similarity matrix between patient-specific networks. The top 10,000 edges of each network were used to calculate all pairwise Jaccard indices. The resulting weighted adjacency matrix was clustered using the Louvain algorithm. Row annotations show the clustering results. Column annotations show the different clinical labels associated to each patient. Red marques show the significant enrichment (Fisher's exact test adjusted p-value < 0.05) of a clinical feature inside a cluster.

Per-cluster consensus networks are associated with cell and tissue context

To further characterize the biology behind the formation of patient communities based on SSN similarity, we built six consensus networks (one for each group) by identifying the most conserved set of interactions in each case (see Methods). Given that a percentile-based statistic was used to determine if an interaction is conserved, consensus network size and density largely depend on the range of possible edges within each cluster. Supplementary Table 1 contains the structural summary statistics of each consensus network.

It can be observed that the number of nodes, edges, and connected components varies accordingly with the degree of similarity shared between SSNs in a cluster, hampering downstream comparative analyses. In consequence, to facilitate the annotation of all consensus networks, we focused solely on each largest connected component to perform subsequent functional enrichment analysis (Fig. 2).

Figure 3 depicts the biological processes significantly associated with each cluster (Fisher's exact test, adjusted p-value < 0.05). It is possible to note that, except for Cluster 3 and Cluster 6, each patient community is enriched by unique biological programs: Cluster 1 is associated with processes related to cilium movement and structure. Cluster 2 is associated with processes regarding chromatin remodeling and histone modification. Cluster 4 is enriched with angiogenic and matrix remodeling processes. Cluster 5 is strongly associated with cell cycle progression and regulation. Finally, Cluster 3 and Cluster 6 share multiple enriched processes related to the activation and regulation of the immune response.

Considering that, despite having dissimilar networks, Cluster 3 and Cluster 6 shared the vast majority of overrepresented immune system-related biological processes, we sought to evaluate whether this phenomenon could be caused by underlying differences in tumor cell composition. We used the R implementation of xCell to estimate cell enrichment scores for 64 immune and stromal cell types from expression data.

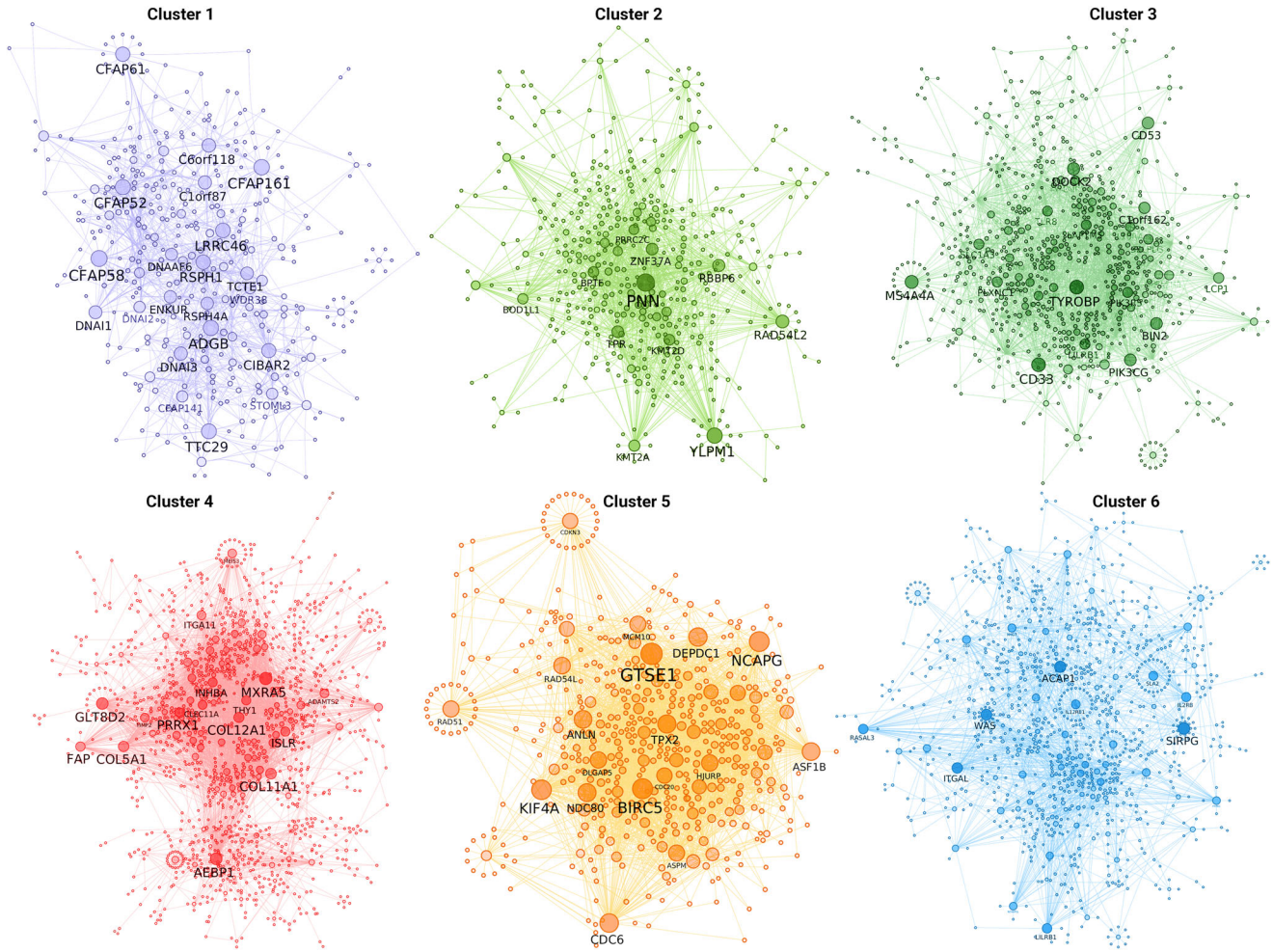

Fig. 2 | Largest connected components of per-cluster consensus networks. The top 1% most frequent edges in each cluster were aggregated to build the corresponding consensus network. Node and label size correspond to the number of incident edges (degree).

We identified 18 immune cell types and 11 stromal cell types with significantly different distributions (Kruskal-Wallis H, adjusted p-value < 0.001) in at least one patient cluster (Fig. 4A). A post-hoc Dunn test of xCell's ImmuneScore shows that, in general, the immune cell composition of Cluster 3 and Cluster 6 is inversely correlated (Fig. 4B), with Cluster 3 being the most immune-depleted group of patients, while Cluster 6 displays the most infiltrated group of tumors. Analogously, the StromalScore distribution of Cluster 3 follows a similar trend by displaying the lowest degree of infiltration (Fig. 4C). In addition, the effect sizes for both cell-score differences highlight that not only are the most extreme differences in immune and stromal composition found between Cluster 3 or Cluster 6 and any other group (Fig. 4D, E, respectively), but also that larger inter-cluster changes in the tumor microenvironment are mainly driven by the immune milieu. Figure 4F summarizes the overall distribution of the individual immune and stromal cell types with significant differences across clusters.

It is worth noting that when performing the same analyses on gene expression-derived clusters, we were able to observe that the same cell types vary across groups of patients (Supplementary Fig. 3A). However, examining the distribution of each cell type across these clusters revealed that the overall cell enrichment signature differed from that seen in SSN-similarity communities (Supplementary Fig. 3B–E). For example, the gene expression-derived clusters display multiple groups with varying levels of immune cell infiltration, ranging from low to high. In contrast, the SSN-based clusters predominantly captured a single low-infiltration and a single high-infiltration group, thus suggesting that SSN-based clustering could be more sensitive to detecting cell marker genes' expression co-variation.

Weighted Gene Degrees capture structural changes associated with cell proliferation

As a way of summarizing the role that an individual gene plays in the overall structure of each SSN, we calculated the weighted degree of each gene as the sum of the weights of all incident edges. To investigate whether specific hallmarks or biological processes rewire their connectivity the most, we conducted gene set enrichment analysis on the standard deviation value of each gene's weighted degree. Noticeably, 5 hallmarks associated with cell proliferation, (MYC Targets V1, MYC Targets V2, E2F Targets, G2M Checkpoint, and Mitotic Spindle) are among the categories with the highest enrichment scores (Fig. 5A).

Comparably, when looking for enriched Biological Processes, we observed a similar pattern where multiple programs related to cell proliferation, were significantly enriched (Fig. 5B). These results suggest that the network of genes regulating cell cycle progression and replication is under constant structural change. Additionally, weighted gene degrees, as a topological feature of GCNs, reflect biological phenomena and provide a stable foundation for downstream analyses.

