Individualized pseudogenes networks for survival prognosis in B-cell acute lymphoblastic leukemia

Arturo Kenzuke Nakamura-García[1], Marieke L. Kuijjer[2,3,4], Jesús Espinal-Enríquez[1*]

1 Computational Genomics Division, National Institute of Genomic Medicine, 14610, Mexico City, Mexico.
2 iCAN Flagship in Digital Precision Cancer Medicine, University of Helsinki, Helsinki, 00290, Finland.
3 Department of Biochemistry and Developmental Biology, University of Helsinki, Helsinki, 00290, Finland.
4 Centre for Molecular Medicine Norway (NCMM), University of Oslo,Oslo, 0318, Norway.

* correspondence: jespinal@inmegen.gob.mx

Keywords:
B-cell acute leukemia/Pseudogenes/single-sample gene co-expression networks/survival risk stratification.

Individualized pseudogenes networks for survival prognosis in B-cell acute lymphoblastic leukemia

Arturo Kenzuke Nakamura-García[1], Marieke L. Kuijjer[2,3,4], Jesús Espinal-Enríquez[1*]

[1*]Computational Genomics Division, National Institute of Genomic Medicine, 14610, Mexico City, Mexico.
[2]iCAN Flagship in Digital Precision Cancer Medicine, University of Helsinki, Helsinki, 00290, Finland.
[3]Department of Biochemistry and Developmental Biology, University of Helsinki, Helsinki, 00290, Finland.
[4]Centre for Molecular Medicine Norway (NCMM), University of Oslo, Oslo, 0318, Norway.

*Corresponding author(s). E-mail(s): jespinal@inmegen.gob.mx;
Contributing authors: anakamura@inmegen.edu.mx; marieke.kuijjer@ncmm.uio.no;

Abstract

B-cell acute lymphoblastic leukemia (B-ALL) is the most common pediatric cancer, with significant advancements in risk stratification driven by next-generation sequencing (NGS). While genomic alterations have been extensively studied in B-ALL, the role of pseudogenes remains largely unexplored due to their historical classification as non-functional sequences. However, recent evidence suggests that pseudogenes may play regulatory roles in cancer. Our previous work identified changes in pseudogene connectivity in B-ALL bone marrow samples compared to normal samples, with specific pseudogene clusters being overexpressed in the malignant phenotype. These findings suggest that pseudogene co-expression patterns may contain biologically and clinically relevant information. To evaluate the role of the coexpression between pseudogenes in B-ALL and its impact on patient outcomes, we constructed single-sample co-expression networks (SSNs) using RNA-seq data from two independent B-ALL cohorts (n = 1,416). Unsupervised clustering of these networks revealed patient subgroups with distinct overall survival (OS) profiles. Differential co-expression and network topology analyses identified *EEF1A1P12* as a central hub, with its coordination state potentially

influencing OS. Specifically, distinct co-expression patterns between *EEF1A1P12* and *EEF1A1P4* were associated with significant survival differences. To translate these findings into a predictive framework, we developed a novel pipeline leveraging co-expression biomarkers for survival risk stratification. This approach identified the interaction between *RPL7P10* and *RPS3AP36* as a robust classifier for B-ALL patient survival. Our results establish pseudogene co-expression as a key molecular feature in B-ALL, with implications for patient stratification and prognostic modeling. This study underscores the importance of integrating pseudogene interactions into leukemia research and network medicine to improve precision oncology approaches.

Keywords: Pseudogenes, B-cell acute leukemia, survival risk stratification, single-sample gene co-expression networks

1 Introduction

B-cell acute lymphoblastic leukemia (B-ALL) is characterized by a clonal expansion of malignant B-cell progenitors and, although it has high cure rates, relapse is still very common [1–3]. Additionally, this type of leukemia is the most common among pediatric patients. Next-generation sequencing has helped to uncover new B-ALL subtypes, which are mostly defined by gene fusions and chromosomal rearrangements[4–9]. However, these works have mainly focused on deregulation among protein-coding genes, leaving the non-coding DNA sequences mostly unexplored. Recent studies have started to explore the potential regulatory roles of pseudogenes, suggesting that they may contribute to the complex gene expression landscapes observed in complex diseases such as B-ALL [10].

A pseudogene is a copy from a protein-coding gene that, due to detrimental alterations in its sequence, has lost its ability to code the original functional protein [11]. Historically, these type of DNA sequences have been considered "junk" DNA, and hence research in their possible roles in gene regulation has been largely hampered. However, it has been demonstrated that pseudogenes can be transcribed and even translated, although their functions diverge from their parental coding gene [12]. Multiple research works have demonstrated that pseudogenes are actively involved in gene regulatory circuits through diverse mechanisms. For example, they can alter the translation of their parental gene through endogenous competition for regulatory elements, or even alter DNA structure to promote or repress the transcription of a sequence (reviewed in [11]). Moreover, pseudogenes have been found to function as important modulators of gene expression in cancer [13]. Given the regulatory role of pseudogenes in gene expression, understanding their interactions within the co-expression landscape can provide valuable insights into the underlying biological processes in which they are involved. Furthermore, analyzing them in the context of complex diseases, such as B-ALL, can reveal novel information about the regulatory re-wiring that drives disease pathology.

Gene co-expression networks (GCNs) are a common tool in the field of systems biology to approach the inherent complexity of biological systems[14]. These networks are

theoretical graphs composed by genes connected between them when their expression profiles are correlated. A co-expression interaction between a pair of genes suggests a coordination which is possibly part of the same biological process [15]. Previous works have described GCNs from multiple cancer types[16–26]. Furthermore, we recently reported an increased co-expression between pseudogenes in hematological cancers compared to normal samples [27]. This suggests that a coordination between these sequences could be involved in the biology of these malignancies, which warrants further exploration. However, these recent results were obtained through the analysis of aggregate gene co-expression networks, constructed by integrating data from multiple samples. Consequently, these pseudogene co-expression networks provide only a summary of the biological data across populations. As a result, these aggregate networks fail to capture the intrinsic biological heterogeneity within the population.

Biological samples are highly heterogeneous, and therefore understanding how their inherent variability affects biological systems is crucial when designing treatments for complex diseases, such as cancer, or when evaluating clinical features of patients, such as survival rates. In general, a one-size-fits-all approach is inadequate for addressing complex diseases [28, 29], necessitating a more personalized approach.

Because of the latter, in the context of systems biology, new approaches have been developed to construct gene co-expression networks at the single-sample resolution. These approaches enable the creation of "single-sample networks" (SSNs), which describe the coordination of genes within individual samples from a population. Various methods have been proposed for this purpose (reviewed in [30]), each differing in how the individual sample networks are calculated. One such method is LIONESS [31], which employs a leave-one-out strategy to estimate each sample's contribution to the aggregated network. This method has been successfully applied to understand gene regulatory mechanisms in various cancers, leading to the identification of potential new regulatory subtypes in leiomyosarcoma [32], sex-biased regulatory patterns affecting prognosis in colon cancer [33], and regulatory signatures associated with poor prognosis in glioblastoma [34].

Due to our previous observation of pseudogene coexpression in leukemias [27], we hypothesized that modeling pseudogene interactions at single-sample resolution would show their contribution to cancer heterogeneity and that this heterogeneity could be associated with clinical characteristics. To answer this, we employed the LIONESS algorithm to construct gene co-expression single-sample networks from an RNA-seq dataset comprising 132 patients enrolled in the TARGET-ALL-P2 project, sourced from the Genomic Data Commons (GDC). Using the landscape of pseudogene co-expression, we grouped patients based on their edge weights in single-sample networks. Subsequently, we compared the overall survival (OS) among these patient groups. Our analysis successfully revealed distinct clusters of samples with significant differences in OS, solely through the examination of pseudogene-pseudogene interactions. Notably, the exclusion of these interactions yielded non-identifiable clusters with statistically significant differences in OS.

Additionally, after identifying the main differences between the clusters with high and low survival, we found *EEF1A1P12* to be a highly connected node that could potentially serve as a target to disrupt the coordination between pseudogenes.

In specific, we identified, we identified the interactions between *EEF1A1P12* and *EEF1A1P4*, as well as between *RPL7P10* and *RPS3AP36*, as co-expression biomarkers associated with survival. Our findings were corroborated using an independent dataset from the MP2PRT-ALL project. These results underscore the importance of including pseudogenes in the analysis of genomic data, as they may contain relevant information to better understand biological systems.

