picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Subject Section

spongEffects: ceRNA modules offer patient-specific insights into the miRNA regulatory landscape

Fabio Boniolo[1,2,†], Markus Hoffmann[1,3,†], Norman Roggendorf[1], Bahar Tercan[4], Jan Baumbach[5,6], Mauro A. A. Castro[7], A. Gordon Robertson[8], Dieter Saur[2], and Markus List[1,*]

[1]Chair of Experimental Bioinformatics, TUM School of Life Sciences, Technical University of Munich, Munich, Germany
[2]Chair of Translational Cancer Research and Institute of Experimental Cancer Therapy, Klinikum Rechts der Isar, School of Medicine, Technical University of Munich, Munich, Germany
[3] Institute for Advanced Study (Lichtenbergstrasse 2 a, D-85748 Garching, Germany), Technical University of Munich, Germany
[4] Institute for Systems Biology, Seattle, WA, USA
[5] Chair of Computational Systems Biology, University of Hamburg, Hamburg, Germany
[6] Computational BioMedicine Lab, University of Southern Denmark, Odense, Denmark
[7] Bioinformatics and Systems Biology Laboratory, Universidade Federal do Paraná, Curitiba, Brazil
[8] BC Cancer Genome Sciences Centre, Vancouver, Canada; Dxige Research Inc., Courtenay BC Canada

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
[*]To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Cancer is one of the leading causes of death worldwide. Despite significant improvements in prevention and treatment, mortality remains high for many cancer types. Hence, innovative methods that use molecular data to stratify patients and identify biomarkers are needed. Promising biomarkers can also be inferred from competing endogenous RNA (ceRNA) networks that capture the gene-miRNA-gene regulatory landscape. Thus far, the role of these biomarkers could only be studied globally but not in a sample-specific manner. To mitigate this, we introduce spongEffects, a novel method that infers subnetworks (or modules) from ceRNA networks and calculates patient- or sample-specific scores related to their regulatory activity.
Results: We show how spongEffects can be used for downstream interpretation and machine learning tasks such as tumor classification and for identifying subtype-specific regulatory interactions. In a concrete example of breast cancer subtype classification, we prioritize modules impacting the biology of the different subtypes. In summary, spongEffects prioritizes ceRNA modules as biomarkers and offers insights into the miRNA regulatory landscape. Notably, these module scores can be inferred from gene expression data alone and can thus be applied to cohorts where miRNA expression information is lacking.
Availability:https://bioconductor.org/packages/devel/bioc/html/SPONGE.html
spongEffects scores calculated for TCGA datasets: https://doi.org/10.6084/m9.figshare.19328885.v1
Contact: markus.list@tum.de
Supplementary information: Supplementary data are available at *Bioinformatics* online.

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture

1 Introduction

Despite recent advances in screening, diagnosis, and prognosis, and an increased understanding of the mechanisms driving tumorigenesis, progression, and maintenance, cancer deaths are estimated to be ~400,000/year by 2040 in the US alone (Rahib *et al.*, 2021), highlighting the need for innovative approaches for diagnosis, monitoring, and treatment. In addition to the traditional study of genome variation and gene expression, microRNAs (miRNAs) have been assessed as potential biomarkers, given their important roles in regulating gene expression both in physiological conditions and in cancer (Liu *et al.*, 2018). Mature miRNAs are short non-coding RNAs with lengths of 20-23 nucleotides that play an important role in regulating gene product abundance (Kartha and Subramanian, 2014). miRNAs are involved in regulating at least half the genes in the human genome (Friedman *et al.*, 2009) and are dysregulated in many diseases (Jiang *et al.*, 2009). They function via the targeting of RNAs that will be degraded or whose translation will be impeded (Bartel, 2009). miRNA-target recognition is facilitated via matching the seed region at the 5' end of the miRNA with the 3' binding site of the target (canonical targeting) or via the action of additional regions outside of the seed that can contribute to the recognition of the target (non-canonical targeting) (McGeary *et al.*, 2019). Salmena et al. proposed the competitive endogenous RNA (ceRNA) hypothesis (Salmena *et al.*, 2011), which suggests that RNAs with miRNA binding sites compete for a limited pool of miRNAs, giving rise to a complex gene regulatory network (Tsang *et al.*, 2010). Past studies have identified genes that act as ceRNAs (Salmena *et al.*, 2011; Tay *et al.*, 2014) and have associated them with different RNA classes, like messenger RNAs (mRNAs), circular RNAs, pseudogenes, transcripts of 3' untranslated regions (UTRs), and long non-coding RNAs (lncRNAs) (Dykes and Emanueli, 2017). In coding genes, effective miRNA targeting (i.e., reducing transcript levels) largely occurs in 3'UTRs (Agarwal *et al.*, 2015). lncRNAs have recently received particular attention in the framework of ceRNAs, suggesting the key role these molecules may play in binding miRNAs and indirectly regulating the expression of protein-coding genes sharing the same miRNA binding site region (Wang *et al.*, 2022). It is important to note that the key actors in these modules are not the miRNAs but rather the ceRNAs that exert influence over the expression of other ceRNAs using miRNAs as a limited resource.

A range of computational methods has been developed to identify potential miRNA-target interactions based on different inputs, such as gene expression data, protein-protein interaction networks, and sequence information (Muniategui *et al.*, 2013). Such methods often result in a large number of inferred interactions that, in return, suggest complex ceRNA networks that are typically difficult to explore. A key problem is to dissect these networks into functional units, or modules, that could play a role in specific tissues or diseases (Choobdar *et al.*, 2019). The identification of ceRNA modules could highlight the biological relevance of miRNA regulation in specific processes, while potentially identifying new prognostic biomarkers and therapeutic targets for clinical applications.

While tools for infering ceRNA networks exist (Zhang *et al.*, 2021), it is not straightforward to identify functional modules with biological or clinical relevance. Moreover, none of the ceRNA module identification methods proposed to date is able to summarize the information content of the modules in a patient- or sample-specific way. A sample-specific measure of ceRNA regulatory activity would be useful for the interpretation of its biological function and make it accessible for downstream analysis tasks such as clustering and classification. Methods that infer personalized ceRNA networks (Wang *et al.*, 2022) are suited to identify individual interactions (edges) that deviate from the norm but do not capture the overall activity of ceRNAs (nodes) in the network.