Predicting overall survival of lung adenocarcinoma patients using Weighted Gene Degrees

After having used unsupervised strategies to mine associations between SSN topology and lung adenocarcinoma heterogeneity, we set out to employ a supervised learning approach to predict each patient's overall survival using network based features. Specifically, we used WGDs as input for a lasso regularized Cox proportional hazards regression model. In general, we

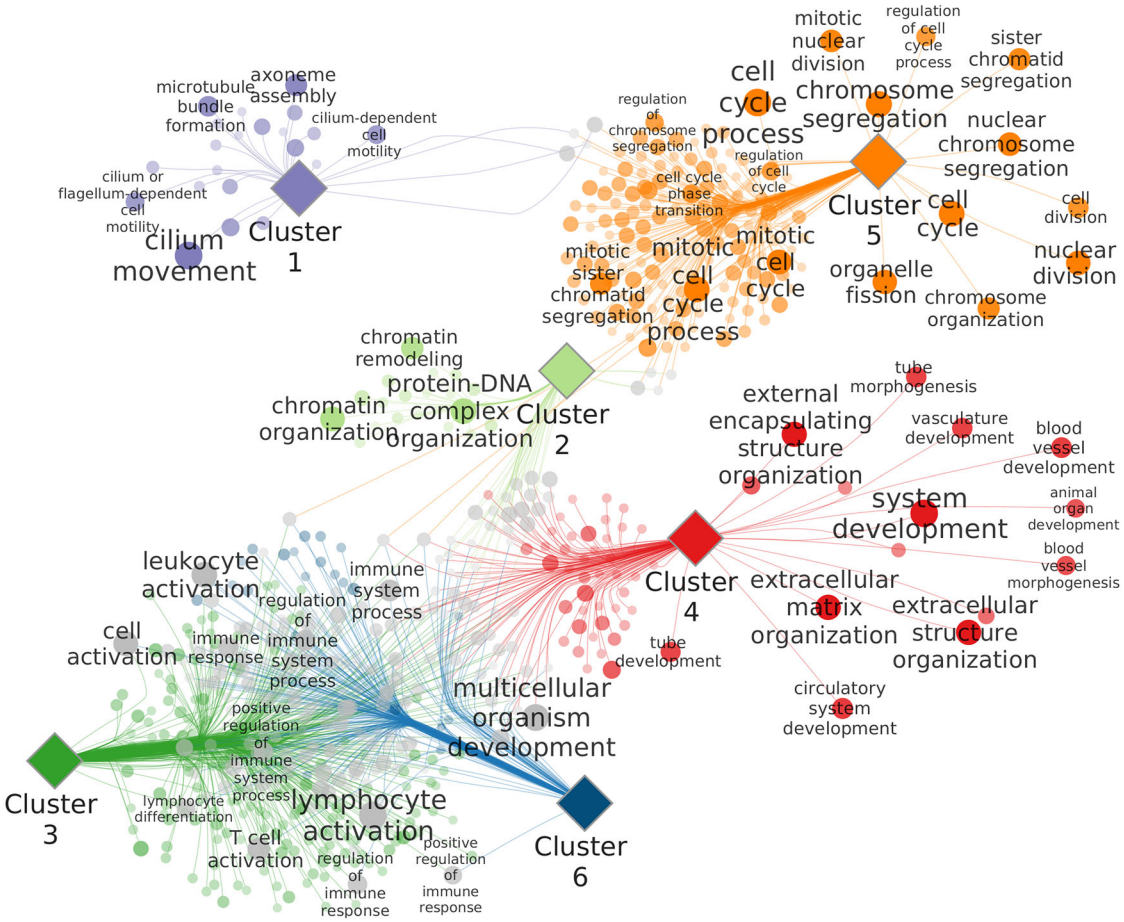

Fig. 3 | Bipartite network of enriched Biological Processes in each consensus network. Diamond-shaped nodes represent different clusters of patients, while circular-shaped nodes represent a Biological Process enriched in its corresponding network's largest connected component. Node and font size and transparency vary according to the scaled -log(adjusted *p*-value) in the enrichment test. Grey-colored nodes indicate Biological Processes that are enriched in more than one Cluster, while nodes with the same color as their neighbor Cluster node represent a Biological Process specific to that group.

benchmarked three different modeling scenarios: Model L, where lasso was applied to the full set of WGDs; Model U+L, where lasso was applied to a set of genes previously selected by a univariate filter; and Model U+L+R, which involved a relaxed fit of the same model. We then compared the overall performance of models built with WGDs to those constructed using only gene expression data, aiming to assess the prognostic potential of network-based features.

To compare the global performance of the three above-mentioned regression models, we computed both Harrel's C-Index as well as the Integrated Time Dependent Brier's Score for 100 models built using different training/testing data partitions (70% and 30% split, respectively). Figure 6 A highlights the obtained C-Index values across the three frameworks. As expected, using a univariate Cox regression as a filtering step prior to lasso shrinkage significantly improved the predictive performance compared to the model using the full WGD matrix (Wilcoxon test, p-value < 0.05). On the other hand, models U+L and U+L+R show very similar predictive capabilities, with the unrelaxed fit showing a slightly higher median C-Index. Conversely, when assessing each model's performance through the Integrated Brier's Score statistic, no significant differences were found (Fig. 6B). Overall, both metrics for evaluating survival predictions show that the three models perform better than at random in the majority of scenarios. Furthermore, when comparing the resulting models trained on WGDs with those built on gene expression data, we found that using WGDs outperformed gene expression data in the three different regularized regression schemes (Wilcoxon test, p-value < 0.05; Supplementary Fig. 4).

One of the main advantages of using regularized regression strategies is their ability to perform feature selection while reducing multicollinearity in high dimensional case scenarios such as those commonly found in "omics" data sets. However, given how lasso selects a random representative variable among a group of correlated predictors[14], dissimilar sets of genes may be selected in different iterations of the model fitting process. Thus, we consider a gene's weighted degree to be stably associated to survival only if it is selected by both the univariate filter and lasso in at least 50 of the 100 random data partitions. Notably, a group of 12 genes were selected to be stably associated with patient survival, with Chordin Like 2 (CHRDL2) being the gene most frequently picked, followed by a tie between VAC14 Component of PIKFYVE Complex (VAC14) and Secreted Phosphoprotein 2 (SPP2) (Fig. 6C). Supplementary Fig. 5 shows the association of all stable genes' weighted degrees with overall patient survival.

Additionally, we focused on a representative U+L+R model to assess overall goodness of fit and local predictive performance on a discrete scale, as it was the best performing model among the three frameworks in a random iteration. The obtained model contained 65 genes, of which 7 were considered stable by the resampling procedure described above. Among the stable features, CHRDL2 and SPP2 are the two genes with the largest negative and positive coefficients, respectively (Fig. 7A).

To validate the sign of association between both genes weighted degree and survival, we stratified the full data set by their median weighted degree. In both cases, the groups with high and low weighted degrees correspond