Results

Pseudogene co-expression landscape shows strong inter-chromosomal communication

To investigate the co-expression interactions between pseudogenes, we first calculated an aggregate co-expression network (as detailed in Methods) using data from 132 patients of the TARGET-ALL-P2 project, available in the GDC. Edges with a $p_{val} < 10^{-8}$ were considered significant for the study. This resulted in a network of 45,114 edges and 11,379 genes, from which, 6,032 edges occurred between 865 pseudogenes (Fig. 1). The edges of this aggregate pseudogene network were used for the subsequent downstream analysis. From now on, this pseudogene-pseudogene co-expression network will be referred as PG_{net}.

The PG_{net} of Fig. 1A shows the chromosome in which those molecules appear. As observed, the largest connected component of the network is formed by pseudogenes from all chromosomes, meanwhile the smallest islands are mostly composed of intra-chromosomal interactions. Additionally, as observed in Fig. 1B, the PG_{net} nodes are importantly grouped by the pseudogenes' families. This last result is particularly relevant in terms of the apparently loose chromosomal assortativity observed in the largest component of the PG_{net}, in which 94.59% (5,394) of edges are inter-chromosomal interactions. Certainly, pseudogenes in the largest connected component correlate with others located on different chromosomes; however, upon further investigation, we found that such correlations occur between pseudogenes that came from the same parental gene.

One possible explanation for the observed co-expression among pseudogenes within the same family is sequencing alignment error. Since a pseudogene is derived from the sequence of a coding gene, there is a certain similarity between both sequences. This feature may produce a challenge for aligners of RNA-seq data in distinguishing if a read belongs to a pseudogene or its parental gene [35]. To ensure this was not the case, we evaluated the correlation between the expression values of pseudogenes with their sequence similarities with respect to their parental gene. This analysis revealed moderate correlations for 14 pseudogene families (Fig. S1A-D and Supplementary Table 1). Additionally, we also found that in these families, members with high sequence similarity (¿50%) could also be lowly expressed (¡1 transcript per million (TPM), Fig. S1E). In fact, in most of the 14 families with significant correlation, the proportion of pseudogenes with high sequence similarity and mean expression below 1 TPM seems to be almost the same or even greater than the proportion of those with expression greater than 1 TPM (Fig. S1E). When these lowly expressed pseudogenes are excluded from the analysis, the moderate correlations found in Fig. S1A tend to drop and to

cut the families to less than 10 members (Fig. S1F). Lastly, when analyzing pairs of co-expressed pseudogenes, we found no strong correlation between their edge weights values and their sequence similarities (Fig. S2).

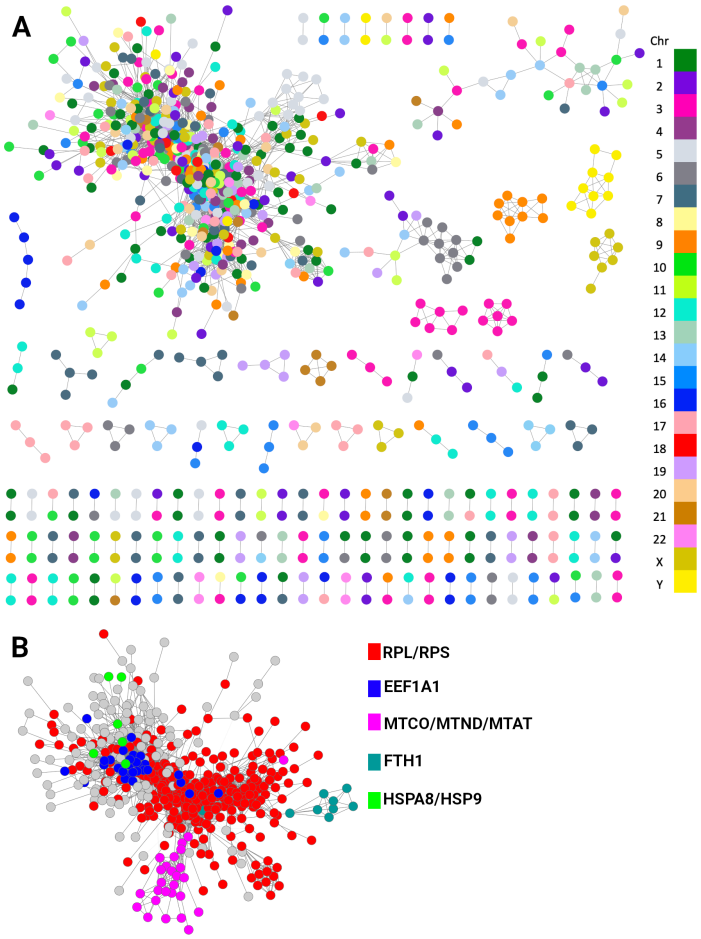

Fig. 1 A) Aggregated gene co-expression network between pseudogenes (PG_{net}) in B-ALL samples. The color of nodes corresponds to the chromosome in which those pseudogenes are located. To note, the largest component contains pseudogenes from all chromosomes, meanwhile the smaller ones are mostly formed of pseudogenes from the same chromosome. B) Largest component of A), with pseudogenes colored according to their families.

SSNs of PG_{net} enable stratification of B-ALL patients into clusters with distinct overall survival

We next implemented the LIONESS algorithm to the PG_{net}, to construct single-sample networks (SSNs) for each of the patients in our discovery cohort [31]. Briefly, LIONESS calculates the contribution of each sample to the aggregated network by

estimating the variation that occurs in the network when that sample is removed from the co-expression calculation, and combines this with a network modeled on all samples except the sample-of-interest to derive the final sample-specific network. This approach hence, allows for the comparison of edges that are similar, and edges that are distinctly relevant between sets of samples.

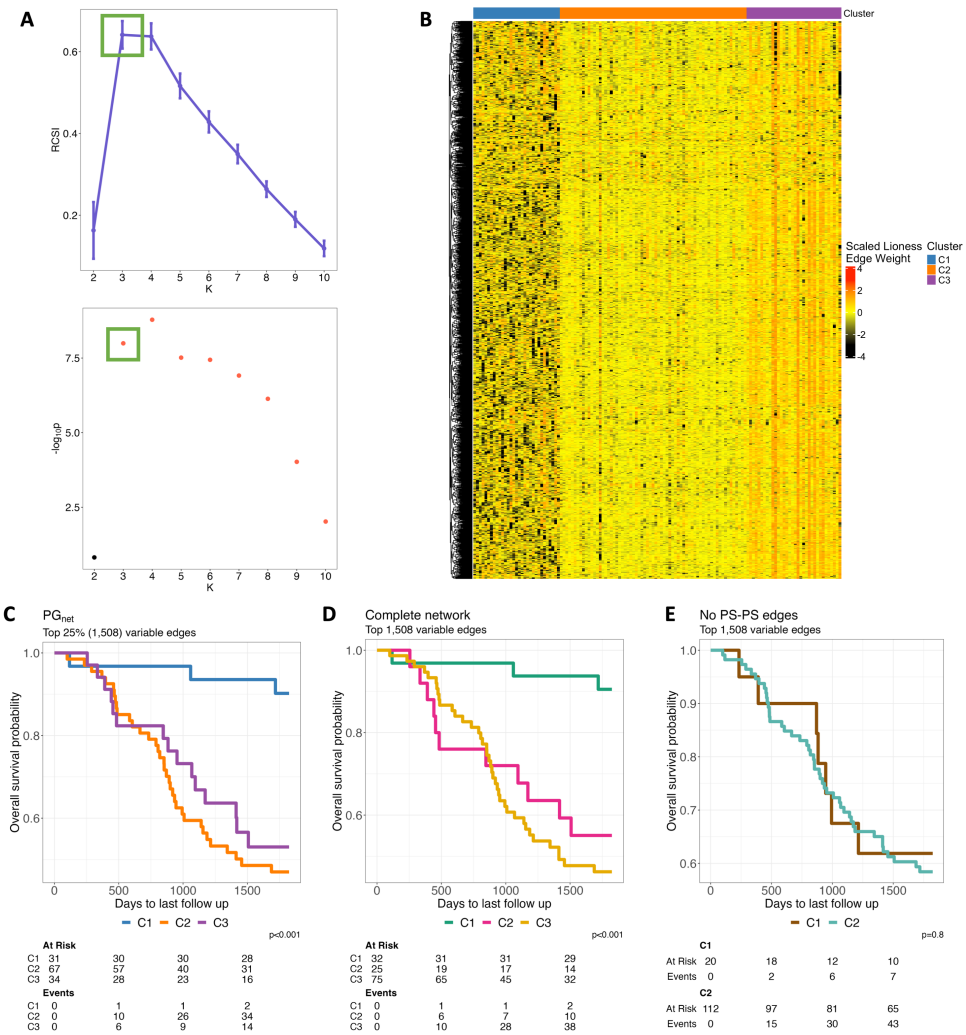

Fig. 2 Clustering analysis of single sample networks of B-ALL samples from TARGET-ALL-P2. A) The M3C algorithm shows K = 3 as the optimal partition for the data. The top figure shows the RCSI scores calculated for up to 10 partitions of the dataset. The bottom figure shows the p-values associated with each K of the top figure. B) Heatmap showing the SSNs' values scaled across samples clustered according to the M3C result. C) Kaplan-Meier analysis in the PG_{net}. D) KM plot in the network containing all types of edges. E) KM in the network without interactions between pseudogenes.