We introduce spongEffects, a new method to (1) extract ceRNA modules from previously inferred ceRNA networks (e.g. with SPONGE (List *et al.*, 2019)) and (2) to assesses their regulatory activity using enrichment scores as surrogates for module activity. These spongEffects scores are calculated at the single patient- or sample level, thus allowing the study of ceRNA effects across groups of samples or patients or even in personalized analyses.

Given gene expression data and a pre-computed ceRNA network, spongEffects performs several steps (Fig. 1): (1) it filters for significant interactions with meaningful effect size and identifies the most important ceRNAs via network centrality analysis; (2) for the subset of nodes with the highest centrality values, it constructs ceRNA modules by incorporating first-degree ceRNA network neighbors; (3) it performs single-sample gene set enrichment using the identified nodes, thus obtaining spongEffect scores; and (4) it uses the spongEffects scores to perform downstream machine learning tasks for classification and biomarker identification.

As one potential use case, we show that spongEffects scores can be used to classify breast cancer subtypes with high accuracy while random modules (obtained after network permutation) do not have a comparable predictive value. SpongEffects modules reveal ceRNAs and miRNAs that may play important roles in cancer biology. In contrast to the ceRNA networks that are typically inferred from matched gene and miRNA expression data, spongEffects scores are inferred exclusively from gene expression data, making spongEffects applicable to data sets for which miRNA expression profiles are missing. As we show using independent data from METABRIC (Curtis *et al.*, 2012), the enrichment methods underlying spongEffects are robust to missing values and across different technologies (e.g. microarray). This is a particular advantage for studying long non-coding RNAs which are often not included in gene expression data due to technical limitations.

2 Material and Methods

spongEffects was implemented in R (version 3.6.2) and is provided to the community as a new function in the SPONGE package in Bioconductor (see Availability). We used the pre-computed inferred ceRNA networks and the network centralities data from SPONGEdb (Hoffmann *et al.*, 2021) as input for spongEffects. We downloaded log2(TPM + 0.001) transformed TCGA expression data (RNAseq) from the XENA Browser (Goldman *et al.*, 2020). In addition, we downloaded log intensity levels (Illumina Human v3 microarray) gene expression data from the 1st and 2nd METABRIC (Molecular Taxonomy of Breast Cancer International Consortium) cohorts (Curtis *et al.*, 2012) from cBioPortal (Gao *et al.*, 2013) and used them as external datasets to validate our findings.

2.1 Defining ceRNA modules

Classical centrality measures, such as degree, closeness, and betweenness, can extract important information in biological networks by identifying hub and bottleneck nodes (**?**). The definitions of such centrality measures have been generalized for applications in weighted networks (Barrat *et al.*, 2004) to acknowledge that not all interactions are equally important. Here we use weighted degree centrality:

$$Degree_i = \sum_j^N x_{i,j} \quad (1)$$

$$Strength_i = \sum_j^N w_{i,j} \quad (2)$$

$$Centrality_i^\alpha = Degree_i \times (\frac{Strength_i}{Degree_i})^\alpha = (Degree_i)^{(1-\alpha)} \times Strength_i^\alpha \quad (3)$$

(1) describes the definition of the degree of centrality of node i in an undirected network of size N. x_{ij} is the binary adjacency matrix

The Author 2015. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

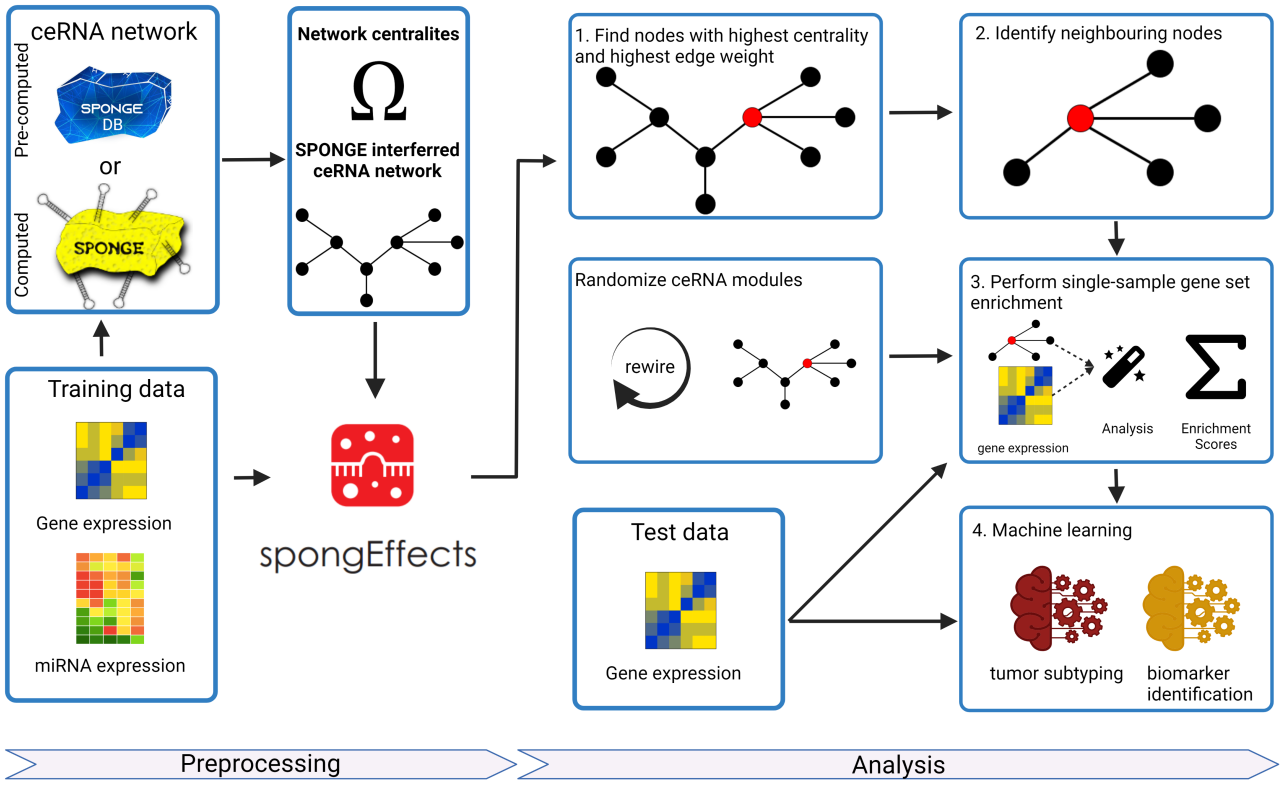

Fig. 1. The workflow of spongEffects. SpongEffects accepts a gene expression matrix and a ceRNA network as input. Subsequently, it 1) filters the network and calculates weighted centrality scores to identify important nodes, 2) identifies first neighbors, 3) runs single-sample gene set enrichment, and 4) prepares the output for further downstream tasks (e.g., machine learning-based classification and extraction of mechanistic biomarkers).