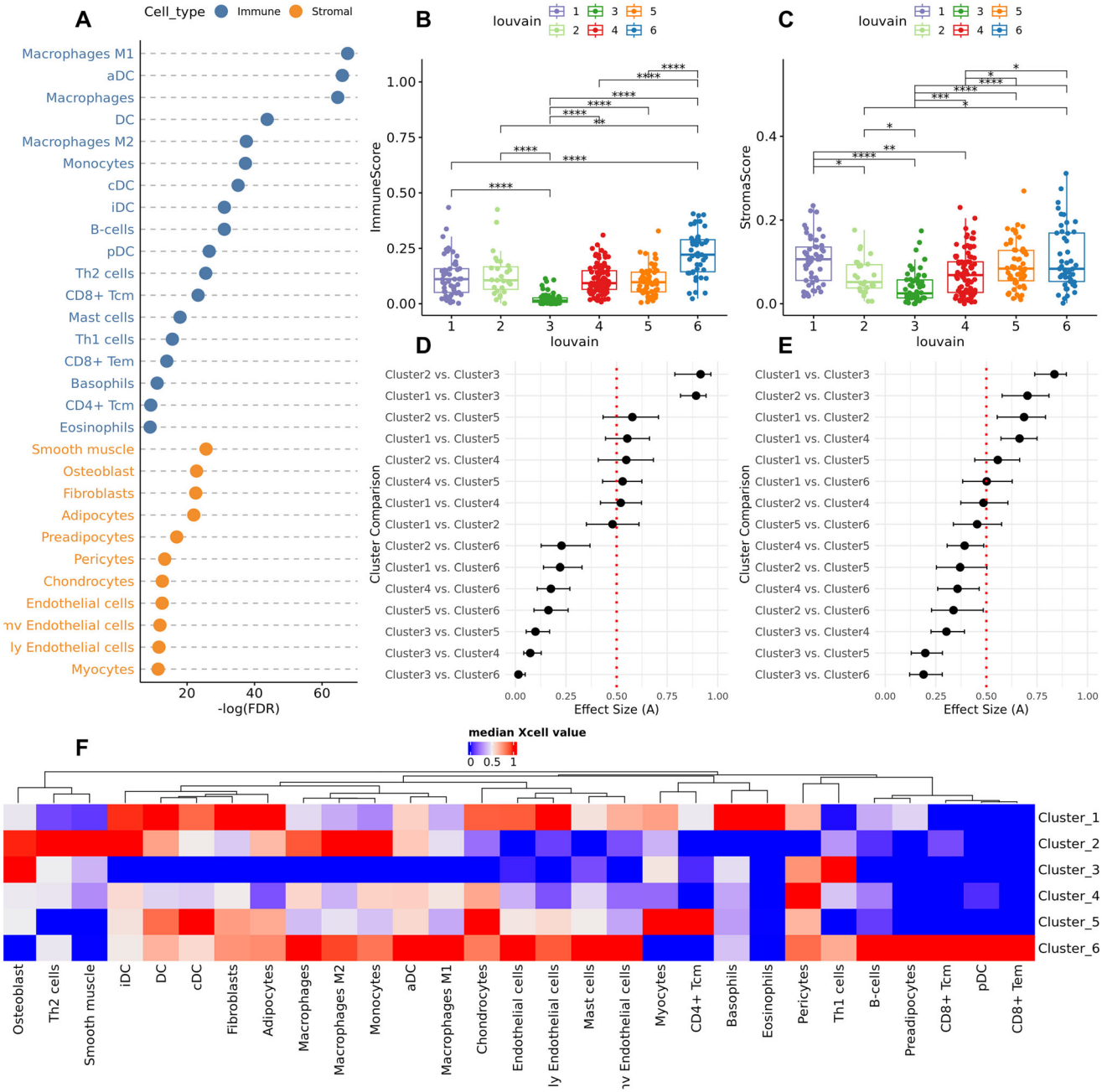

Fig. 4 | Differences in cell enrichment scores across network-derived clusters. **A** Dot-chart of Kruskal-Wallis test results comparing xCell's immune and stromal cell type enrichment scores between patient communities. The significance threshold was set at adjusted p-value < 0.001. **B**, **C** Distribution of estimated ImmuneScore and StromalScore values, respectively. Pairwise significant differences were calculated through post-hoc Dunn testing. **D**, **E** Forest plot depicting the effect sizes of all pairwise inter-cluster comparisons of ImmuneScore and StromalScore, respectively. Effect sizes were calculated using Vargha and Delaney's A statistic. Dotted red lines indicate the value where there are no differences between groups (i.e., points farther away from the line indicate more extreme differences). Error bars represent the 95% CI obtained by 1999 bootstrap iterations. **F** Heatmap of median Xcell values for all cell types with significant differences obtained in the Kruskal-Wallis test.

with the sign of the coefficients obtained in the model and show significant differences in overall survival probabilities (Fig. 7D, E). When evaluating the full model, we observed a significant difference in the survival distribution of both risk groups obtained by stratifying the linear predictors of the unseen test data (Fig. 7B). Moreover, the assessment of 1-, 2-, 3-, and 5-year overall survival prediction yielded AUC scores of 0.63, 0.63, 0.64, and 0.67, respectively (Fig. 7F). Finally, the time-dependent Brier score on all possible time-points showed a progressive loss of predictive performance over time, reaching an asymptotic behavior after the 1000-day mark, likely due to the lack of events with longer time spans (Fig. 7C).

Discussion

This work aimed to characterize the link between GCN topology and heterogeneous phenotypes of patients with lung adenocarcinoma. To this end, we modelled single sample GCNs for patients in the LUAD cohort of the TCGA data base and analyzed them through both unsupervised and supervised methods. Here we focused on patient-specific network similarity and node weighted degree centrality as the topological features used to identify patterns in the co-expression landscape at a population level.

When focusing on shared gene interactions, a clear community structure can be observed in the patient-similarity graph. As previously

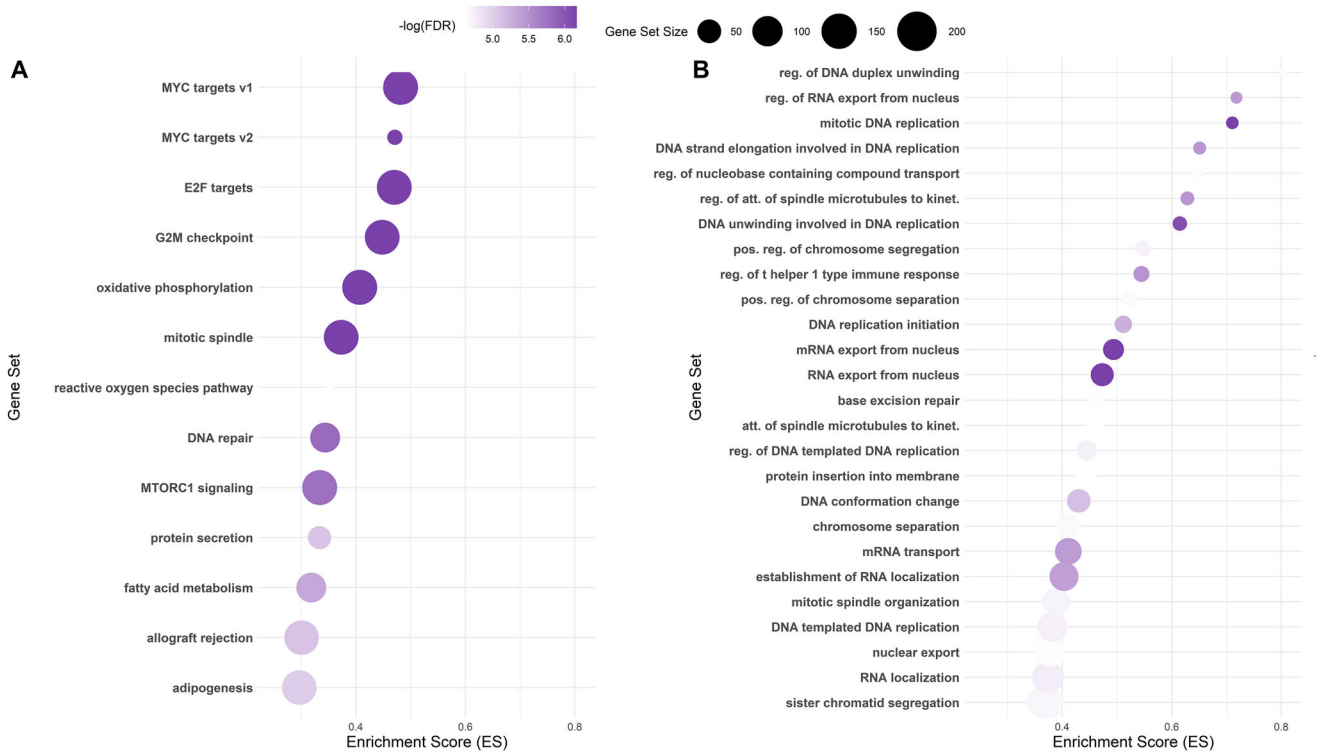

Fig. 5 | Gene Set Enrichment Analysis (GSEA) results of significant (FDR < 0.01). A Hallmark signatures, and **B** Biological Processes, according to the standard deviation of each gene's weighted degree across all samples.

mentioned, differing clinical backgrounds were not sufficient to explain the observed grouping. Nevertheless, Cluster 1 showed a significant over-representation of T1 tumors (i.e., growths ≤ 3 cm surrounded by lung or pleura with no local invasion[15]). Concomitantly, Cluster 1 had the consensus network with the highest degree of conservation among its patients and displayed enriched functional processes associated with cilium genesis, structure, and maintenance. Ciliated cells are considered the most predominant cell type in the lung airway and the most terminally differentiated, showing very limited proliferative potential[16]. Furthermore, ciliated cells have been proposed as one of the cell types capable of forming precursory lesions that trigger the origin of pulmonary LUAD tumors[17]. Together, this suggests that tumors sampled from patients in Cluster 1 are likely to represent a cancerous phenotype more indicative of the earlier stages of progression compared to the rest of the cohort, as the SSNs of these patients illustrate the regular transcriptional activity that is most conserved in ciliated lung epithelia.

On the other hand, networks in Cluster 5, which showed a significant overrepresentation of T1 neoplasms and female patients, are primarily composed of genes involved in multiple biological processes related to cell cycle regulation and progression. Gender biases have been observed in lung cancer development, response to treatment, and mortality[18]. Specifically, the link between female sex hormones and cell proliferation programs in cancer has been previously reported. For example, not only is estrogen capable of inducing cell proliferation in lung cancer, but it can also be actively produced by NSCLC cells[19]. Conversely, progesterone has been shown to inhibit lung cancer growth in mice, while Progesterone Receptor-positive (PR+) NSCLCs in female patients frequently present an inverse correlation between PR activity and overall disease progression[20]. Additionally, sex-biased genetic factors are well established to play a role in the progression of lung cancer. For instance, female patients are more likely to have driver mutations in genes such as EGFR or KRAS[21]. We hypothesize that our analysis suggests the presence of a conserved gene co-expression signature associated to tumor growth and proliferation, predominantly among female patients, that likely reflects the cumulative impact that heterogenous factors contribute to gender differences in LUAD patient outcomes.

Next, owing to the shared biological programs associated with consensus networks of Cluster 3 and Cluster 6, we identified significant differences in the overall tumor microenvironment composition across all patient communities. Above all, Cluster 3 consistently showed a depletion of both immune and stromal cell infiltrates, whereas Cluster 6 displayed the opposite pattern, with frequent high levels of immune infiltration.

It is important to note that large effect sizes in immune infiltration are unique to Cluster 3 and 6. While some differences in stromal components are apparent for the remaining groups, their effect sizes are low, suggesting that the clustering results are not solely driven by cellular composition but reflect broader dysregulations in cancer gene-networks.

Notwithstanding, our analysis was not able to identify any significant association between the groups with high and low immune infiltration and patient clinical outcomes. This is likely due to the fact that the relationship between immune infiltration and patient prognosis varies in a context specific manner, thus reflecting the functional plasticity displayed by immune system cells in the tumor microenvironment. However, the establishment immunosuppressive phenotypes are known to be a crucial event that favours tumor malignancy. Hence, the identification of clusters of patients with diverse immune system activity highlights the potential of SSNs as a tool for probing the network of genes that drive immunosuppressive states, thereby warranting further research.