To group samples with similar features in their SSNs, we implemented the clustering algorithm M3C, which calculates a Relative Clustering Stability Index (RCSI) and evaluates whether different numbers of partitions in the data are more plausible compared to the null hypothesis that the data contain no clusters [36]. Since M3C recommends to use a set of variable features that are approximately normally distributed, we first filtered the top 25% most variable edges from the PG_{net} and then scaled these edges to Z-scores across samples (by subtracting the mean edge weight from the sample weight and dividing it by the edge's standard deviation).

Running the M3C algorithm with these data returned a value of 3 for the optimal number of clusters (K) (Fig. 2A) to partition the SSNs into one cluster with highly negative values (C1 in Fig. 2B), a second one with values very close to zero (C2), and a third cluster with highly positive values (C3). When performing a Kaplan-Meier (KM) analysis between these clusters, we found that C2 and C3 had worse OS compared to C1 (the cluster characterized by negative values on its SSNs) (Fig. 2C).

Validation of pseudogene-driven survival clusters

Despite this intriguing finding, we questioned whether the clustering of samples truly reflected biologically meaningful differences driven by the SSNs of pseudogenes, or if it was instead influenced by methodological factors. Specifically, co-expression network methods, such as LIONESS applied to Spearman correlation, inherently highlight sample-to-sample variations, which might lead to cluster formation even in the absence of strong biological signals. Additionally, selecting the most variable interactions for clustering could enhance the detection of differences, regardless of their biological relevance. To assess whether the observed clusters were a true biological phenomenon rather than an artifact of network construction, we repeated our analysis using a complete network (including all genes and their significant interactions) and a version that excluded pseudogene-pseudogene interactions.

We filtered the top 1,508 most variable edges in both cases, matching the number of edges previously analyzed in the PG_{nets}. Our analysis showed that the complete network (including co-expression between pseudogenes) resulted in similar clustering to the analysis of only PG_{nets} (Fig. 2D). In contrast, when we excluded co-expression between pseudogenes, the M3C algorithm identified 2 as the optimal number of clusters, with no difference in their overall survival (Fig. 2E). Performing PCA on these three sets of edges revealed that the network without pseudogene co-expression had less variability in its principal components (Fig. S3) compared to the other two networks.

To determine if the clustering of samples based on highly variable edges was just related to this variability, but not due to underlying biological relevance, we tested 1,000 sets of 1,508 randomly selected edges from the complete network using the same M3C algorithm. None of these null models produced clusters with differences in overall survival (Fig. S4), supporting the robustness of our findings.

Additionally, to validate our findings, we applied the same pipeline to a second external dataset containing 1,284 samples from the MP2PRT-ALL project [37]. Although this dataset has a large sample number, they are primarily from patients with favorable genetic features [37]. In addition, only 6% of the included patients died

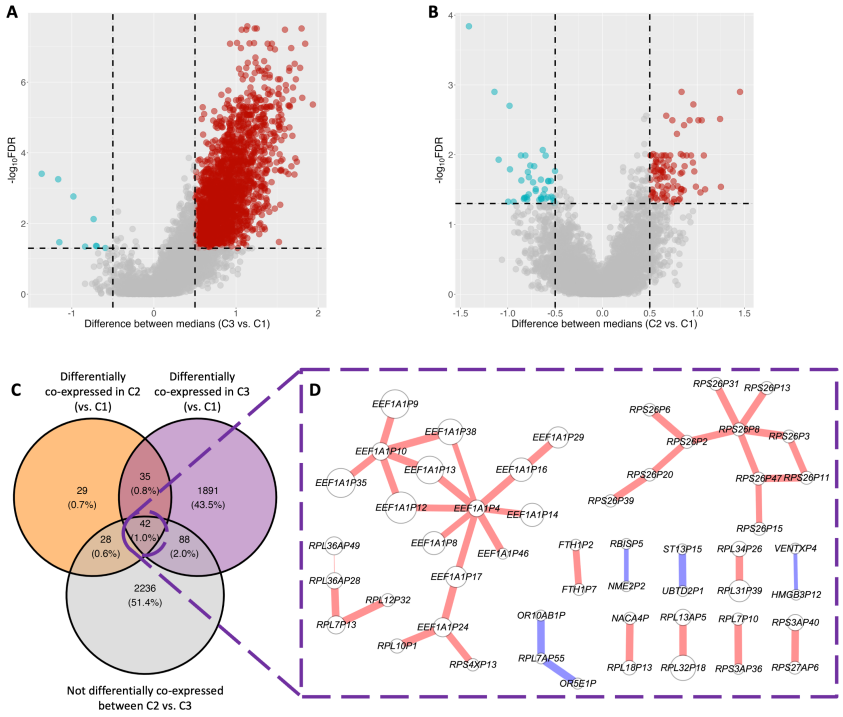

Fig. 3 Volcano plots showing the differential co-expression analysis between Clusters from the PG_{nets}, A) the comparison between C3 and C1. B) Comparison between C2 and C1. C) Venn diagram of differentially co-expressed edges between high (C1) and low survival (C2 and C3, orange and purple circles) clusters, versus not differentially co-expressed edges between C2 and C3. D) Resulting network from the intersection in C. Color of pseudogenes refers to their differential expression values between Clusters 3 V. 1: red represents overexpression, meanwhile blue pseudogenes are underexpressed. Analogously, edge color shows the differential co-expression between edges of Clusters 3 V. 1. Size of nodes is proportional to their degree in the PG_{net}.

within the first 5 years of follow up, which is a striking contrast with the TARGET-ALL-P2 cohort. Despite these differences, due to the lack of public data with complete annotation of follow-up times and with pseudogenes in expression matrix, this was the best dataset we could find and decided to use it as a validation cohort.

In the analysis of the MP2PRT dataset we found similar results. Here, the PG_{net} consisted of 19,244 edges between 1,763 pseudogenes. The clustering analysis of the top 25% variables edges in the PG_{net} identified two clusters, each characterized by distinct survival probabilities. However, here, the SSNs *without* pseudogene interactions also formed clusters with different survival outcomes. This indicates that co-expression between protein-coding genes could potentially also be associated with survival in some cohorts of B-ALL (Fig. S5). It is important to highlight that while survival in MP2PRT was not exclusively associated to pseudogene co-expression (as opposed to the analysis in TARGET), our independent analysis in this cohort does validate that

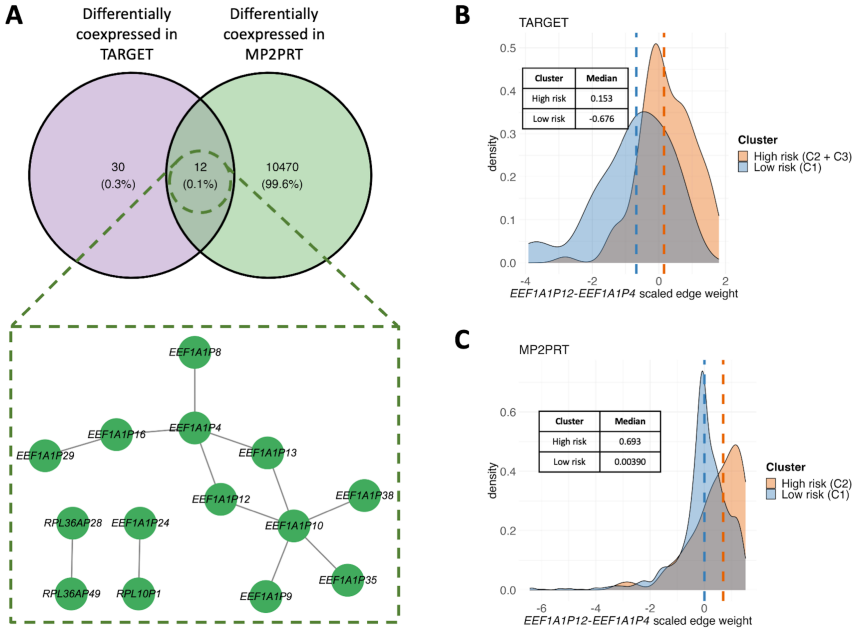

Fig. 4 Comparison of differentially co-expressed edges between cohorts. A) 12 edges are consistently present in differential sets between high and low survival groups on both datasets. B) and C) show weight distribution for *EEF1A1P12-EEF1A1P4* for TARGET and MP2PRT, respectively, by risk group.

the interactions between pseudogenes contribute to biological heterogeneity and to distinguish groups with different survival probabilities.