describing the connection of nodes i and j (i.e., $x_{ij} = 1$ if i and j are connected in the graph and 0 otherwise). (2) formalizes the weighted degree centrality measure, or node strength of node i in a network of size N. In particular, $w_{ij} > 0$ if nodes i and j are connected and it corresponds to the weight of the edge between i and j. Centrality is defined in (3). Opsahl et al., describes α as "a positive tuning parameter that can be set according to the research setting and data. If [α) is between 0 and 1, then having a high degree is taken as favorable, whereas if it is set above 1, a low degree is favorable" (Opsahl, 2009). SpongEffects uses weighted degree centrality as implemented in the R package tnet (version 3.0.16) (Opsahl, 2009) with the alpha parameter set to 1. In the context of a ceRNA network, we thus prioritize ceRNAs with large sum of weights (using multiple sensitivity correlation (mscor) as effect size). We define ceRNA modules as subnetworks consisting of ceRNA genes with highest weighted centrality scores (central node) and their first-degree neighbors.

2.2 spongEffects scores

SpongeEffects scores are calculated using one of three gene set enrichment approaches, namely single sample Gene Set Enrichment Analysis (ssGSEA) and Gene Set Variation analysis (GSVA) algorithms as implemented in the GSVA package (version 1.34.0) (Hänzelmann *et al.*, 2013), and Overall Expression (OE) (Jerby-Arnon *et al.*, 2018). These methods belong to the class of unsupervised and single-sample assessment tools (Hänzelmann *et al.*, 2013) and do not require initial knowledge of existing phenotype groups, which is often unavailable. Importantly, these approaches allow the calculation of the spongEffects even when some of the genes in the modules are not present in the input expression matrix, and result in module-by-sample spongEffects matrices that can be used for downstream analyses. When comparing results between GSVA, ssGSEA and OE (Suppl. Fig. 1) we did not observe any major differences in the results. Nevertheless, the original GSVA publication highlights that the choice of the optimal single-sample enrichment tool is highly dependent on the specific task and dataset and cannot be defined a priori.

2.3 Random Forest for subtype classification

We used Random Forest as implemented in the caret R package (version 6.0.90) (Kuhn, 2008) to classify tumor samples in the respective subtypes, using spongEffects scores as input. The optimal number of features (i.e., spongEffects modules) was randomly sampled as each split was defined via grid search in a repeated (3x) 5-fold cross-validation procedure. We calculated the Gini index to identify the modules that drove the predictions as implemented in the randomForest package (version 4.6.14) (Liaw *et al.*, 2002).

2.4 Randomization of the ceRNA networks

To validate our findings and evaluate whether the previously defined modules are robust, we sampled random modules. For each original module, we sample a random module of the same size such that the size distribution of the random modules matches the real modules exactly. Subsequently, we trained a classifier as previously described.

3 Results and Discussion

The spongEffects method offers sample-specific insights into the miRNA regulatory landscape and highlights ceRNAs heavily engaged in miRNA cross-talk and thus implicated in many biological processes as biomarkers. To illustrate the potential of spongEffect scores for functional

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

interpretation and machine learning, we exemplarily consider the task of breast cancer subtyping.

3.1 SPONGE modules predict breast cancer subtypes

Breast cancer, the leading cause of cancer-related deaths in women (Sung *et al.*, 2021), is a heterogeneous disease characterized by five subtypes (Luminal A, Luminal B, Basal, HER2-positive, and Normal) with different genetic prognostic profiles (Perou *et al.*, 2000). Mechanisms of miRNA dysregulation play a role in breast cancer (Mulrane *et al.*, 2013), offering the chance to validate miRNA mature strands as potential biomarkers (Hamam *et al.*, 2017). Here, we use spongEffects to study differences in miRNA regulation at the patient-specific level based on data from The Cancer Genome Consortium (Cancer Genome Atlas Network, 2012) (TCGA-BRCA, number of samples (n) = 1,063) and the Molecular Taxonomy of Breast Cancer International Consortium (Curtis *et al.*, 2012) (METABRIC cohorts 1 and 2, n = 1,905). In both cohorts, we selected all patients with tumor stages I, II, III, and IV, and removed all samples for which the subtype was missing or not identified as one of the five classes of interest (LumA, LumB, Her2, Basal, Normal-like). This step resulted in two final datasets: (1) TCGA-BRCA (n = 944), which we used for model training, and (2) METABRIC (n = 1,699), which we used as an external validation set.

SPONGE modules were calculated based on the TCGA-BRCA ceRNA network publicly available in SPONGEdb (Hoffmann *et al.*, 2021). The original network, containing ~3x107 edges, was filtered (mscor > 0.1, adjusted p-value < 0.01) down to 702,026 edges to preserve significant associations with non-negligible effect sizes (List *et al.*, 2019). Weighted centrality measures were calculated for each node in the network (see Material and Methods). For the purpose of this analysis, we chose to focus on lncRNAs as central nodes in the sponge modules, motivated by the importance these molecules have in ceRNA networks, their potential as biomarkers (Wang *et al.*, 2022), and their relevance in breast cancer subtypes (Su *et al.*, 2014). The 750 lncRNAs with the highest weighted centrality measures were selected and used as central nodes to define the sponge modules. We retain modules containing 10 to 200 genes after removing genes missing in the expression matrices, as previously suggested (Hänzelmann *et al.*, 2013).

SpongEffects scores were calculated for the two datasets separately (see Materials and Methods for details) using OE, GSVA, and ssGSEA which resulted in models with similar subtype classification performances (Suppl. Fig. 1), showing the robustness of the ceRNA modules. Here, we show the results for spongEffects scores calculated via OE, given its previous application in similar bulk transcriptional profiles (Jerby-Arnon *et al.*, 2018).