Since unsupervised clustering strategies based on edge-level information from sample-specific networks failed to stratify patients by overall survival, we shifted our focus to supervised analyses of node-level structural information.

Of note, previous studies have tried a similar approach wherein the degree of nodes in correlation-based SSNs (i.e., co-expression networks) is used as a summary statistic to uncover biologically relevant associations. For example, Chen et. al. performed SSN modelling using the SWEET method (a SSN inference method based on Pearson Correlation and genome-wide sample weights) and built a Network Degree Matrix to study whether SSN

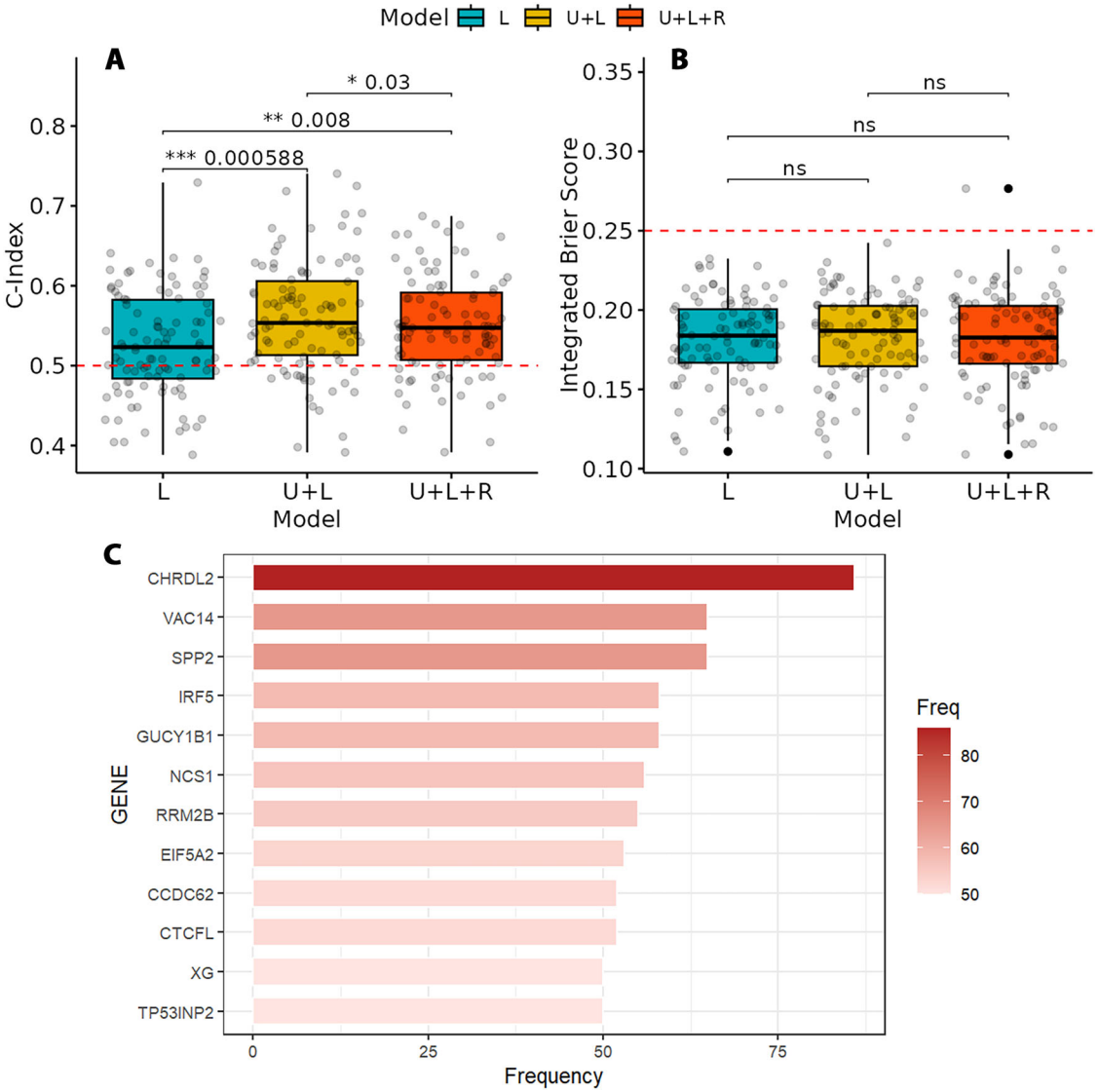

Fig. 6 | Overall predictive performance, uncertainty, and stability of lasso gene selection in 100 random training/testing data partitions. 3 modelling frameworks were contrasted: Model L, where lasso is applied to the entire set of genes, Model U+L, where a univariate filter is applied a priori to lasso, and model U+L+R, where the fit obtained on the U+L model is relaxed (known as relaxed lasso). **A**, **B** Overall performance and uncertainty measured by (**A**), Concordance Index (C-Index), and (**B**), Integrated Brier Score. Red dotted lines mark the score in performance where model prediction are indistinct from a random guesser. Performance differences were calculated through Wilcoxon signed rank tests. **C** Set of genes considered stable predictors of survival. Each gene is plotted against the frequency with which they were selected by both the univariate and lasso filters in the random resampling procedure.

degree distributions were correlated with genetic dependency scores of 580 cell lines. More specifically in the LUAD-TCGA cohort, they utilized the same approach to discover "dark" genes with degree differences between two defined subtypes that were not explained by differential expression analysis[22]. Another example of a similar network-analytical approach was employed by Dai et al., whereby cell-specific networks built using a mutual dependency criterion to establish gene-gene correlations were further summarized by calculating a Network Degree Matrix for downstream analyses, analogous to standard single-cell RNA-Seq pipelines[23].

To our knowledge, this work is the first to employ the same network analysis strategy using LIONESS-based fully-connected mutual information networks.

Working under the driving hypothesis that SSN weighted degrees are enriched with biological information, we sought to investigate whether they could further be predictive of LUAD patient survival. To this end, we used gene weighted degrees as input for regularized Cox proportional hazards regression approaches and validated the model building procedure and performance through random resampling methods. The proposed methodology highlighted not only that weighted degrees generally outperformed gene expression data in terms of predictive power, but also that 12 genes whose structural importance in individual GCNs is robustly associated to survival, regardless of the established training and testing data partition. Notably, all 12 genes (CHRDL2, SPP2, VAC14, IRF5, GUCY1B1, NCS1, RRM2B, EIF5A2, CCDC62, CTCFL, XG and TP53INP2) have been previously associated to various types of cancer[24–35], with 7 (CHRDL2, SPP2, IRF5, NCS1, RRM2B, EIF5A2, CTCFL) including reports specific to lung cancer[25,36–41].

It is worth mentioning that the functional assessment of sets of genes selected by lasso based on gene enrichment analysis may be hindered by the removal of collinearity between variables in the model fitting step. However, there is an interesting overlap in the body of literature that places these genes in the context of cancer research. For example, CHRDL2 and SPP2, the two most stable predictors of survival in this analysis, have been shown to mainly operate as antagonists of Bone Morphogenetic Proteins (BMPs), a group of