Differential co-expression analysis shows *EEF1A1P12* as a hub pseudogene and its interaction with *EEF1A1P4* as a biomarker of survival in B-ALL

To identify the specific pseudogene-pseudogene edges that differentiate clusters in the data, we performed a Wilcoxon test on the scaled values of the single-sample networks and calculated the median differences between groups with high and poor survival (e.g., the difference is calculated as the median of C2 minus the median of C1, $C2-C1$)

In both the TARGET and MP2PRT datasets, we found that clusters with poor survival had more positive values than the clusters with high survival (Fig. 3A,B and Fig. S6). Notably, the comparison between C1 and C2 in the TARGET network showed a mix of interactions that were both positively and negatively co-expressed (Fig. 3A,B, left).

Since our goal was to identify edges with clear differences across groups with varying survival outcomes, we filtered the network to retain edges that met the following criteria: a) significant difference in both comparisons (C3 vs. C1 and C2 vs. C1), and

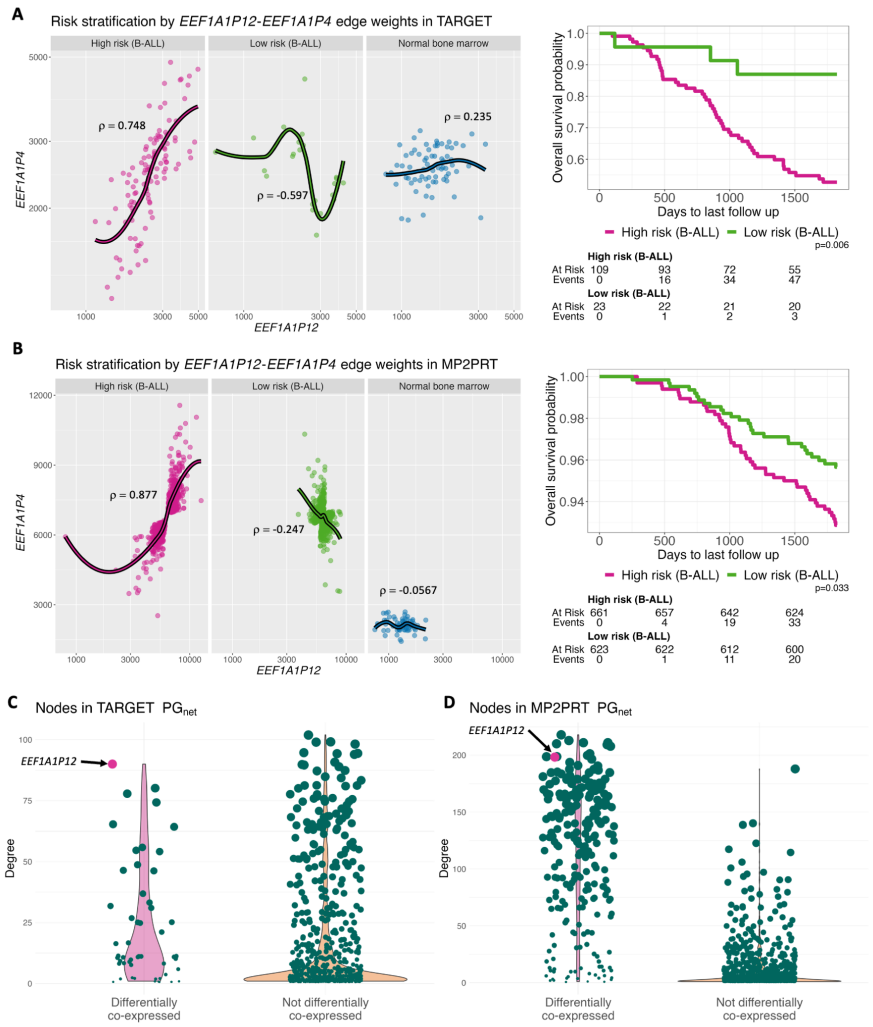

Fig. 5 EEF1A1P12 is a highly relevant node in PG_{net} associated with survival in B-ALL. A) Left: *EEF1A1P12* and *EEF1A1P4* expression in TARGET's high- and low-survival risk groups, and in normal bone marrow. Spearman's correletion coefficient is shown by sample group; right: survival analysis between risk groups in TARGET. B) The same analysis of *EEF1A1P12* and *EEF1A1P4* performed on MP2PRT data. C) and D) shows node degree of differentially and not differentially co-expressed pseudogenes from both TARGET PG_{net} and MP2PRT PG_{net}. Point size reflect degree values. *EEF1A1P12* is highlighted in pink.

b) no significant difference in the comparison between C3 and C2, as these two groups have similar OS and their differences should not be associated with this clinical variable (Fig. 3C). For the MP2PRT analysis we only compared C2 vs. C1, since here we only identified two clusters. The resulting network in the TARGET analysis is depicted in Fig. 3D.

To robustly identify consistent co-expressed edges associated with survival, we intersected the edges that were differentially co-expressed in both the TARGET and MP2PRT PG_{nets}. This resulted in a network comprising 12 edges and 14 pseudogenes, most of which were pseudogenes of the eukaryotic elongation factor 1 alpha 1 (*EEF1A1*) (Fig. 4A).

We decided to evaluate the potential of the edges in Fig. 4A to separate the cohorts into two different risk groups (since these are robustly found in TARGET and MP2PRT PG_{nets}). We used the edge weight medians of the good prognosis groups as a threshold to divide the samples into groups with high and low survival risks (Fig. 4B and C, and Supplementary Table S2). We hypothesized that stable predictors of survival could be identified by using the median weights of the good prognosis groups, as these values likely represent a threshold for distinguishing different coordination states that result in varying clinical outcomes. Using these thresholds, we aimed to identify edges whose coordination patterns are highly different between risk groups, serving as stable and biologically relevant predictors of survival.

This analysis, performed independently between cohorts, identified that the edge between *EEF1A1P12* and *EEF1A1P4* is consistently associated with different survival outcomes. To note, from the 12 edges analyzed, this was the only one that showed significant differences of survival probabilities between groups (Supplementary Table S2). In addition, it was one of the only two edges of *EEF1A1P12* in this network. Splitting the cohorts into high- and low-risk groups by the median weight of C1 (in both cohorts the good survival group is called C1), resulted in groups with different coordination states Fig. 5C and D). Moreover, the high-risk groups have a strong positive correlation between the pseudogenes. Meanwhile, the low risk groups shows moderate (in TARGET) and weak (MP2PRT) negative correlations. Lastly, in the normal bone marrow we found a weak positive correlation in TARGET and a null correlation in MP2PRT. This confirms that different coordination states are associated with different survival outcomes, and suggests that the interaction between *EEF1A1P12* and *EEF1A1P4* is associated with a mechanism not present in the normal bone marrow.

Lastly, to identify topologically relevant pseudogenes, we analyzed their degree in the PG_{nets}. Interestingly, this analysis revealed that *EEF1A1P12* is a highly connected node the networks from both datasets, ranking 10th in TARGET and 12th in MP2PRT. Among the top 20 nodes by degree, it was the only pseudogene in the differentially co-expressed edges consistently found in the analyses of both cohorts (5A and B).

Proof of concept for a co-expression risk classification model: *RPL7P10-RPS3AP36* is a stable biomarker for risk-stratification of B-ALL

So far, our analysis has shown that co-expression between pseudogenes is a widespread phenomenon in B-ALL. Furthermore, we have identified the pseudogene *EEF1A1P12* as a hub node in GCNs that also contributes to the risk stratification of B-ALL patients by its interaction with *EEF1A1P4*. However, our co-expression biomarkers, derived from complex operations on a population level, could be difficult to translate