Fig. 2 shows that spongeEffects scores over all found modules differ markedly between subtypes. We note that the distribution of the basal subtypes diverges a normal-like distribution suggesting a hidden subgroup of patients. After disecting this distribution using the R package mixtools (version 1.2.0) (Benaglia *et al.*, 2010), we find that the smaller group is enriched in stroma-high tumors (Fig. 2 and Suppl. Fig. 2.a), as calculated via ESTIMATE (Yoshihara *et al.*, 2013), and in extracellular matrix (ECM)-related genes, which have been associated with Basal invasion programs under the regulation of KRT14 related genes (Hanley *et al.*, 2020) (Suppl. Fig. 2.b). Analogous studies have identified similar Basal subgroups from different data types and cohorts (Asleh *et al.*, 2022). This highlights the ability of spongEffects to identify a subset of samples with potential prognostic value.

We trained a random forest classifier on the TCGA-BRCA dataset and validated the resulting model on the METABRIC dataset. We repeated this analysis using different numbers of lncRNAs (200, 250, 500, 750, 1000, 1500, and 2000) to study the robustness of our approach and to highlight

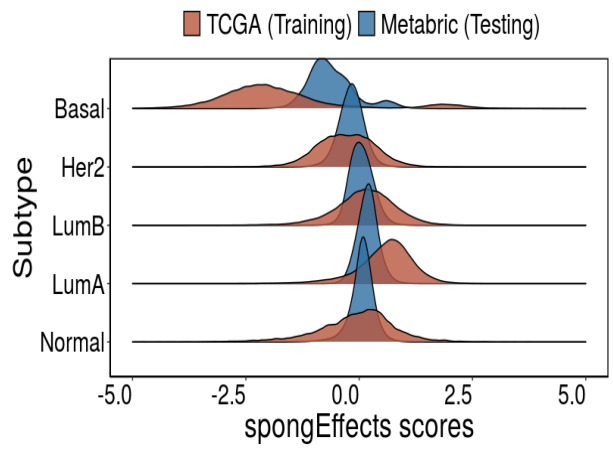

Fig. 2. Distribution of the spongEffects scores over all modules in the TCGA-BRCA (training) and METABRIC (testing) dataset. With the exception of Basal breast cancer, the scores for the various breast cancer subtypes follow a normal distribution. For the Basal subtype, the bimodality of the distribution hints at the existence of a subgroup of patients with characteristics different from the rest of the class (Suppl. Fig. 2).

the relatively small influence of the number of initial lncRNAs (and thus modules) considered on the final results (Suppl. Fig. 3). Moreoever, we compared our results to random modules and a baseline model trained on the expression values (not on spongEffects scores) of the 25 lncRNAs common to both datasets. The latter shows if the spongEffect scores offer additional value over the expression level of the ceRNAs alone. The subset accuracy metric (Ghamrawi and McCallum, 2005) was used to evaluate the classification performance along with sensitivity and specificity (confusion matrices in Suppl. Fig. 4). The spongEffects-based classification model outperformed the random module-based and the lncRNA-based models both in training and testing (Fig. 3.a) and preserved good specificity and sensitivity across all breast cancer subtypes (Suppl. Fig. 5.b).

3.2 ceRNA modules identify key biological mechanisms

Upon closer inspection of the top 25 ceRNA modules across breast cancer subtypes in both datasets (Fig. 3. b, Suppl. Fig. 6), we observe that Basal samples showed lower scores for modules centered around lncRNAs that were previously suggested to act as miRNA sponges. For example, CACNA1G-AS1 was shown to decrease proliferation, EMT transition, and migration in hepatocellular carcinoma (Yang *et al.*, 2019) and non-small cell lung cancer (Yu *et al.*, 2018), and to downregulate p53 levels in colorectal cancer (Wei *et al.*, 2020). On the contrary, DNM3OS has been shown to promote tumor progression and SNAIL-mediated EMT in gastric cancers (Wang *et al.*, 2019) and to confer resistance to radiotherapy by mediating DNA damage response mechanisms in esophageal squamous cell carcinoma. TPM1-AS has been shown to regulate alternative splicing of TPM1 and regulate TPM1-mediated migration of cancer cells in esophageal cancer (Huang *et al.*, 2017). Modules such as LINC00461 show strong enrichment in Basal samples. LINC00461 has been shown to promote migration and invasion in breast cancer (Zhang *et al.*, 2019; Dong *et al.*, 2019) and to play a role in progression in other cancer types such as hepatocellular carcinoma (Ji *et al.*, 2019) and renal cell carcinoma (Chen *et al.*, 2019), or in mediating radiosensitivity in lung adenocarcinoma (Hou *et al.*, 2020). In TCGA-BRCA, where patient-matched miRNA and gene expression data is available, we ranked miRNAs based on how many times they were predicted by SPONGE to target the genes in the modules and noted several miRNA families shared across modules (Suppl. Fig. 7.a). 13 of the 51 miRNAs that were frequently associated with modules were