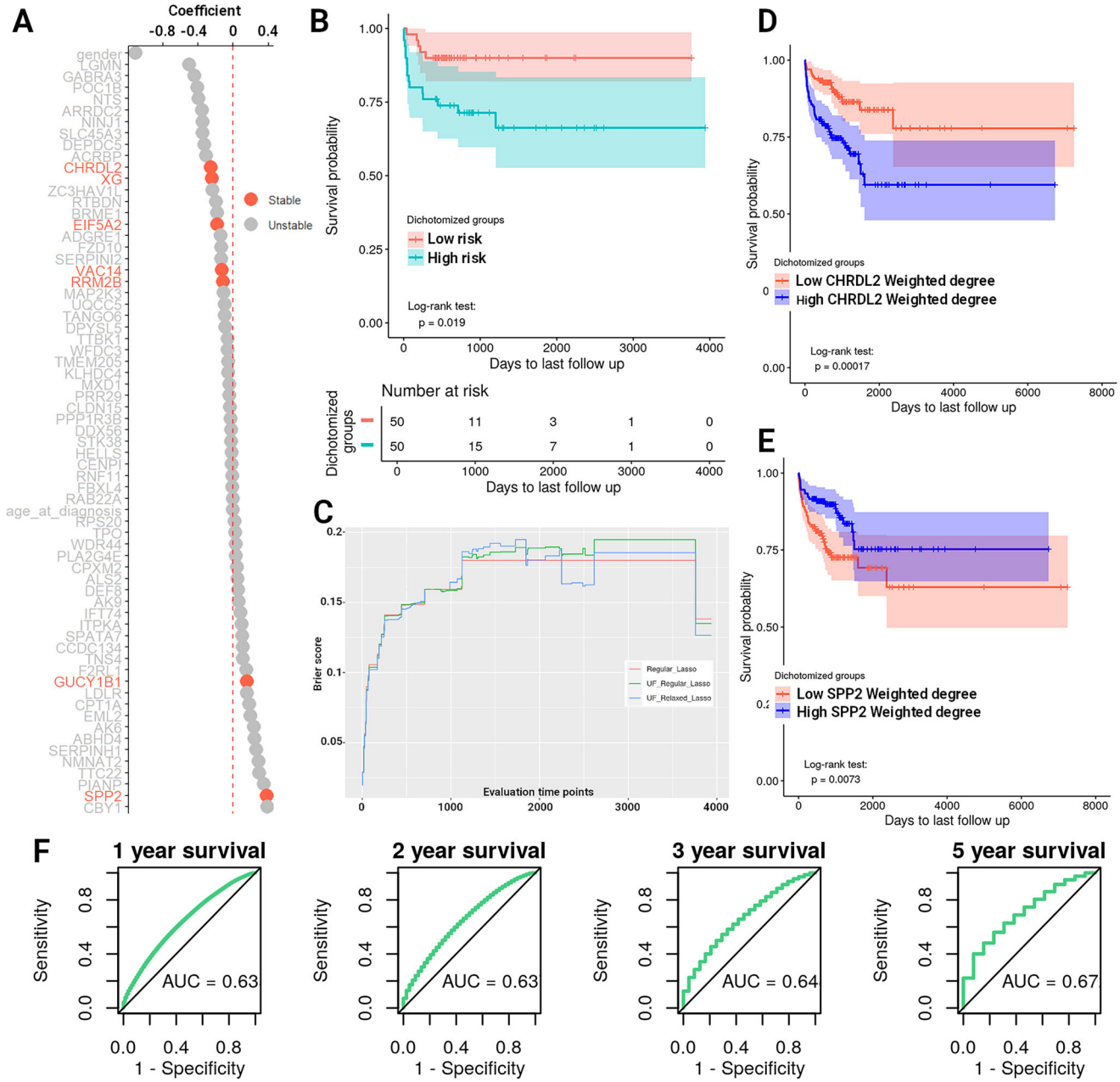

Fig. 7 | Model diagnostics of a representative relaxed model. A Coefficient plot of 65 genes selected after univariate and lasso filtering. Genes highlighted in orange were considered to be stable predictors. **B** Kaplan-Meier plot of the stratified risk scores predicted for the test data set using the relaxed fit. **C** Time-Dependent Brier Score curve for all available time points in the test data set. **D**, **E** Kaplan-Meier plots of CHRDL2 and SPP2 stratified by their median weighted degree in the entire cohort. **F** AUC scores for prediction of 1-,2-,3-,5- year overall survival in the test data.

growth factor cytokines that belong to the transforming growth factor β (TGF-β) superfamily of proteins. While the baseline interactions between CHRDL2/SPP2 and the BMP signalling pathway have been more extensively described in the context of bone development and homeostasis, multiple studies have linked both genes to cancer-promoting processes mediated by BMP signalling. For example, overexpression of CHRDL2 has been shown to promote proliferation and metastasis in osteosarcoma tissues and cell lines by blocking the BMP-9/PI3K/AKT pathway[42]. Similarly, in colorectal cancer, CHRDL2 overexpression promotes cell proliferation and inhibits apoptosis by blocking BMP2-driven Smad1/5 signaling[43]. Conversely, BMP2 blockade via SPP2 administration has been shown to inhibit tumor cell growth in hepatocellular carcinoma[44], osteosarcoma[45], prostate cancer[46], pancreatic cancer[47] and lung cancer[48].

The role of BMP signaling in NSCLC growth[49], and more recently in LUAD metastasis and cell differentiation[50] has been demonstrated to be of great relevance. However, added layers of complexity are known to take part in the BMP-induced carcinogenic process. One such case is the crosstalk between BMP, TGF-β and WNT/β-catenin pathways, which together contribute to the formation of cancer-related phenotypes[51]. We hypothesize that other genes highlighted in our analysis, such as EIF5A, TP53INP2 and CTCFL/BORIS are reflective of this crossroad between signalling routes. For instance, EIF5A has been shown to promote epithelial to mesenchymal transition (EMT), treatment resistance and metastasis through direct dysregulation of TGF-β signaling in a variety of cancers[52–55]. TP53INP2, an autophagy-related protein, was observed to modulate EMT, migration and invasion of bladder cancer cells[56], as well as the malignant progression of

colorectal cancer[35] through β-catenin dependent signalling. In parallel, CTCFL/BORIS, a crucial regulator of the malignant effects of cancer stem cells, is known to primarily exert its tumorigenic actions through both Wnt/β-catenin and NOTCH pathways[57]. However, TGF-β dependent signalling of CTCFL/BORIS has also been observed to be relevant in melanoma invasiveness[58], and neuroblastoma migration[59].

Together, given that SSN analysis is capable of representing traits of the population's underlying heterogeneity as structural features of transcriptome-wide GCNs, we believe that the underscored set of genes associated to survival indeed portray the complex and context-specific nature of the transcriptional regulation that BMP, TGF-β and WNT/β-catenin signaling pathways may exert to influence LUAD clinical outcomes. However, further research needs to be addressed to support this hypothesis.

Cancer is a putative heterogenous disease. A myriad of studies addressing the discrete relationship between clinical-, cellular- and molecular-context specificity have motivated the development of new analytical strategies capable of embracing such phenotypic complexity. To this end, the use of patient-specific GCNs could aid bridge the gap between the studies of minute changes in transcriptional activity/regulation and cancer heterogeneity.

In this work, we applied unsupervised and supervised methods to analyze changes in the co-expression patterns at an individual level for patients with LUAD. We hereby summarise the two principal findings of this study:

1. Using a metric of SSN similarity between patients, a graph-clustering strategy highlighted 6 different subtypes of LUAD. We found that clinical phenotypes were not the main driver of SSN similarity, however, two clusters (Cluster 1 and Cluster 5) displayed a significant enrichment of specific tumor classification and gender. On the other hand, cellular and molecular phenotypes were more clearly divisive between clusters, showing very distinct immune infiltration composition as well as enriched biological processes in their most conserved edges.
2. By calculating the weighted degree of each gene in an individual network, we observed that relevant biological signals consistent with known cancer biology can be retrieved. Furthermore, by using this node-centrality metric, we were able to capture complex traits related to each gene's structural role in a network and use them as features in a supervised learning scheme with the goal of finding novel network-based predictors of survival. Our approach underscored 12 genes (CHRDL2, SPP2, VAC14, IRF5, GUCY1B1, NCS1, RRM2B, EIF5A2, CCDC62, CTCFL, XG and TP53INP2) with known associations to the origin and progression of cancer, whose topological relevance in a network is predictive of patient survival in LUAD.

Regarding the limitations of our research, it should be stressed that all of these analyses were performed using the TCGA-LUAD cohort only. Given that SSN's place population heterogeneity as the main subject of study, new workflows integrating multiple sources of variation in every step of the procedure, including the individual network building stage, should be considered in order to validate that the results here obtained are not strictly dependent on the choice of data set. However, such a strategy involving multiple data sets and larger sample sizes are posed to be an expensive computational task when using mutual information as a network building method. Thus, careful experimental design should be taken into account in regards to the specific biological questions at hand.

The choice of an "ideal" set of background samples from which SSNs will be based on, is a major challenge of multiple SSN building methods that has been previously raised[60]. While the resampling methods employed in this work are effective means to attenuate data-dependent effects in predictive modelling workflows, an adequate independent validation analysis of the employed unsupervised and supervised methods needs to be pursued. Thus, we envision expanding our analysis pipeline to the entire NSCLC landscape as well as cancers with similar clinical outcomes as a first perspective avenue.

Finally, we contend that not only will the workflow implemented in this study significantly contribute to the growing methodological and analytical toolkit that patient-specific network understanding demands, but also that the biological insights uncovered will guide future endeavors in the fields of network-medicine and lung cancer research.

Methods

RNA-sequencing data acquisition and preprocessing

Following the methodology established by Andonegui et al.[10], clinical and RNA-sequencing (RNA-Seq) data corresponding to patients within the Lung Adenocarcinoma (LUAD) project in The Cancer Genome Atlas (TCGA) Research Network: https://www.cancer.gov/tcgawere extracted using the TCGAbiolinks R package[61]. Subsequently, all transcripts that were not protein-coding genes were filtered out, and genes with a zero-prevalence higher than 50% across all samples were removed. The resulting expression matrix was normalized to correct for transcript length effects, GC content, and library size. Furthermore, in order to focus on overall patient survival and explore its association with SSNs' topological features, samples belonging to patients whose survival data were not properly annotated were excluded. Thus, the analysis centered solely on patients who had been followed for more than a year or who had died before this period elapsed. Next, the scattering of expression profiles was visualized using principal component analysis (PCA) to verify the proper preprocessing of the samples. Finally, in order to focus on LUAD disease heterogeneity, only samples from primary tumor tissue were kept for SSN building and downstream analyses (Fig. 8).

Single sample network reconstruction and modeling

The LIONESS equation requires two objects as input: the first is a GCN built using all samples (network a) and the second is a GCN built without sample q (network $a - q$). To build the two input GCNs, we use ARACNE's implementation of Mutual Information (MI) as the measurement of the statistical dependence between the expression of all pair-wise combinations of genes[62]. In order to handle the inputs for both ARACNE and LIONESS in a computationally efficient manner, we integrated LIONESS into a multi-core backend for ARACNE[10] so that all MI calculations are done in parallel, while each SSN is built sequentially.