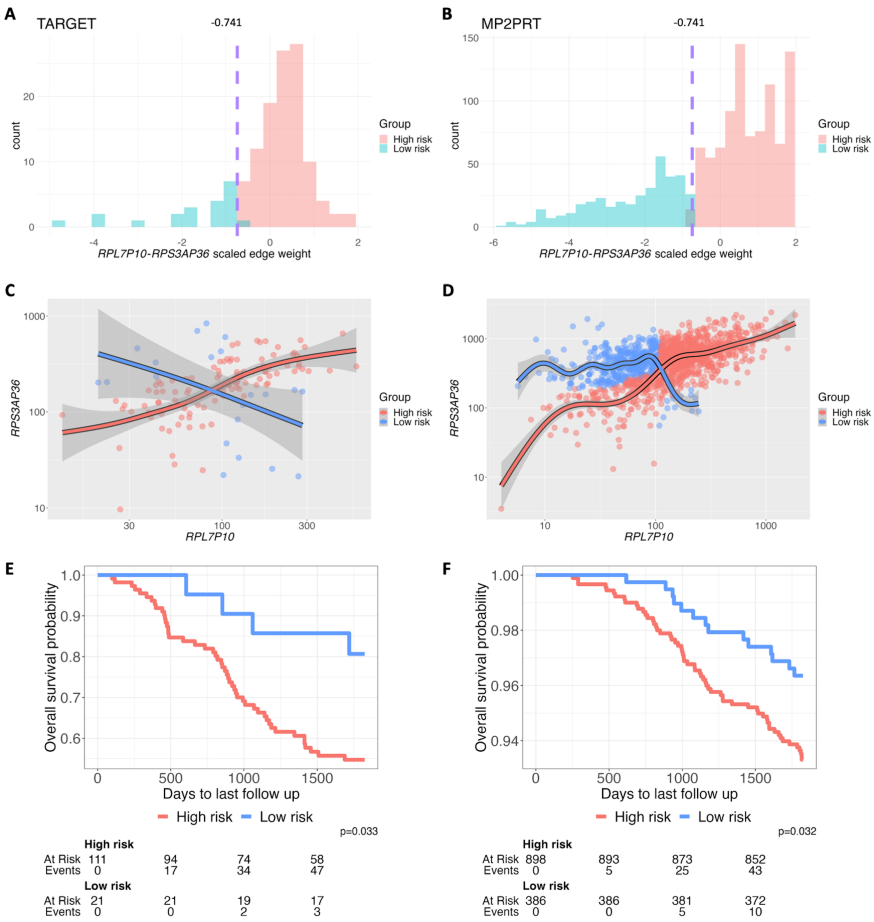

Fig. 6 Single-sample edge weights of *RPL7P10-RPS3AP36* are associated with different outcomes. Histograms of weight distribution by risk group are shown in A) (TARGET) and B) (MP2PRT). Scatter plots of *RPL7P10* and *RPS3AP36* show a divergent coordination state between risk groups in both TARGET (C) and MP2PRT (D). Overall survival comparison between risk groups confirm that differences in coordination of *RPL7P10* and *RPS3AP36* are associated with different survival probability in both TARGET (E) and MP2PRT (F)

to a clinical setting in which variables of different nature (often derived from clinical or cytogenetic analysis) are commonly used to risk-stratify patients.

Still, our results on pseudogenes interactions show promise to contribute to this end. Since our co-expression analysis is based on a model calculated at the population level, we decided to design a methodology that helps to classify patients based on their relative risk to a reference set of samples using co-expression biomarkers. This methodology is different from our previous independent analysis, and implies performing a single analysis using samples from both cohorts (integrating samples across cohorts). Here, we considered the TARGET cohort as our reference set, since it is

where the pseudogene co-expression is more relevant, and the MP2PRT samples as the new samples to classify.

Since this methodology is based on a relative risk assessment to the TARGET cohort, we decided to focus on the potential use of the 42 differentially co-expressed edges identified in Fig. 3D. As in our previous analysis of Fig. 5, we selected the median value in C1 as a threshold to split the data into two risk groups.

Since the LIONESS calculation is only comparable across samples included in the same model of the aggregate network, we need to include all the samples of interest in the calculation. However, given the intrinsic differences between patients of TARGET and MP2PRT, integrating both cohorts on the same model could introduce noise to the calculation, and would surely alter the interpretation of the LIONESS-Spearman edge weights (with respect to the interpretation we already have from TARGET).

To address these limitations, we performed an *in-silico* clinical setting. We treated each MP2PRT sample as a new sample to be compared with the reference dataset (TARGET). Since our analysis is done per sample in MP2PRT, and we are only interested in the relative risk to TARGET, we only have to include one new sample from MP2PRT in the new LIONESS model. Thus, for each new sample, we normalized its raw counts using the raw counts of the reference dataset and then calculated its LIONESS value for the 42 edges. By repeating this pipeline for each new sample independently, we ensure that the original model of TARGET is not significantly altered and hence, the interpretation of the edge weights of our prior analysis (Fig. 3) remains the same. After calculating the new LIONESS values for MP2PRT (that are relative to TARGET), we splitted the samples according to the median weights of the TARGET C1, per edge, to assess their predictive power across cohorts in this new context.

From the 42 edges analyzed, the only one that showed consistent results was *RPL7P10-RPS3AP36* (Supplementary table 3). The median value of this edge in cluster 1 was -0.741, and higher values were associated with a greater risk of survival (Fig. 6 A, 6 B and Supplementary Table 3). When analyzing the expression values of *RPL7P10* and *RPS3AP36*, we observed that the high-risk samples have a positive trend, meanwhile the low-risk groups have a negative trend in TARGET and a seemingly null trend in MP2PRT (Figs. 6C and 6D). This was also confirmed by the Pearson's correlation coefficients, which were strongly positive for the high-risk groups, but null for the MP2PRT low risk group and moderate negative in the low risk group from TARGET (Supplementary Fig. 7A). Interestingly, the Spearman correlation coefficients of the groups showed stronger coefficients values, suggesting that these trends are primarily monotonic rather than purely linear (Supplementary Fig. 7B). Finally, when performing a Kaplan-Meier analysis, we see that the theoretical high- and low-risk groups have statistically significant different survival probabilities (Figs. 6E and 6F), confirming that different coordination patterns of *RPL7P10-RPS3AP36* are associated with different survival probabilities.

Discussion

This study, for the first time as far as the authors are aware, classifies B-ALL patients exclusively based on co-expression interactions between pseudogenes.

Interactions between pseudogenes are not well understood, with most research focusing on their interactions with regulatory elements or coding genes. For instance, Carron, Coletta, and Lourenço analyzed co-expression modules to uncover potential functions for pseudogenes by examining their connectivity patterns with protein-coding genes[38]. Similarly, Chang et al.[39] reported ceRNA networks involving lncRNAs, pseudogenes, and miRNAs influencing lung cancer linked to carbon nanotube exposure. Therefore, more research is needed to understand the biological implications of pseudogene co-expression. Given the long-standing belief that pseudogenes are merely "junk" DNA, we aimed to investigate their potential utility at the co-expression level.

We constructed Gene Co-Expression Networks (GCNs) and Single-Sample Networks (SSNs) from 1,416 B-ALL samples from the GDC, spanning two independent projects (TARGET-ALL-P2 and MP2PRT-ALL). Clustering analysis of SSNs comprising only pseudogene-pseudogene (PS-PS) interactions demonstrated that these edges could produce clusters of samples with different overall survival in both datasets.

Interestingly, the pseudogene networks (PG_{nets}) exhibited greater variability than other edges, suggesting that including these interactions in co-expression analysis captures more heterogeneity. In the TARGET dataset, clustering without pseudogene interactions did not reveal differences in overall survival, whereas clustering with the whole network that incorporated all pseudogene interactions did. Although we did find significant associations when excluding pseudogene interactions from the MP2PRT dataset, indicating that there could be other mechanisms contributing to survival in this cohort, significant survival differences were still found in clusters formed with pseudogene edges, confirming our findings from the TARGET cohort.

Several factors could account for the differing results across datasets. The MP2PRT dataset has nearly ten times more samples, providing greater statistical power and yielding more significant edges, also between protein-coding genes, and thus, it is likely to find more relevant edges between coding genes there. Furthermore, the MP2PRT-ALL dataset included patients with standard or high risk for relapse but with genetic features associated with favorable outcome. This dataset had only 74 patients (6%) who died within the first five years and 115 patients (9%) who died in the study's 16-year span. This bias towards favorable outcomes may partly explain the discrepancies with the TARGET analysis, which includes data from 53 patients (out of 132) who died within the first five years of follow-up.

Notably, despite the clinical differences between TARGET and MP2PRT, we have very consistent results. For example, heterogeneity in both cohorts is greatly explained by the pseudogenes co-expression (Fig. S3), and, per our clustering analysis, this heterogeneity is linked with the most important clinical outcome, overall survival. Additionally, 28% of edges from the network in Fig 3D (TARGET analysis) are also differentially co-expressed between the differing survival groups in MP2PRT (Fig. 4A), supporting their importance to B-ALL biology. Furthermore, we identified *EEF1A1P12* as a hub node in both cohorts, which seems pivotal for the coexpression

network of pseudogenes, with a possible role in affecting survival in B-ALL (given its differential co-expression profile). Lastly, through different approaches, we identified two co-expression biomarkers of survival.