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

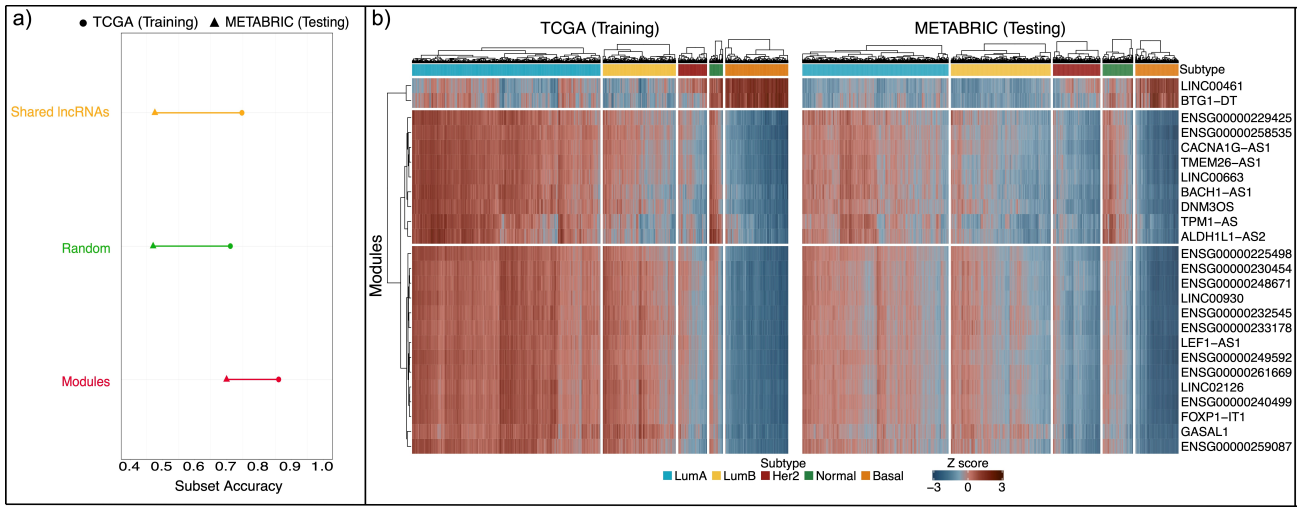

Fig. 3. Overview of Results. Overview of performance metrics for the random forest models calibrated on SPONGE modules (red), randomly defined modules (green), and lncRNAs common to both the TCGA (training) and METABRIC (testing) datasets (yellow). Subset accuracy (the proportion of samples that have all their labels classified correctly) values in training and testing for the three models. b) Heatmaps showing the standardized spongEffects scores for the top 25 most predictive modules calculated on the TCGA and METABRIC datasets. The heatmaps show a clear separation of the Basal subtype from the remaining ones, potentially pointing to the role of miRNA-regulated modules in the aggressive and proliferative phenotype observed in this subtype. .

also considered highly predictive in a subtype classification model based on miRNA expression data from the TCGA-BRCA dataset (Suppl. Fig. 8) and showed clear differences in expression between subtypes (Suppl. Fig. 7.b). This highlights how our approach offers insights into miRNA and ceRNA regulation even when miRNA expression data are unavailable. Focusing on the miRNA expression differences between subtypes offers a potential explanation for the spongEffects scores we observed.

3.3 Predictive SPONGE modules can identify potential biomarkers in breast cancer

Here, we exemplarily describe the CACNA1G-AS1 module and miRNAs that have been experimentally validated in more detail. Specifically, we investigated its role in the Basal subtype, one of the most aggressive breast cancer subtypes for which effective targeted therapies are still missing (Foulkes *et al.*, 2010; Lehmann *et al.*, 2011).

Genes in module CACNA1G-AS1 have lower expression levels in the Basal subtype when compared to the other subtypes in TCGA-BRCA and METABRIC (Suppl. Fig. 7.a and Suppl. Fig. 7.b) and contribute to the biology of this subtype. For example, expression of TBC1D9 has been shown to be inversely correlated to proliferation and grading in Basal breast cancer (Kothari *et al.*, 2021), while genes such as MYB and ZBTB16 are tumor suppressors whose promoter regions are often hypermethylated in Basal breast cancers (He *et al.*, 2020; Roll *et al.*, 2013).

In order to check whether we could exploit our modules to explain such changes, we further investigated the most frequent miRNAs in the CACNA1G-AS1 module (Suppl. Fig. 7) and checked their expression in the TCGA-BRCA dataset. Interestingly, miR-301b-3p and miR-130b-3p have higher expression in the Basal subtypes, giving a potential explanation for the spongEffects score of the genes they target (Suppl. Fig. 9.c). The same approach can be applied to module LINC00461, where genes belonging to these modules are known to be upregulated in Basal cancers and contribute to its highly proliferative and aggressive phenotype (Suppl. Fig. 9.d and Suppl. Fig. 9.e).

4 Conclusion and Outlook

SpongEffects is a new method for extracting ceRNA modules from pre-computed ceRNA networks. To the best of our knowledge, spongEffects

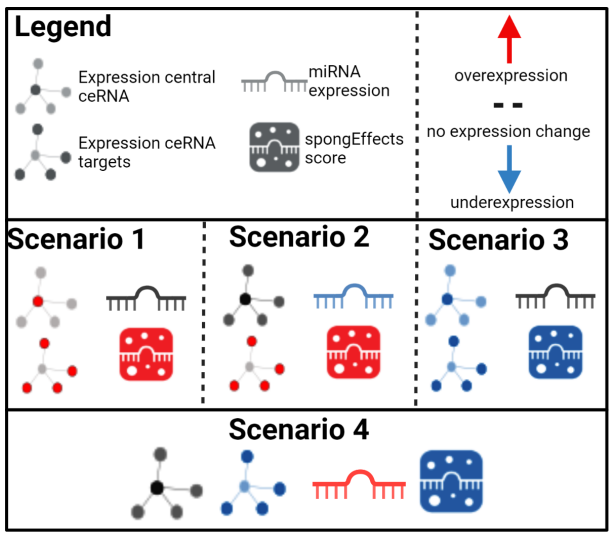

Fig. 4. Understanding increments and reductions in spongEffects scores. Increments in spongEffects scores can result from: Scenario 1) an increase in the expression of a module's central ceRNA while miRNA expression levels remain constant, leading the to upregulation of the module's target ceRNAs, and/or Scenario 2) a decrease in the expression of a module's miRNAs that results in higher expression of the the module ceRNAs. Negative spongEffects scores can result from: Scenario 3) a decrease of expression in the module's central ceRNAwhile while miRNA expression levels remain constant, leading to the downregulation of the module's target ceRNAs, and/or Scenario 4) an increase of expression in a module's miRNAs that results in lower expression of the the module's target ceRNAs.

is the first method that offers a sample- or patient-specific score that can quantify the regulatory activity of a ceRNA and its associated miRNAs.

We expect spongEffects scores to reflect two overlapping regulatory mechanisms. The first is miRNA-mediated ceRNA-target regulation, where an increase in the expression of a ceRNA leads to miRNA depletion, relieving other target genes from miRNA repression and, conversely, a decrease in ceRNA expression leaves more miRNA copies that can act as repressors of their target genes (Fig- 4 top). The second mechanism

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

is the classical miRNA regulation, where an increase in the expression of a miRNA leads to the repression of its target genes and vice versa (Fig. 4 bottom). While it is possible to verify the former in specific settings, e.g., in the presence of miRNA expression data, it is challenging to disentangle the two effects computationally. Single-sample enrichment tools estimate the enrichment or depletion of the sponge modules with respect to the expression of all the other genes in the expression matrix. In spite of the difficult interpretation, spongeEffects scores highlight important biomarkers. We have calculated spongEffects scores for samples from 22 different cancer types from TCGA and made the scores publicly available so that the community can use them to infer ceRNA and miRNA regulatory activities and gain insights into the gene regulatory landscape where the cohort size is too small for robustly inferring ceRNA networks or where supporting miRNA expression data is missing.