By using this workflow, we were able infer a total of 334 patient-specific LUAD GCNs. To allow the individual inspection of each network's node- and edge-level topological description, we created fully-connected SSNs as well as subsets of the top 10,000, 50,000 and 100,000 edges ranked by their weight (i.e., LIONESS score).

Clustering of patient network-similarity matrix

To assess whether the higher order structure of Individual GCNs is shared among phenotypically similar patients, we constructed a Patient network-similarity matrix A, where each entry a_{ij} contains the Jaccard-similarity index between the sets of edges of the filtered networks i and j. A is an $n \times n$ symmetrical matrix where n is the number of SSNs, making it suitable for representation as an adjacency matrix of an undirected weighted graph. In this graph, nodes represent patients and edges represent their SSN similarity score. This notation enables us to reframe the problem of patient clustering based on SSN similarity as one of detecting community structures in a graph. We performed community detection using the Louvain algorithm for weighted networks as implemented in the igraph R package[63,64]. Varying resolution values were allowed for the detection of smaller communities, as it is known that graph clustering methods based on modularity maximization can be biased towards detecting larger clusters in a network[65].

Finally, we compared the patient communities derived from SSN-similarity clustering with those obtained using gene expression data alone. To achieve this, we applied k-means clustering on the principal components that accounted for up to 90% of the total variance in the expression data. The optimal number of clusters was determined using the NbClust() function from the NbClust R package[66].

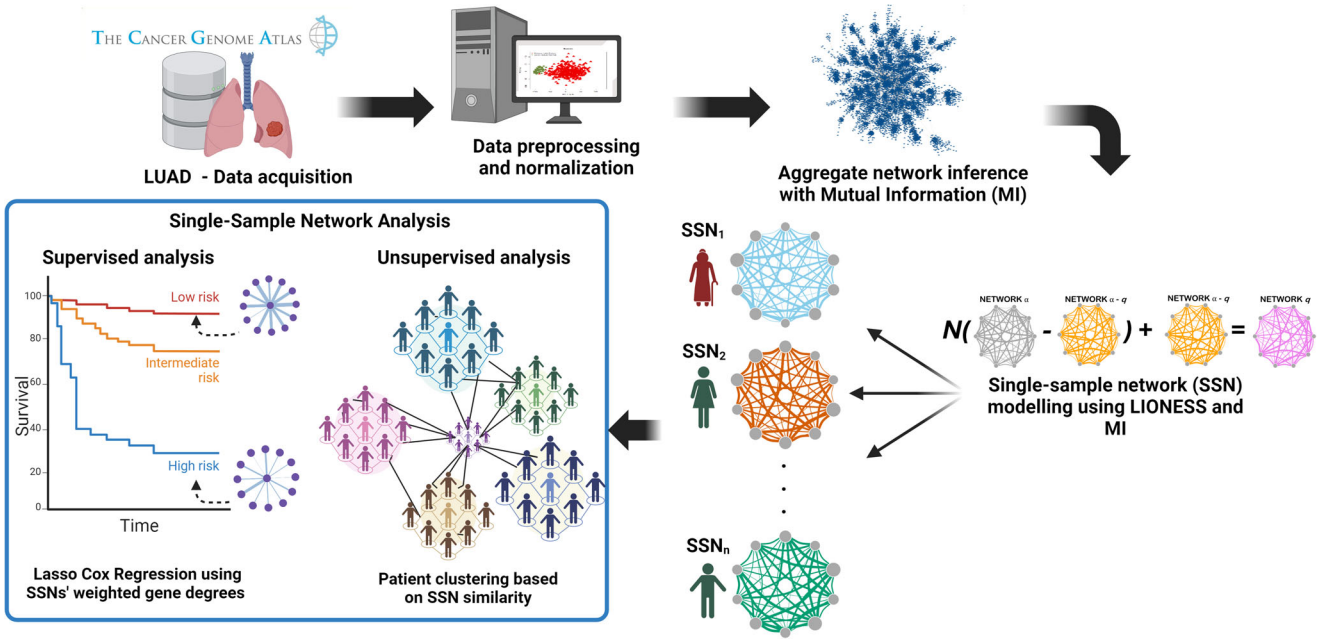

Fig. 8 | Schematic overview of the methods used in this study. Image created with BioRender.

Biological assessment of patient network-similarity clusters

Communities obtained from the Patient-similarity-graph were probed to shed light into the traits that drive clinical, cellular and molecular phenotypic heterogeneity. Thus, to biologically interpret each patient-cluster, we seek to answer three questions:

1. Are patients in a given cluster significantly enriched for particular clinical features?
2. Do patient-clusters have distinct cellular compositions?
3. What are the main biological processes that drive the formation of heterogenous phenotypes in the population?

To address the first question, we used hypergeometric testing followed by FDR correction to investigate statistically significant (adjusted p-value < 0.05) over-representation of gender, disease stage, tumor size and extent (T), lymph node involvement (N), and presence of distant metastasis (M). Moreover, we used a log-rank test to look for significant differences (p-value < 0.05) in the survival distributions of patients between clusters.

For the second question, we utilize the R implementation of Xcell[67] to calculate cell enrichment scores for each sample using normalized expression values to account for transcript length. We then conduct Kruskal-Wallis tests with post-hoc Dunn's test to identify any statistically significant associations between different cell types' enrichment score (Xcell value) and patient communities.

Lastly, to extract the most representative biological processes in each community, we repurposed the CoDiNA (Co-Expression Differential Network Analysis) framework to identify the most conserved edges within each cluster of patient-specific networks[68]. In essence, we obtain the entire edge-space of a given cluster of networks and examine the frequency with which an edge appears throughout all samples. We then consider an edge to be highly conserved in a set of networks if it is within the 99th percentile of most frequent edges. Thus, by aggregating the most conserved interactions, we are able to construct consensus networks that portray the most active set of sub-graphs in each cluster.

Subsequent biological annotation of each per-cluster consensus network is achieved by querying the presence of significantly enriched (adjusted p-value < 0.01) Biological Processes in its largest connected component. The detection of the largest connected components and functional enrichment analysis were carried out using the clusterMaker[69] and gProfiler[70] Cytoscape plugins[71], respectively.

Calculation of genes' weighted degrees in single-sample networks

The interaction between genes are severely altered during the transition from healthy to diseased states. These alterations can be better summarised by examining the centrality of individual nodes in each network. For example, several studies have previously used the weighted in-degrees of patient-specific regulatory networks as proxy of the amount of regulation that a given gene receives from all transcription factors[13,72,73]. Drawing inspiration from this approach, we built an $m \times n$ matrix of weighted degrees (WDs) for all m protein coding genes in n fully-connected SSNs. In the context of individual GCNs, the weighted degree of a gene serves as a summary statistic indicating the extent of potential active interactions within a SSN, encompassing both direct and indirect regulation, as well as patterns of coregulatory behaviour. Weighted degrees for all genes in each SSN were calculated using the igraph R package.

To validate that the resulting WD matrix reflects true biology of LUAD, we performed pre-ranked gene set enrichment analysis (GSEA[74,75]) on the standard deviation values of all WD scores. This allows us to investigate whether genes with the most variable patterns of connectivity are associated to specific Biological Processes[76] or Hallmark gene sets[77]. We use FDR < 0.01 as our significance cutoff in each enrichment test.

Predictive modelling of patient overall survival through penalized regression

To identify the sets of genes whose weighted degree in a network is associated to survival in patients with LUAD, we utilized the WD matrix as features for a penalized Cox proportional hazards regression model while correcting for gender and age at diagnosis[78]. Model construction was performed on a 70% training split, while validation was carried out on the remaining 30% hold-out test data set. In particular, given the high dimensionality of the WD matrix, we make use of the L1 penalty as it is known that this type of regularization performs both feature selection and model fitting in a manner that favours sparser models and counters overfitting[79,80].

We built two different regression models using the Least Absolute Shrinkage and Selection Operator (lasso) penalty. The first involves using the complete WD matrix when constructing the model, i.e. we let lasso have full control over the selection of genes that are relevant to the final fit. The second model involves a univariate Cox proportional hazards screening step for each gene prior to the shrinkage procedure. Thus, only genes with a

significant univariate association to survival (log-rank test, p-value < 0.05) were kept for subsequent lasso selection. The motivation behind this double filtering step is to first reduce the amount of competing noisy variables that could bias the shrinkage of coefficients, followed by a second reduction of the model's complexity by filtering out correlated variables through the L1 penalty.

Despite lasso's effectiveness in dealing with high dimensional data, it has been reported that lasso tends to over-shrink true large coefficients, thereby yielding biased estimates that are not directly suitable for generalization on test data[81]. To address this issue, the Relaxed lasso has been proposed as a de-biasing procedure in which lasso is only used as a variable screening step rather than a model selector. Consequently, the active set of variables is fitted through the Ordinary Least Squares solution without any penalization[81,82]. Owing to this, we inspect if relaxing the models after lasso variable selection produces a better fit. We utilized the relax.glmnet() function from the glmnet R package[83,84] with gamma = 0 for model predictions and evaluation, as this produces the most relaxed fit possible.

Fitting a regularized model requires the fine-tuning of the penalization parameter lambda. To achieve this, we find the optimal values of lambda for each model variant through 10-fold cross-validation. In particular, we choose the value of lambda that is one standard deviation away from the one that maximizes Harrel's Concordance Index (C-Index[85]) in the cross-validation procedure.