First, performing two independent analyses across cohorts, we demonstrated that a group of samples with a positive coordination between *EEF1A1P12* and *EEF1A1P4* is associated with worse survival compared to groups with a negative (or even null) coordination state (Fig. 5). The second biomarker was identified by simulating a clinical scenario to estimate the risk of a patient in MP2PRT relative to patients in TARGET. Through this analysis, the interaction between *RPL7P10* and *RPS3AP36* showed consistent results in both cohorts. As in the previous analysis, a positive coordination state was associated with worse survival compared to the opposite state. The consistency of these findings across cohorts is highly significant, as it underscores the potential utility of these interactions as prognostic biomarkers, even under distinct clinical contexts and in cohorts with a bias towards favorable outcomes (e.g. MP2PRT).

The fact that both approaches identified different biomarkers is a reflection of the intrinsic differences of the cohorts and methodologies used. Notably, the consistency of the interaction between *EEF1A1P12* and *EEF1A1P4* across cohorts in the first approach is due to the independent nature of this analysis, which evaluates the relative risk among samples within the same cohort. In contrast, the second approach assessed the risk of MP2PRT samples relative to TARGET samples. Thus, the interpretations of these approaches are inherently different and do not necessarily yield the same results. These methodological differences highlight the value of complementary approaches, as each reveals distinct but biologically relevant interactions associated with survival.

EEF1A1P12 has been associated with osteosarcoma patient survival[40] and treatment response in multiple myeloma[41]. It is also differentially expressed in 3D culture models of ovarian cancer cells [42]. Our results suggest it plays a pivotal part in the global communication between pseudogenes in B-ALL, with a possible implication in the patients' survival. However, understanding the biological meaning of its role as a hub node in PG_{nets} seems essential to understand the increased co-expression of pseudogenes in B-ALL. No information of *RPS3AP36*, *PL7P10* and *EEF1A1P4* was found on literature in the context of cancer.

Regardless of pseudogene expression levels, our findings suggest a mechanism coordinating pseudogenes, conferring aggressive features to cancer cells. In previous research on hematological cancers [27], we found that pseudogene interactions mostly occurred through inter-chromosomal interactions, unlike coding genes that mainly interact intra-chromosomally. Thus, the mechanism coordinating pseudogenes could be independent or even anti-correlated with the loss of inter-chromosomal communication in cancer.

A possible explanation for the co-expression (or even detected expression) of pseudogenes could be a bias in RNA-seq read mapping due to high sequence similarity between pseudogenes and their parental genes. We addressed this possible issue here (Fig. S1, Fig. S2 and Supplementary Table 1) and found no clear sign that our results were driven by this possibility.

Upon further testing, these co-expression edges could become part of novel prognostic models. Our approach to analyze the risk of new samples relative to a reference dataset provides an example of how co-expression signatures can be used to classify new patients. Future research could focus on developing co-expression biomarkers using more accessible technologies than RNA-Seq, such as PCR, to facilitate their adoption in clinical settings. In addition, these new classification models could be integrated with drug repositioning approaches to recommend personalized therapies for high-risk patients, thus indirectly finding drugs that possibly target the co-expression biomarkers.

Future research involves exploring targeted therapies to modulate the high-risk coordination state of gene pairs. Identifying drivers of a high-risk coordination state, such as transcription factors (TFs) or other regulatory elements, could provide a means to alter these interactions. For example, blocking a TF responsible for the high-risk coordination state may shift the network to a low-risk configuration. Alternatively, directly modifying the expression of one gene in the pair could disrupt or enhance the interaction, depending on the desired therapeutic outcome.

Finally, integrating these perspectives into the field of network medicine could lead to novel strategies for risk stratification and therapeutic interventions. For instance, expanding traditional drug recommendation tools to incorporate co-expression and regulatory network biomarkers, would enable the identification of drugs capable of re-wiring a patient's networks to improve treatment response and prognosis.

Concluding remarks

- This study represents the first attempt, to the best of our knowledge, to classify B-ALL patients based solely on co-expression interactions between pseudogenes.
- Single-Sample Co-expression Networks from 1,416 B-ALL samples revealed that pseudogene-pseudogene (PS-PS) interactions could produce clusters of samples with differing overall survival in two different datasets.
- Pseudogene networks (PG_{nets}) exhibited greater variability than other edges, suggesting that including these interactions in co-expression analysis captures more heterogeneity.
- Clustering without pseudogene interactions failed to reveal differences in overall survival in the TARGET dataset, while clustering with the whole network did. Similar results were observed in the MP2PRT dataset.
- In both datasets, PG_{nets} formed clusters with distinct survival outcomes, indicating the potential clinical relevance of these edges.
- Further analysis identified EEF1A1P12 as a hub node in both datasets, suggesting a pivotal role in coordinating multiple pseudogenes.
- Different coordination states between *EEF1A1P12* and *EEF1A1P4* are associated with different survival probabilities in B-ALL. Patients in which this pseudogene pair is anti-correlated have better survival outcomes.
- The interaction between *RL7P10* and *RPS3AP36* is a stable biomarker of survival in B-ALL with promising application in a classification model to estimate the relative survival risk of new samples. As in the previous case, a positive correlation is associated with worse survival probability.

- The mechanism coordinating pseudogenes in B-ALL remains unclear, but our findings suggest a potential biological mechanism driving pseudogene co-expression in cancer, independent of sequence similarity.

Our study sheds light on the importance of exploring single-sample co-expression networks in cancer research, particularly in the context of the role that pseudogenes may have in B-cell acute lymphoblastic leukemia. By leveraging these networks, we identified potential biomarkers for survival and uncovered novel insights into the complex interplay of pseudogene interactions in cancer. This underscores the significance of integrating single-sample co-expression analysis into cancer studies, offering valuable opportunities for identifying clinically relevant markers and understanding the underlying mechanisms driving cancer progression.

Methods

The code for reproducing the results of this work can be found at: https://github.com/AKNG97/PSs-SSNs-in-B-ALL.git. Every step of the pipeline was executed in R.

Data acquisition

The RNA-seq data used in this work were retrieved from Genomic Data Commons (GDC, https://portal.gdc.cancer.gov). Expression data from B-ALL samples were obtained from the MP2PRT-ALL and TARGET-ALL-P2 projects. Only cancer samples categorized as "Primary Blood-Derived Cancer - Bone Marrow" with complete survival follow-up information were included. Normal bone marrow samples were obtained from the TARGET-AML project. The types of counts used included TPM and unstranded counts.

A complementary gene annotation file was obtained from Biomart [43].

Data pre-processing

Unstranded counts from cancer and normal samples were processed together. Only genes present in both the expression matrix and the external gene annotation were considered. All genes with duplicated HGNC symbols were removed. When multiple replicates of a sample were found, we summed the expression of genes between replicates with the function "collapseReplicates" from the "DESeq2" [44] package. Next, we filtered out lowly expressed genes in the following order:

1. Genes on the first quartile of mean expression.
2. Genes with 0 counts in at least 50% + 1 samples.
3. Genes with average expression less than 10.

Network inference

For the co-expression analysis, raw counts were normalized with the following pipeline:

1. Within sample normalization by gene length and GC content with the EDASeq package [45].

2. Between sample normalization by trimmed mean of M values (TMM) with the NOISeq package [46].
3. Batch effect removal by ARSyN from the NOISeq package [46].

Gene lenght and GC content were retrieved from Biomart. Normalized counts from cancer samples were used to calculate the aggregate gene co-expression network using the Spearman's rank correlation coefficient. Correlations with a p-value ¡ 10^{-8} were considered as statistically significant. The LIONESS equation [31] was applied on the aggregate gene co-expression network to obtain the single-sample edge weights.

Sequence similarity analysis in pseudogene families

Sequence similarity analysis was performed only with the TARGET-ALL-P2 data.

The analysis between pseudogenes and parental genes used the TPMs counts for the B-ALL samples. Sequences were obtained from Biomart [43] using the function "getSequence" by Ensembl IDs. The type of sequence retrieved was "gene_exon_intron". Sequence similarity scores were calculated using the Smith-Waterman algorithm with the function "smith_waterman" from the "text.alignment" package [47].

For the analysis of sequence similarity between pairs of pseudogenes, we analyzed communities from the PG_{net} identified through the Louvain algorithm (from the "igraph" [48] package) to generate the communities.

Clustering analysis

Hierarchical clustering analysis was performed using the M3C package. The function "M3C" uses a Monte Carlo simulation to generate a null distribution of stability scores while preserving the structure of the input data. From this, a Relative Clustering Stability Index (RCSI) is calculated to evaluate the plausibility of different numbers of partitions (K) in the data compared to the null hypothesis that the data contains no clusters (K=1) [36]. The function returns an optimal value of K using the RCSI and its associated p-value.