Using SPONGE-inferred ceRNA networks and RNA-seq data for TCGA breast cancer as inputs, we demonstrate how spongEffects scores can capture important aspects of cancer biology and be used in downstream machine learning tasks such as subtype prediction and biomarker identification. Using independent microarray-based gene expression data from the METABRIC consortium, we show that the inferred modules generalize to different cohorts. A particular strength of our approach is that spongEffects modules, once inferred from a ceRNA network, can also be computed for datasets or cohorts that lack miRNA data. For the modules that were predictive in subtype classification, spongEffects scores capture contributions of two different regulatory mechanisms: ceRNA regulation and miRNA regulation (see Interpretation of spongEffects scores). While it is challenging to separate the two, we observe that, in practice, our method returns a robust representation of microRNA-mediated regulatory effects.

Many gene regulatory mechanisms of lncRNAs are poorly understood and need to be experimentally discovered (Statello *et al.*, 2021). The spongEffects modules could help prioritize lncRNAs that have an expected impact on cancer biology and miRNA regulation, informing experimental validation and unraveling currently uncharacterized gene regulatory mechanisms of lncRNAs in combination with miRNAs. Of particular interest here is to address the question of lncRNAs' modes of action. For instance, it is currently debated if lncRNAs are exported outside of the nucleus, which would be a requirement for the established Argonaute-dependent mechanism of miRNA regulation (McGeary *et al.*, 2019). lncRNAs with an experimentally verified role as ceRNAs could then serve as important biomarkers and potential therapeutic drug targets. We note that while we have applied spongEffects here to ceRNA networks inferred with SPONGE, our method is generally applicable to arbitrary ceRNA networks and potentially also to gene-regulatory networks, where similar concepts have already been explored (Castro *et al.*, 2016; Chagas *et al.*, 2019). We foresee two interesting directions for further research. First, recent studies on the inference of regulons (Castro *et al.*, 2016; Fletcher *et al.*, 2013) offer the chance to integrate methods for the combined analysis of transcription factors and miRNA regulation. Second, the surge of single-cell sequencing technologies allows the study of regulatory mechanisms at a higher resolution than what was possible before (Van de Sande *et al.*, 2020; Ma *et al.*, 2020; Gibbs *et al.*, 2022), paving the way to the design of methods with the potential of uncovering new insights into the complexity of tumor biology.

5 Conflict of Interest Disclosure

The authors have no conflicts of interest to declare.

Acknowledgements

We want to thank Christina Trummer for designing the spongEffects logo. We further thank Marcel H. Schulz for constructive feedback.
The results published and shown here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga Images were created with https://www.biorender.com. Parts of images were designed using resources from Flaticon.com.

Funding

This project is funded by the European Union's Horizon 2020 research and innovation program (No. 777111). This publication reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information it contains. This work was further supported by the German Federal Ministry of Education and Research (BMBF) within the framework of the e:Med research and funding concept (No. 01ZX1908A). With the support of the Technical University Munich – Institute for Advanced Study, funded by the German Excellence Initiative.

References

Agarwal, V., Bell, G. W., Nam, J.-W., and Bartel, D. P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. *Elife*, **4**.

Asleh, K., Negri, G. L., Spencer Miko, S. E., Colborne, S., Hughes, C. S., Wang, X. Q., Gao, D., Gilks, C. B., Chia, S. K. L., Nielsen, T. O., and Morin, G. B. (2022). Proteomic analysis of archival breast cancer clinical specimens identifies biological subtypes with distinct survival outcomes. *Nat. Commun.*, **13**(1), 896.

Barrat, A., Barthélemy, M., Pastor-Satorras, R., and Vespignani, A. (2004). The architecture of complex weighted networks. *Proc. Natl. Acad. Sci. U. S. A.*, **101**(11), 3747–3752.

Bartel, D. P. (2009). MicroRNAs: target recognition and regulatory functions. *Cell*.

Benaglia, T., Chauveau, D., Hunter, D. R., and Young, D. S. (2010). mixtools: An R package for analyzing mixture models. *J. Stat. Softw.*, **32**, 1–29.

Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, **490**(7418), 61–70.

Castro, M. A. A., de Santiago, I., Campbell, T. M., Vaughn, C., Hickey, T. E., Ross, E., Tilley, W. D., Markowetz, F., Ponder, B. A. J., and Meyer, K. B. (2016). Regulators of genetic risk of breast cancer identified by integrative network analysis. *Nat. Genet.*, **48**(1), 12–21.

Chagas, V. S., Groeneveld, C. S., Oliveira, K. G., Trefflich, S., de Almeida, R. C., Ponder, B. A. J., Meyer, K. B., Jones, S. J. M., Robertson, A. G., and Castro, M. A. A. (2019). RTNduals: an R/Bioconductor package for analysis of co-regulation and inference of dual regulons. *Bioinformatics*, **35**(24), 5357–5358.

Chen, L., Zhang, T., Zhang, S., Huang, J., Zhang, G., Xie, K., Wang, J., Wu, H., and Dai, G. (2019). Identification of long Non-Coding RNA-Associated competing endogenous RNA network in the differentiation of chicken preadipocytes. *Genes*, **10**(10).

Choobdar, S., Ahsen, M. E., Crawford, J., Tomasoni, M., Fang, T., Lamparter, D., Lin, J., Hescott, B., Hu, X., Mercer, J., Natoli, T., Narayan, R., DREAM Module Identification Challenge Consortium, Subramanian, A., Zhang, J. D., Stolovitzky, G., Kutalik, Z., Lage, K., Slonim, D. K., Saez-Rodriguez, J., Cowen, L. J., Bergmann, S., and Marbach, D. (2019). Assessment of network module identification across complex diseases. *Nat. Methods*, **16**(9), 843–852.

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., Gräf, S., Ha, G., Haffari, G., Bashashati, A., Russell, R., McKinney, S., METABRIC Group, Langerød, A., Green, A., Provenzano, E., Wishart, G., Pinder, S., Watson, P., Markowetz, F., Murphy, L., Ellis, I., Purushotham, A., Børresen-Dale, A.-L., Brenton, J. D., Tavaré, S., Caldas, C., and Aparicio, S. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, **486**(7403), 346–352.

Dong, L., Qian, J., Chen, F., Fan, Y., and Long, J. (2019). LINC00461 promotes cell migration and invasion in breast cancer through mir-30a-5p/integrin β3 axis. *J. Cell. Biochem.*, **120**(4), 4851–4862.

Dykes, I. M. and Emanueli, C. (2017). Transcriptional and post-transcriptional gene regulation by long non-coding RNA. *Genomics Proteomics Bioinformatics*, **15**(3), 177–186.