Model validation and feature stability analysis

To obtain a comprehensive evaluation of each model, we sought to quantify and compare all regressions' goodness-of-fit, predictive performance and overall uncertainty.

As a goodness-of-fit test, and as a first illustration of each model's prognostic power, we predicted risk scores for patients in the hold-out test data, and dichotomized them into High- and Low-risk groups using the median value as a threshold. Subsequently, we performed a log-rank test to check for significant differences between both risk groups' survival distributions.

Next, to evaluate a survival model's overall predictive performance, we focus on both global and time-discrete evaluation metrics. For a global assessment, we utilize the C-Index as well as the Integrated Time-dependent Brier's Score (IBS). To evaluate a model's performance at discrete time points, we use the area under the receiver operator characteristic curve (AUROC) for 1-, 2-, 3- and 5-year overall survival predictions as well as the Time-dependent Brier's Score (BS) for all possible time points.

Consequently, to test each model-building procedure and measure the final overall performance's uncertainty, we created 100 different training/validation data splits and repeated all filtering and hyperparameter-tuning steps, assessing the global discriminatory power in each run. Furthermore, we utilized this resampling-based method to identify the sets of genes most frequently selected by lasso and the univariate filter. The frequency with which a gene was selected was then used as a representative measurement of the robustness of its association to patient survival.

Finally, we performed the same analysis pipeline using the original gene expression data as input, to contrast the predictive performance of the resulting network-derived features.

Data availability

Clinical and RNA-sequencing (RNA-Seq) data corresponding to patients within the Lung Adenocarcinoma (LUAD) project in The Cancer Genome Atlas (TCGA) Research Network: https://www.cancer.gov/tcga.

Code availability

The pipeline for SSN reconstruction using ARACNE in parallel and the LIONESS method is available on GitHub and can be found at https://github.com/PatricioLOPSA/LIONESS-MI.

Received: 6 August 2024; Accepted: 10 April 2025;
Published online: 09 May 2025

References

1. Sung, H. et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: a cancer J. clinicians* **71**, 209–249 (2021).
2. Perez-Moreno, P., Brambilla, E., Thomas, R. & Soria, J.-C. Squamous Cell Carcinoma of the Lung: Molecular Subtypes and Therapeutic Opportunities. *Clin. Cancer Res.* **18**, 2443–2451 (2012).
3. Wang, B.-Y. et al. The comparison between adenocarcinoma and squamous cell carcinoma in lung cancer patients. *J. Cancer Res. Clin. Oncol.* **146**, 43–52 (2020).
4. Bade, B. C. & Dela Cruz, C. S. Lung Cancer 2020: Epidemiology, Etiology, and Prevention. *Clin. Chest Med.* **41**, 1–24 (2020).
5. Du, W. & Elemento, O. Cancer systems biology: Embracing complexity to develop better anticancer therapeutic strategies. *Oncogene* **34**, 3215–3225 (2015).
6. Serin, E. A. R., Nijveen, H., Hilhorst, H. W. M. & Ligterink, W. Learning from Co-expression Networks: Possibilities and Challenges. *Front. Plant Sci.* **7**, 444 (2016).
7. Espinal-Enríquez, J., Fresno, C., Anda-Jáuregui, G. & Hernández-Lemus, E. RNA-Seq based genome-wide analysis reveals loss of inter-chromosomal regulation in breast cancer. *Sci. Rep.* **7**, 1760 (2017).
8. García-Cortés, D., de Anda-Jáuregui, G., Fresno, C., Hernández-Lemus, E. & Espinal-Enríquez, J. Gene Co-expression Is Distance-Dependent in Breast Cancer. *Front. Oncol.* **10**, 1232 (2020).
9. Zamora-Fuentes, J. M., Hernández-Lemus, E. & Espinal-Enríquez, J. Gene Expression and Co-expression Networks Are Strongly Altered Through Stages in Clear Cell Renal Carcinoma. *Front. Genet.* **11**, 578679 (2020).
10. Andonegui-Elguera, S. D., Zamora-Fuentes, J. M., Espinal-Enríquez, J. & Hernández-Lemus, E. Loss of Long Distance Co-Expression in Lung Cancer. *Front. Genet.* **12**, 625741 (2021).
11. Nakamura-García, A. K. & Espinal-Enríquez, J. The network structure of hematopoietic cancers. *Sci. Rep.* **13**, 19837 (2023).
12. Kuijjer, M. L., Tung, M. G., Yuan, G., Quackenbush, J. & Glass, K. Estimating Sample-Specific Regulatory Networks. *iScience* **14**, 226–240 (2019).
13. Lopes-Ramos, C. M. et al. Gene Regulatory Network Analysis Identifies Sex-Linked Differences in Colon Cancer Drug Metabolism. *Cancer Res.* **78**, 5538–5547 (2018).
14. Freijeiro-González, L., Febrero-Bande, M. & González-Manteiga, W. A Critical Review of LASSO and Its Derivatives for Variable Selection Under Dependence Among Covariates. *Int. Stat. Rev.* **90**, 118–145 (2022).
15. Lababede, O. & Meziane, M. A. The Eighth Edition of TNM Staging of Lung Cancer: Reference Chart and Diagrams. *Oncologist* **23**, 844–848 (2018).
16. Sainz de Aja, J., Dost, A. F. M. & Kim, C. F. Alveolar progenitor cells and the origin of lung cancer. *J. Intern. Med.* **289**, 629–635 (2021).
17. Park, W. Y. et al. Ciliated adenocarcinomas of the lung: A tumor of non-terminal respiratory unit origin. *Mod. Pathol.* **25**, 1265–1274 (2012).
18. May, L., Shows, K., Nana-Sinkam, P., Li, H. & Landry, J. W. Sex Differences in Lung Cancer. *Cancers* **15**, 3111 (2023).
19. Pietras, R. J. et al. Estrogen and growth factor receptor interactions in human breast and non-small cell lung cancer cells. *Steroids* **70**, 372–381 (2005).
20. Ishibashi, H. et al. Progesterone Receptor in Non–Small Cell Lung Cancer—A Potent Prognostic Factor and Possible Target for Endocrine Therapy. *Cancer Res.* **65**, 6450–6458 (2005).
21. Florez, N. et al. Lung Cancer in Women: The Past, Present, and Future. *Clin. Lung Cancer* **25**, 1–8 (2024).
22. Chen, H.-H. et al. SWEET: A single-sample network inference method for deciphering individual features in disease. *Brief. Bioinforma.* **24**, bbad032 (2023).