M3C was run with a set of top variable features normalized to Z-scores across samples. For the analysis of the PG_{net} we used the top 25% most variable edges. Then, for the analyses of the complete network and the network without PS-PS edges, we used the same amount of top variable features as in the analysis of PG_{net}. PCA analysis of input data is computed automatically in each run of M3C.

Differential co-expression analysis

Differential co-expression analysis was performed by the median differences between groups. The significance of these differences was assessed using the Wilcoxon test. P-values were adjusted for multiple testing using the false discovery rate (FDR) method. Edges with an absolute median difference greater than 0.5 and an FDR < 0.05 were considered differentially co-expressed.

Co-expression biomarkers identification and validation

After evaluating the PG_{net} from TARGET and MP2PRT, we identified the shared differentially co-expressed edges across both analyses (Fig. 4A). These edges represent pairs of pseudogenes that highly contribute to B-ALL heterogeneity and whose weights are potentially associated with different survival probabilities. Given the significant differences in patient composition between the two cohorts, a first analysis was performed independently for each cohort. Here, we used the median edge weights on the good prognosis groups to split the datasets. This threshold, being representative of the good prognosis cluster, was expected to highlight biologically relevant differences tied to survival probabilities. The analysis was conducted on a per-edge, per-cohort basis.

A second analysis was conducted to evaluate the relative risk of new samples against a reference dataset. Due to the higher heterogeneity observed in the TARGET data, we selected it as the reference to test the relative risk of the MP2PRT samples. For this purpose, we repeated our data processing pipeline, from raw counts to single-sample networks inference, including the reference data alongside each individual MP2PRT sample. By excluding all MP2PRT samples except one at a time, we ensured that the single-sample networks were not influenced by the variability of other MP2PRT samples. This approach allowed us to model the variability of the new samples in contrast to the reference dataset. This analysis serves as a proof of concept for a risk classification model based on single-sample co-expression networks. We focused on the 42 edges shown in Fig. 3D, as these represent the set of edges most strongly associated with cluster separation in the TARGET cohort.

Supplementary information.

- **Supplementary Table S1.** Spearman correlation between expression values and sequence similarity (between pseudogenes and parental genes) by family.
- **Supplementary Table S2.** Survival analysis of the 12 commonly differentially co-expressed edges. Sample groups were separated by the C1 median weight, per cohort. Significant p-values on the log-rank test are highlighted in blue. Groups of high and low survival risk are highlighted in pink and green, respectively.
- **Supplementary Table S3.** Survival analysis of the 42 differentially co-expressed edges in TARGET, across cohorts. Sample groups were separated by the C1 median weight (Threshold) from TARGET, across cohorts. Significant p-values on the log-rank test are highlighted in green. Groups of high and low survival risk are highlighted in red and blue, respectively.
- **Supplementary Fig. S1.** Correlation analysis between sequence similarity and mean expression (TPMs) of pseudogenes with their parental genes in the TARGET dataset. A)Scatter plot of p-value and correlation value of analyzed families. Families of parental genes and pseudogenes in which a significant correlation (-log10(p-value) < 2 & absolute Rho > 0.5) was found are shown in blue. B), C) and D) show scatter plots of sequence similarity (between each PS and its parental gene) and mean expression, PSs above 1 TPM are shown in red and PSs below 1 in blue. E) Proportions of members with high sequence similarity (greater than 0.5) above and below detection level (1 TPM) from the 14 families with significant correlations (A). F) Spearman's correlation coefficient when all the members are considered (full

data) versus when only members above detection level are considered; here, only families from E) in which at least 10 members were above 1 TPM were considered.

- **Supplementary Fig. S2.** Correlation analysis between sequence similarity and edge weight in the aggregated network of the TARGET dataset. A to G show scatter plots by community. H) show the scatter plot of the complete network (6,032 edges).
- **Supplementary Fig. S3.** Principal components analysis performed on different network subsets. Top figures show the analysis on the TARGET network, bottom figures show the analysis using the MP2PRT data.
- **Supplementary Fig. S4.** Histogram of p-values from Kaplan-Meier analysis of clusters formed by randomly subsetting (1,000 times) 1,508 edges from the complete TARGET network.
- **Supplementary Fig. S5.** Kaplan-Meier analysis of clusters in MP2PRT data. A) Top 25% (4,811 edges) most variable PS-PS edges B) Top 4,811 most variable edges from the network without PS-PS edges. C) Top 4,811 most variable edges from the complete network containing all classes of edges.
- **Supplementary Fig. S6.** Volcano plot showing the differential co-expression analysis between clusters in the MP2PRT data formed by the analysis of the PGnets.
- **Supplementary Fig. S7.** Coefficients of Pearson (A) and Spearman (B) correlations between samples of high and low survival risk by stratification using the single sample edge weights of RPL7P10-RPS3AP36.

Acknowledgements. Arturo Kenzuke Nakamura-Garcia is a doctoral student from Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM). This work is part of his PhD thesis.

Declarations

- **Funding:** AKNG received a PhD scholarship from SECIHTI (number 806341).
- **Conflict of interest/Competing interests**. The authors declare no conflict of interest/competing interests.
- **Ethics approval and consent to participate:** Not applicable
- **Consent for publication:** Not applicable
- **Data availability:** Clinical and RNA-sequencing (RNA-Seq) data corresponding to patients within the TARGET-ALL-P2, MP2PRT-ALL and TARGET-AML projects in the Genomic Data Commons (GDC): https://portal.gdc.cancer.gov/
- **Materials availability:** Not applicable
- **Code availability:** The pipeline for reproducing the results from this manuscript is available on GitHub and can be found at https://github.com/AKNG97/PSs-SSNs-in-B-ALL.git
- **Author contribution:** A.K.N.G performed computational pipelines, drafted the figures, drafted the manuscript. 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.

References

[1] Ratti, S., Lonetti, A., Follo, M.Y., Paganelli, F., Martelli, A.M., Chiarini, F., Evangelisti, C.: B-all complexity: is targeted therapy still a valuable approach for pediatric patients? Cancers **12**(12), 3498 (2020)

[2] Roberts, K.G.: Genetics and prognosis of all in children vs adults. Hematology 2014, the American Society of Hematology Education Program Book **2018**(1), 137–145 (2018)

[3] Bhojwani, D., Yang, J.J., Pui, C.-H.: Biology of childhood acute lymphoblastic leukemia. Pediatric Clinics **62**(1), 47–60 (2015)

[4] Gu, Z., Churchman, M.L., Roberts, K.G., Moore, I., Zhou, X., Nakitandwe, J., Hagiwara, K., Pelletier, S., Gingras, S., Berns, H., *et al.*: Pax5-driven subtypes of b-progenitor acute lymphoblastic leukemia. Nature genetics **51**(2), 296–307 (2019)

[5] Li, J.-F., Dai, Y.-T., Lilljebjörn, H., Shen, S.-H., Cui, B.-W., Bai, L., Liu, Y.-F., Qian, M.-X., Kubota, Y., Kiyoi, H., *et al.*: Transcriptional landscape of b cell precursor acute lymphoblastic leukemia based on an international study of 1,223 cases. Proceedings of the National Academy of Sciences **115**(50), 11711–11720 (2018)

[6] Liu, Y.-F., Wang, B.-Y., Zhang, W.-N., Huang, J.-Y., Li, B.-S., Zhang, M., Jiang, L., Li, J.-F., Wang, M.-J., Dai, Y.-J., *et al.*: Genomic profiling of adult and pediatric b-cell acute lymphoblastic leukemia. EBioMedicine **8**, 173–183 (2016)

[7] Lilljebjörn, H., Henningsson, R., Hyrenius-Wittsten, A., Olsson, L., Orsmark-Pietras, C., Von Palffy, S., Askmyr, M., Rissler, M., Schrappe, M., Cario, G., *et al.*: Identification of etv6-runx1-like and dux4-rearranged subtypes in paediatric b-cell precursor acute lymphoblastic leukaemia. Nature communications **7**(1), 11790 (2016)

[8] Zhang, J., McCastlain, K., Yoshihara, H., Xu, B., Chang, Y., Churchman, M.L., Wu, G., Li, Y., Wei, L., Iacobucci, I., *et al.*: Deregulation of dux4 and erg in acute lymphoblastic leukemia. Nature genetics **48**(12), 1481–1489 (2016)

[9] Jeha, S., Choi, J., Roberts, K.G., Pei, D., Coustan-Smith, E., Inaba, H., Rubnitz, J.E., Ribeiro, R.C., Gruber, T.A., Raimondi, S.C., *et al.*: Clinical significance of novel subtypes of acute lymphoblastic leukemia in the context of minimal residual disease–directed therapy. Blood cancer discovery **2**(4), 326–337 (2021)

[10] Cheetham, S.W., Faulkner, G.J., Dinger, M.E.: Overcoming challenges and dogmas to understand the functions of pseudogenes. Nature Reviews Genetics **21**(3), 191–201 (2020)