Fletcher, M. N. C., Castro, M. A. A., Wang, X., de Santiago, I., O'Reilly, M., Chin, S.-F., Rueda, O. M., Caldas, C., Ponder, B. A. J., Markowetz, F., and Meyer, K. B. (2013). Master regulators of FGFR2 signalling and breast cancer risk. *Nat. Commun.*, **4**, 2464.

Foulkes, W. D., Smith, I. E., and Reis-Filho, J. S. (2010). Triple-negative breast cancer. *N. Engl. J. Med.*, **363**(20), 1938–1948.

Friedman, R. C., Farh, K. K.-H., Burge, C. B., and Bartel, D. P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.*, **19**(1), 92–105.

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., Cerami, E., Sander, C., and Schultz, N. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.*, **6**(269), l1.

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

Ghamrawi, N. and McCallum, A. (2005). Collective multi-label classification. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, CIKM '05, pages 195–200, New York, NY, USA. Association for Computing Machinery.

Gibbs, C. S., Jackson, C. A., Saldi, G.-A., Tjärnberg, A., Shah, A., Watters, A., De Veaux, N., Tchourine, K., Yi, R., Hamamsy, T., Castro, D. M., Carriero, N., Gorissen, B. L., Gresham, D., Miraldi, E. R., and Bonneau, R. (2022). High performance single-cell gene regulatory network inference at scale: The inferelator 3.0.

Goldman, M. J., Craft, B., Hastie, M., Repečka, K., McDade, F., Kamath, A., Banerjee, A., Luo, Y., Rogers, D., Brooks, A. N., Zhu, J., and Haussler, D. (2020). Visualizing and interpreting cancer genomics data via the xena platform. *Nat. Biotechnol.*, **38**(6), 675–678.

Hamam, R., Hamam, D., Alsaleh, K. A., Kassem, M., Zaher, W., Alfayez, M., Aldahmash, A., and Alajez, N. M. (2017). Circulating microRNAs in breast cancer: novel diagnostic and prognostic biomarkers. *Cell Death Dis.*, **8**(9), e3045.

Hanley, C. J., Henriet, E., Sirka, O. K., Thomas, G. J., and Ewald, A. J. (2020). Tumor-Resident stromal cells promote breast cancer invasion through regulation of the basal phenotype. *Mol. Cancer Res.*, **18**(11), 1615–1622.

Hänzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*, **14**, 7.

He, J., Wu, M., Xiong, L., Gong, Y., Yu, R., Peng, W., Li, L., Li, L., Tian, S., Wang, Y., Tao, Q., and Xiang, T. (2020). BTB/POZ zinc finger protein ZBTB16 inhibits breast cancer proliferation and metastasis through upregulating ZBTB28 and antagonizing BCL6/ZBTB27. *Clin. Epigenetics*, **12**(1), 82.

Hoffmann, M., Pachl, E., Hartung, M., Stiegler, V., Baumbach, J., Schulz, M. H., and List, M. (2021). SPONGEdb: a pan-cancer resource for competing endogenous RNA interactions. *NAR Cancer*, **3**(1).

Hou, J., Wang, Y., Zhang, H., Hu, Y., Xin, X., and Li, X. (2020). Silencing of LINC00461 enhances radiosensitivity of lung adenocarcinoma cells by down-regulating HOXA10 via microRNA-195. *J. Cell. Mol. Med.*, **24**(5), 2879–2890.

Huang, G.-W., Zhang, Y.-L., Liao, L.-D., Li, E.-M., and Xu, L.-Y. (2017). Natural antisense transcript TPM1-AS regulates the alternative splicing of tropomyosin I through an interaction with RNA-binding motif protein 4. *Int. J. Biochem. Cell Biol.*, **90**, 59–67.

Jerby-Arnon, L., Shah, P., Cuoco, M. S., Rodman, C., Su, M.-J., Melms, J. C., Leeson, R., Kanodia, A., Mei, S., Lin, J.-R., Wang, S., Rabasha, B., Liu, D., Zhang, G., Margolais, C., Ashenberg, O., Ott, P. A., Buchbinder, E. I., Haq, R., Hodi, F. S., Boland, G. M., Sullivan, R. J., Frederick, D. T., Miao, B., Moll, T., Flaherty, K. T., Herlyn, M., Jenkins, R. W., Thummalapalli, R., Kowalczyk, M. S., Cañadas, I., Schilling, B., Cartwright, A. N. R., Luoma, A. M., Malu, S., Hwu, P., Bernatchez, C., Forget, M.-A., Barbie, D. A., Shalek, A. K., Tirosh, I., Sorger, P. K., Wucherpfennig, K., Van Allen, E. M., Schadendorf, D., Johnson, B. E., Rotem, A., Rozenblatt-Rosen, O., Garraway, L. A., Yoon, C. H., Izar, B., and Regev, A. (2018). A cancer cell program promotes T cell exclusion and resistance to checkpoint blockade. *Cell*, **175**(4), 984–997.e24.

Ji, D., Wang, Y., Li, H., Sun, B., and Luo, X. (2019). Long non-coding RNA LINC00461/miR-149-5p/LRIG2 axis regulates hepatocellular carcinoma progression. *Biochem. Biophys. Res. Commun.*, **512**(2), 176–181.

Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., Li, M., Wang, G., and Liu, Y. (2009). miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res.*, **37**(Database issue), D98–104.

Kartha, R. V. and Subramanian, S. (2014). Competing endogenous RNAs (ceRNAs): new entrants to the intricacies of gene regulation. *Front. Genet.*, **5**, 8.

Kothari, C., Clemenceau, A., Ouellette, G., Ennour-Idrissi, K., Michaud, A., C-Gaudreault, R., Diorio, C., and Durocher, F. (2021). TBC1D9: An important modulator of tumorigenesis in breast cancer. *Cancers*, **13**(14).

Kuhn, M. (2008). Building predictive models inRUsing thecaretpackage. *J. Stat. Softw.*, **28**(5).

Lehmann, B. D., Bauer, J. A., Chen, X., Sanders, M. E., Chakravarthy, A. B., Shyr, Y., and Pietenpol, J. A. (2011). Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J. Clin. Invest.*, **121**(7), 2750–2767.

Liaw, A., Wiener, M., and Others (2002). Classification and regression by randomforest. *R news*, **2**(3), 18–22.

List, M., Dehghani Amirabad, A., Kostka, D., and Schulz, M. H. (2019). Large-scale inference of competing endogenous RNA networks with sparse partial correlation. *Bioinformatics*, **35**(14), i596–i604.