23. Dai, H., Li, L., Zeng, T. & Chen, L. Cell-specific network constructed by single-cell RNA sequencing data. *Nucleic Acids Res.* **47**, e62 (2019).
24. Wang, L. et al. CHRDL2 promotes cell proliferation by activating the YAP/TAZ signaling pathway in gastric cancer. *Free Radic. Biol. Med.* **193**, 158–170 (2022).
25. Tu, Y., Chen, C. & Fan, G. Association between the expression of secreted phosphoprotein - related genes and prognosis of human cancer. *BMC Cancer* **19**, 1230 (2019).
26. Wang, C.-H. et al. TIG1 Inhibits the mTOR Signaling Pathway in Malignant Melanoma Through the VAC14 Protein. *Anticancer Res.* **43**, 2635–2643 (2023).
27. Roberts, B. K., Collado, G. & Barnes, B. J. Role of interferon regulatory factor 5 (IRF5) in tumor progression: Prognostic and therapeutic potential. *Biochimica et. Biophysica Acta (BBA) - Rev. Cancer* **1879**, 189061 (2024).
28. Sadeghi, M. et al. Network-Based and Machine-Learning Approaches Identify Diagnostic and Prognostic Models for EMT-Type Gastric Tumors. *Genes* **14**, 750 (2023).
29. Wang, G.-C. et al. The Role of NCS1 in Immunotherapy and Prognosis of Human Cancer. *Biomedicines* **11**, 2765 (2023).
30. Xue, L. et al. Ribonucleotide reductase subunit M2B deficiency leads to mitochondrial permeability transition pore opening and is associated with aggressive clinicopathologic manifestations of breast cancer. *Am. J. Transl. Res.* **10**, 3635–3649 (2018).
31. Zhong, X. et al. Targeting eIF5A2 inhibits prostate carcinogenesis, migration, invasion and metastasis in vitro and in vivo. *Bioengineered* **11**, 619–627 (2020).
32. Xu, C. et al. Clinical Eosinophil-Associated Genes can Serve as a Reliable Predictor of Bladder Urothelial Cancer. *Front. Mol. Biosci.* **9**, 963455 (2022).
33. Debaugny, R. E. & Skok, J. A. CTCF and CTCFL in cancer. *Curr. Opin. Genet. Dev.* **61**, 44–52 (2020).
34. Meynet, O. et al. Xg expression in Ewing's sarcoma is of prognostic value and contributes to tumor invasiveness. *Cancer Res.* **70**, 3730–3738 (2010).
35. Shi, K. et al. TP53INP2 modulates the malignant progression of colorectal cancer by reducing the inactive form of β-catenin. *Biochemical Biophysical Res. Commun.* **690**, 149275 (2024).
36. Wu, I. & Moses, M. A. BNF-1, a novel gene encoding a putative extracellular matrix protein, is overexpressed in tumor tissues. *Gene* **311**, 105–110 (2003).
37. Feng, D.-d et al. Transcription factor E2F1 positively regulates interferon regulatory factor 5 expression in non-small cell lung cancer. *Onco. Targets Ther.* **12**, 6907–6915 (2019).
38. Uchida, A. et al. Involvement of dual-strand of the miR-144 duplex and their targets in the pathogenesis of lung squamous cell carcinoma. *Cancer Sci.* **110**, 420–432 (2019).
39. Cho, E.-C. et al. Tumor suppressor FOXO3 regulates ribonucleotide reductase subunit RRM2B and impacts on survival of cancer patients. *Oncotarget* **5**, 4834–4844 (2014).
40. Chen, C., Zhang, B., Wu, S., Song, Y. & Li, J. Knockdown of EIF5A2 inhibits the malignant potential of non-small cell lung cancer cells. *Oncol. Lett.* **15**, 4541–4549 (2018).
41. Hong, J. A. et al. Reciprocal binding of CTCF and BORIS to the NY-ESO-1 promoter coincides with derepression of this cancer-testis gene in lung cancer cells. *Cancer Res.* **65**, 7763–7774 (2005).
42. Chen, H. et al. CHRDL2 promotes osteosarcoma cell proliferation and metastasis through the BMP-9/PI3K/AKT pathway. *Cell Biol. Int.* **45**, 623–632 (2021).
43. Sun, J. et al. Overexpression of colorectal cancer oncogene CHRDL2 predicts a poor prognosis. *Oncotarget* **8**, 11489–11506 (2016).
44. Lao, L. et al. Secreted phosphoprotein 24kD (Spp24) inhibits growth of hepatocellular carcinoma in vivo. *Environ. Toxicol. Pharmacol.* **51**, 51–55 (2017).
45. Chen, H. et al. Secreted phosphoprotein 24 kD (Spp24) inhibits the growth of human osteosarcoma through the BMP-2/Smad signaling pathway. *J. Orthop. Res.: Off. Publ. Orthop. Res. Soc.* **41**, 1803–1814 (2023).
46. Lao, L. et al. Secreted Phosphoprotein 24 kD Inhibits Growth of Human Prostate Cancer Cells Stimulated by BMP-2. *Anticancer Res.* **36**, 5773–5780 (2016).
47. Li, C.-S. et al. Secreted phosphoprotein 24 kD (Spp24) inhibits growth of human pancreatic cancer cells caused by BMP-2. *Biochemical Biophysical Res. Commun.* **466**, 167–172 (2015).
48. Lee, K.-B. et al. Effects of the bone morphogenetic protein binding protein spp24 (secreted phosphoprotein 24 kD) on the growth of human lung cancer cells. *J. Orthop. Res.: Off. Publ. Orthop. Res. Soc.* **29**, 1712–1718 (2011).
49. Langenfeld, E., Hong, C. C., Lanke, G. & Langenfeld, J. Bone Morphogenetic Protein Type I Receptor Antagonists Decrease Growth and Induce Cell Death of Lung Cancer Cell Lines. *PLOS ONE* **8**, e61256 (2013).
50. Wu, C.-K. et al. BMP2 promotes lung adenocarcinoma metastasis through BMP receptor 2-mediated SMAD1/5 activation. *Sci. Rep.* **12**, 16310 (2022).
51. Zhou, W., Yan, K. & Xi, Q. BMP signaling in cancer stemness and differentiation. *Cell Regeneration* **12**, 37 (2023).
52. Hao, F. et al. EIF5A2 Is Highly Expressed in Anaplastic Thyroid Carcinoma and Is Associated With Tumor Growth by Modulating TGF-β Signals. *Oncol. Res.* **28**, 345–355 (2020).
53. Wu, R., Zhong, Q., Liu, H. & Liu, S. MicroRNA-577/EIF5A2 axis suppressed the proliferation of DDP-resistant nasopharyngeal carcinoma cells by blocking TGF-β signaling pathway. *Chem. Biol. Drug Des.* **102**, 815–827 (2023).
54. Zhao, G. et al. EIF5A2 controls ovarian tumor growth and metastasis by promoting epithelial to mesenchymal transition via the TGFβ pathway. *Cell Biosci.* **11**, 70 (2021).
55. Zhang, Z. et al. HERC3 regulates epithelial-mesenchymal transition by directly ubiquitination degradation EIF5A2 and inhibits metastasis of colorectal cancer. *Cell Death Dis.* **13**, 1–12 (2022).
56. Zhou, Z. et al. TP53INP2 Modulates Epithelial-to-Mesenchymal Transition via the GSK-3β/β-Catenin/Snail1 Pathway in Bladder Cancer Cells. *OncoTargets Ther.* **13**, 9587–9597 (2020).
57. Soltanian, S. & Dehghani, H. BORIS: A key regulator of cancer stemness. *Cancer Cell Int.* **18**, 154 (2018).
58. Janssen, S. M. et al. BORIS/CTCFL promotes a switch from a proliferative towards an invasive phenotype in melanoma cells. *Cell Death Discov.* **6**, 1 (2020).
59. Makani, V. K. K., Mendonza, J. J., Edathara, P. M., Yerramsetty, S. & Pal Bhadra, M. BORIS/CTCFL expression activates the TGFβ signaling cascade and induces Drp1 mediated mitochondrial fission in neuroblastoma. *Free Radic. Biol. Med.* **176**, 62–72 (2021).
60. De Marzio, M., Glass, K. & Kuijjer, M. L. Single-sample network modeling on omics data. *BMC Biol.* **21**, 296 (2023).
61. Colaprico, A. et al. TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* **44**, e71 (2016).
62. Margolin, A. A. et al. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinforma.* **7**, S7 (2006).
63. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *Inter J. Complex Syst.* 1695 https://igraph.org (2006).
64. Csárdi, G. et al. *igraph: Network Analysis and Visualization in R* (2024). https://CRAN.R-project.org/package=igraph. R package version 2.0.3.
65. Lancichinetti, A. & Fortunato, S. Limits of modularity maximization in community detection. *Phys. Rev. E.* **84**, 066122 (2011).

66. Charrad, M., Ghazzali, N., Boiteau, V. & Niknafs, A. NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *J. Stat. Softw.* **61**, 1–36 (2014).
67. Aran, D. Cell-Type Enrichment Analysis of Bulk Transcriptomes Using xCell. *Methods Mol. Biol. (Clifton, N.J.)* **2120**, 263–276 (2020).
68. Morselli Gysi, D. et al. Whole transcriptomic network analysis using Co-expression Differential Network Analysis (CoDiNA). *PloS One* **15**, e0240523 (2020).
69. Morris, J. H. et al. clusterMaker: A multi-algorithm clustering plugin for Cytoscape. *BMC Bioinforma.* **12**, 436 (2011).
70. Kolberg, L. et al. G:Profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). *Nucleic Acids Res.* **51**, W207–W212 (2023).
71. Shannon, P. et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **13**, 2498–2504 (2003).
72. Lopes-Ramos, C. M. et al. Regulatory Network of PD1 Signaling Is Associated with Prognosis in Glioblastoma Multiforme. *Cancer Res.* **81**, 5401–5412 (2021).
73. Belova, T. et al. Heterogeneity in the gene regulatory landscape of leiomyosarcoma. *NAR Cancer* **5**, zcad037 (2023).
74. Subramanian, A. et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
75. Mootha, V. K. et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34**, 267–273 (2003).
76. Ashburner, M. et al. Gene Ontology: Tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
77. Liberzon, A. et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
78. Tibshirani, R. The lasso method for variable selection in the Cox model. *Stat. Med.* **16**, 385–395 (1997).
79. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B (Methodol.)* **58**, 267–288 (1996).
80. Ying, X. An Overview of Overfitting and its Solutions. *J. Phys.: Conf. Ser.* **1168**, 022022 (2019).
81. Su, W., Bogdan, M. & Candes, E. False Discoveries Occur Early on the Lasso Path (2016). 1511.01957.
82. Meinshausen, N. Relaxed Lasso. *Computational Stat. Data Anal.* **52**, 374–393 (2007).
83. Friedman, J. H., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **33**, 1–22 (2010).
84. Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *J. Stat. Softw.* **39**, 1–13 (2011).
85. Harrell, F. E. Evaluating the Yield of Medical Tests. *JAMA: J. Am. Med. Assoc.* **247**, 2543 (1982).

Acknowledgements

This work serves as fulfillment of Patricio López-Sánchez for obtaining a M.Sc. degree in the Posgrado en Ciencias Biológicas UNAM, within the field of knowledge of Biomedicine. We thank the Consejo Nacional de Ciencias, Humanidades y Tecnologías (CONAHCyT) for funding and support of this research (Scholarship No. 825718).

Author contributions

P.L.S. performed computational pipelines, drafted the figures, drafted the manuscript. F.A.M. discussed the manuscript and supervised. E.H.L. designed and supervised the statistical analyses. M.K. discussed the manuscript, supervised and provided computational facilities. J.E.E. conceived and directed the project, revised the manuscript, provided computational facilities and supervised.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41540-025-00522-0.

Correspondence and requests for materials should be addressed to Jesús Espinal-Enríquez.