[11] Nakamura-García, A.K., Espinal-Enríquez, J.: Pseudogenes in cancer: State of the art. Cancers **15**(16), 4024 (2023)

[12] Ji, Z., Song, R., Regev, A., Struhl, K.: Many lncrnas, 5'utrs, and pseudogenes are translated and some are likely to express functional proteins. elife **4**, 08890 (2015)

[13] Poliseno, L.: Pseudogenes: newly discovered players in human cancer. Science signaling **5**(242), 5–5 (2012)

[14] Hernández-Lemus, E., Reyes-Gopar, H., Espinal-Enríquez, J., Ochoa, S.: The many faces of gene regulation in cancer: a computational oncogenomics outlook. Genes **10**(11), 865 (2019)

[15] Van Dam, S., Vosa, U., Graaf, A., Franke, L., Magalhaes, J.P.: Gene co-expression analysis for functional classification and gene–disease predictions. Briefings in bioinformatics **19**(4), 575–592 (2018)

[16] Yang, Y., Han, L., Yuan, Y., Li, J., Hei, N., Liang, H.: Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. Nature communications **5**(1), 3231 (2014)

[17] Tovar, H., García-Herrera, R., Espinal-Enríquez, J., Hernández-Lemus, E.: Transcriptional master regulator analysis in breast cancer genetic networks. Computational biology and chemistry **59**, 67–77 (2015)

[18] Anda-Jáuregui, G., Velázquez-Caldelas, T.E., Espinal-Enríquez, J., Hernández-Lemus, E.: Transcriptional network architecture of breast cancer molecular subtypes. Frontiers in physiology **7**, 215436 (2016)

[19] 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. Scientific reports **7**(1), 1760 (2017)

[20] Alcalá-Corona, S.A., Espinal-Enríquez, J., Anda-Jáuregui, G., Hernández-Lemus, E.: The hierarchical modular structure of her2+ breast cancer network. Frontiers in physiology **9**, 377436 (2018)

[21] Velazquez-Caldelas, T.E., Alcalá-Corona, S.A., Espinal-Enríquez, J., Hernandez-Lemus, E.: Unveiling the link between inflammation and adaptive immunity in breast cancer. Frontiers in immunology **10**, 431428 (2019)

[22] Kalamohan, K., Gunasekaran, P., Ibrahim, S.: Gene coexpression network analysis of multiple cancers discovers the varying stem cell features between gastric and breast cancer. Meta Gene **21**, 100576 (2019)

[23] 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. Frontiers in genetics **11**, 578679 (2020)

[24] García-Cortés, D., Hernández-Lemus, E., Espinal-Enríquez, J.: Luminal a breast cancer co-expression network: Structural and functional alterations. Frontiers in genetics **12**, 629475 (2021)

[25] Li, Y.-K., Hsu, H.-M., Lin, M.-C., Chang, C.-W., Chu, C.-M., Chang, Y.-J., Yu, J.-C., Chen, C.-T., Jian, C.-E., Sun, C.-A., *et al.*: Genetic co-expression networks contribute to creating predictive model and exploring novel biomarkers for the prognosis of breast cancer. Scientific Reports **11**(1), 7268 (2021)

[26] Kasavi, C.: Gene co-expression network analysis revealed novel biomarkers for ovarian cancer. Frontiers in Genetics **13**, 971845 (2022)

[27] Nakamura-García, A.K., Espinal-Enríquez, J.: The network structure of hematopoietic cancers. Scientific Reports **13**(1), 19837 (2023)

[28] Gough, A., Stern, A.M., Maier, J., Lezon, T., Shun, T.-Y., Chennubhotla, C., Schurdak, M.E., Haney, S.A., Taylor, D.L.: Biologically relevant heterogeneity: metrics and practical insights. Slas Discovery: Advancing Life Sciences R&D **22**(3), 213–237 (2017)

[29] Dagogo-Jack, I., Shaw, A.T.: Tumour heterogeneity and resistance to cancer therapies. Nature reviews Clinical oncology **15**(2), 81–94 (2018)

[30] Deschildre, J., Vandemoortele, B., Loers, J.U., De Preter, K., Vermeirssen, V.: Evaluation of single-sample network inference methods for precision oncology. npj Systems Biology and Applications **10**(1), 18 (2024)

[31] Kuijjer, M.L., Tung, M.G., Yuan, G., Quackenbush, J., Glass, K.: Estimating sample-specific regulatory networks. Iscience **14**, 226–240 (2019)

[32] Belova, T., Biondi, N., Hsieh, P.-H., Lutsik, P., Chudasama, P., Kuijjer, M.L.: Heterogeneity in the gene regulatory landscape of leiomyosarcoma. NAR cancer **5**(3), 037 (2023)

[33] Lopes-Ramos, C.M., Kuijjer, M.L., Ogino, S., Fuchs, C.S., DeMeo, D.L., Glass, K., Quackenbush, J.: Gene regulatory network analysis identifies sex-linked differences in colon cancer drug metabolism. Cancer research **78**(19), 5538–5547 (2018)

[34] Lopes-Ramos, C.M., Belova, T., Brunner, T.H., Ben Guebila, M., Osorio, D., Quackenbush, J., Kuijjer, M.L.: Regulatory network of pd1 signaling is associated with prognosis in glioblastoma multiforme. Cancer research **81**(21), 5401–5412 (2021)

[35] Szabelska-Beresewicz, A., Zyprych-Walczak, J., Siatkowski, I., Okoniewski, M.:

Ambiguous genes due to aligners and their impact on rna-seq data analysis. Scientific Reports **13**(1), 21770 (2023)

[36] John, C.R., Watson, D., Russ, D., Goldmann, K., Ehrenstein, M., Pitzalis, C., Lewis, M., Barnes, M.: M3c: Monte carlo reference-based consensus clustering. Scientific reports **10**(1), 1816 (2020)

[37] Chang, T.-C., Chen, W., Elsayed, A., Pounds, S.B., Shago, M., Rabin, K.R., Raetz, E.A., Devidas, M., Cheng, C., Angiolillo, A.L., et al.: Genomic determinants of outcome in acute lymphoblastic leukemia: A Children's Oncology Group study. American Society of Clinical Oncology (2023)

[38] Carron, J., Della Coletta, R., Lourenço, G.J.: Pseudogene transcripts in head and neck cancer: Literature review and in silico analysis. Genes **12**(8), 1254 (2021)

[39] Chang, G., Xie, D., Hu, J., Wu, T., Cao, K., Luo, X.: Identification of candidate lncrna and pseudogene biomarkers associated with carbon-nanotube-induced malignant transformation of lung cells and prediction of potential preventive drugs. International Journal of Environmental Research and Public Health **19**(5), 2936 (2022)

[40] Liu, F., Xing, L., Zhang, X., Zhang, X.: A four-pseudogene classifier identified by machine learning serves as a novel prognostic marker for survival of osteosarcoma. Genes **10**(6), 414 (2019)

[41] Gruber, F., Keats, J.J., McBride, K., Runge, K., Wuest, D., Hadzi, T., Derome, M., Lonial, S., Khalil, I., Hayete, B., *et al.*: Bayesian network models of multiple myeloma: drivers of high risk and durable response. Blood **128**(22), 4406 (2016)

[42] Kerslake, R., Belay, B., Panfilov, S., Hall, M., Kyrou, I., Randeva, H.S., Hyttinen, J., Karteris, E., Sisu, C.: Transcriptional landscape of 3d vs. 2d ovarian cancer cell models. Cancers **15**(13), 3350 (2023)

[43] Durinck, S., Spellman, P.T., Birney, E., Huber, W.: Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. Nature protocols **4**(8), 1184–1191 (2009)

[44] Love, M.I., Huber, W., Anders, S.: Moderated estimation of fold change and dispersion for rna-seq data with deseq2. Genome biology **15**, 1–21 (2014)

[45] Risso, D., Schwartz, K., Sherlock, G., Dudoit, S.: Gc-content normalization for rna-seq data. BMC bioinformatics **12**, 1–17 (2011)

[46] Tarazona, S., Furió-Tarí, P., Turrà, D., Pietro, A.D., Nueda, M.J., Ferrer, A., Conesa, A.: Data quality aware analysis of differential expression in rna-seq with noiseq r/bioc package. Nucleic acids research **43**(21), 140–140 (2015)

[47] Wijffels, J.: Text.alignment: Text Alignment with Smith-Waterman. (2023). R package version 0.1.4. https://CRAN.R-project.org/package=text.alignment

[48] Csardi, G., Nepusz, T.: The igraph software package for complex network research. InterJournal **Complex Systems**, 1695 (2006)