Liu, H., Lei, C., He, Q., Pan, Z., Xiao, D., and Tao, Y. (2018). Nuclear functions of mammalian MicroRNAs in gene regulation, immunity and cancer. *Mol. Cancer*, **17**(1), 64.

Ma, A., Wang, C., Chang, Y., Brennan, F. H., McDermaid, A., Liu, B., Zhang, C., Popovich, P. G., and Ma, Q. (2020). IRIS3: integrated cell-type-specific regulon inference server from single-cell RNA-Seq. *Nucleic Acids Res.*, **48**(W1), W275–W286.

McGeary, S. E., Lin, K. S., Shi, C. Y., Pham, T. M., Bisaria, N., Kelley, G. M., and Bartel, D. P. (2019). The biochemical basis of microRNA targeting efficacy. *Science*, **366**(6472).

Mulrane, L., McGee, S. F., Gallagher, W. M., and O'Connor, D. P. (2013). miRNA dysregulation in breast cancer. *Cancer Res.*, **73**(22), 6554–6562.

Muniategui, A., Pey, J., Planes, F. J., and Rubio, A. (2013). Joint analysis of miRNA and mRNA expression data. *Brief. Bioinform.*, **14**(3), 263–278.

Opsahl, T. (2009). *Structure and evolution of weighted networks*. Ph.D. thesis, Queen Mary, University of London.

Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S. X., Lønning, P. E., Børresen-Dale, A. L., Brown, P. O., and Botstein, D. (2000). Molecular portraits of human breast tumours. *Nature*, **406**(6797), 747–752.

Rahib, L., Wehner, M. R., Matrisian, L. M., and Nead, K. T. (2021). Estimated projection of US cancer incidence and death to 2040. *JAMA Netw Open*, **4**(4), e214708.

Roll, J. D., Rivenbark, A. G., Sandhu, R., Parker, J. S., Jones, W. D., Carey, L. A., Livasy, C. A., and Coleman, W. B. (2013). Dysregulation of the epigenome in triple-negative breast cancers: basal-like and claudin-low breast cancers express aberrant DNA hypermethylation. *Exp. Mol. Pathol.*, **95**(3), 276–287.

Salmena, L., Poliseno, L., Tay, Y., Kats, L., and Pandolfi, P. P. (2011). A ceRNA hypothesis: the rosetta stone of a hidden RNA language? *Cell*, **146**(3), 353–358.

Statello, L., Guo, C.-J., Chen, L.-L., and Huarte, M. (2021). Gene regulation by long non-coding RNAs and its biological functions. *Nat. Rev. Mol. Cell Biol.*, **22**(2), 96–118.

Su, X., Malouf, G. G., Chen, Y., Zhang, J., Yao, H., Valero, V., Weinstein, J. N., Spano, J.-P., Meric-Bernstam, F., Khayat, D., and Esteva, F. J. (2014). Comprehensive analysis of long non-coding RNAs in human breast cancer clinical subtypes. *Oncotarget*, **5**(20), 9864–9876.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.

Tay, Y., Rinn, J., and Pandolfi, P. P. (2014). The multilayered complexity of ceRNA crosstalk and competition. *Nature*, **505**(7483), 344–352.

Tsang, J. S., Ebert, M. S., and van Oudenaarden, A. (2010). Genome-wide dissection of microRNA functions and cotargeting networks using gene set signatures. *Mol. Cell*, **38**(1), 140–153.

Van de Sande, B., Flerin, C., Davie, K., De Waegeneer, M., Hulselmans, G., Aibar, S., Seurinck, R., Saelens, W., Cannoodt, R., Rouchon, Q., Verbeiren, T., De Maeyer, D., Reumers, J., Saeys, Y., and Aerts, S. (2020). A scalable SCENIC workflow for single-cell gene regulatory network analysis. *Nat. Protoc.*, **15**(7), 2247–2276.

Wang, P., Guo, Q., Qi, Y., Hao, Y., Gao, Y., Zhi, H., Zhang, Y., Sun, Y., Zhang, Y., Xin, M., Zhang, Y., Ning, S., and Li, X. (2022). LncACTdb 3.0: an updated database of experimentally supported ceRNA interactions and personalized networks contributing to precision medicine. *Nucleic Acids Res.*, **50**(D1), D183–D189.

Wang, S., Ni, B., Zhang, Z., Wang, C., Wo, L., Zhou, C., Zhao, Q., and Zhao, E. (2019). Long non-coding RNA DNM3OS promotes tumor progression and EMT in gastric cancer by associating with snail. *Biochem. Biophys. Res. Commun.*, **511**(1), 57–62.

Wei, L.-J., Bai, D.-M., Wang, Z.-Y., and Liu, B.-C. (2020). Upregulated lncRNA CACNA1G-AS1 aggravates the progression of colorectal cancer by downregulating p53. *Eur. Rev. Med. Pharmacol. Sci.*, **24**(1), 130–136.

Yang, J., Li, C., Li, H., and E, C. (2019). LncRNA CACNA1G-AS1 facilitates hepatocellular carcinoma progression through the miR-2392/C1orf61 pathway. *J. Cell. Physiol.*, **234**(10), 18415–18422.

Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., Torres-Garcia, W., Treviño, V., Shen, H., Laird, P. W., Levine, D. A., Carter, S. L., Getz, G., Stemke-Hale, K., Mills, G. B., and Verhaak, R. G. W. (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.*, **4**, 2612.

Yu, P.-F., Kang, A.-R., Jing, L.-J., and Wang, Y.-M. (2018). Long non-coding RNA CACNA1G-AS1 promotes cell migration, invasion and epithelial-mesenchymal transition by HNRNPA2B1 in non-small cell lung cancer. *Eur. Rev. Med. Pharmacol. Sci.*, **22**(4), 993–1002.

Zhang, H., Hua, Y., Jiang, Z., Yue, J., Shi, M., Zhen, X., Zhang, X., Yang, L., Zhou, R., and Wu, S. (2019). Cancer-associated fibroblast-promoted LncRNA DNM3OS confers radioresistance by regulating DNA damage response in esophageal squamous cell carcinoma. *Clin. Cancer Res.*, **25**(6), 1989–2000.

Zhang, J., Liu, L., Xu, T., Zhang, W., Li, J., Rao, N., and Le, T. D. (2021). Time to infer miRNA sponge modules. *Wiley Interdiscip. Rev. RNA*, page e1686.

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture