cture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,30)10 (-5,30)(1,0)10 (0,35)(0,-1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,30

Published online 31 July 2009 *Nucleic Acids Research, 2009, Vol. 37, No. 12* **1–9**
doi:10.1093/nar/gkn000

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] *

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

Received January 1, 2009; Revised February 1, 2009; Accepted March 1, 2009

ABSTRACT

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. 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.

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 (1), 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 (2). Mature miRNAs are short non-coding RNAs with lengths of 20-23 nucleotides that play an important role in regulating gene product abundance (3). miRNAs are involved in regulating at least half the genes in the human genome (4) and are dysregulated in many diseases (5). They function via the targeting of RNAs that will be degraded or whose translation will be impeded (6). 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) (7). Salmena et al. proposed the competitive endogenous RNA (ceRNA) hypothesis (8), which suggests that RNAs with miRNA binding sites compete for a limited pool of miRNAs, giving rise to a complex gene regulatory network (9). Past studies have identified genes that act as ceRNAs (8, 10) and have associated them with different RNA classes, like messenger RNAs (mRNAs), circular RNAs, pseudogenes,

*To whom correspondence should be addressed. Email: markus.list@tum.de

© 2008 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

cture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,-30)10 (-5,-30)(1,0)10 (0,-35)(0,1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,

"output" — 2022/6/22 — 5:14 — page 2 — #2
cture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,30)10 (-5,30)(1,0)10 (0,35)(0,-1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,30

transcripts of 3' untranslated regions (UTRs), and long non-coding RNAs (lncRNAs) (11). In coding genes, effective miRNA targeting (i.e., reducing transcript levels) largely occurs in 3'UTRs (12). 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 (13). It is important to note that the key actors in these modules are not the miRNAs but rather the ceRNAs that influence the expression of other ceRNAs by 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 (14). 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 (15). 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 (16), 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 (13) 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.

Here, we introduce spongEffects, a new method to (a) extract ceRNA modules from previously inferred ceRNA networks (e.g. with SPONGE (17)) and (b) to assesses their regulatory activity using enrichment scores as surrogates for module activity. These spongEffects scores (enrichment scores of the modules) 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): (a) it filters for significant interactions with meaningful effect size and identifies the most important ceRNAs via network centrality analysis; (b) for the subset of nodes with the highest centrality values, it constructs ceRNA modules by incorporating first-degree ceRNA network neighbors; (c) it performs single-sample gene set enrichment using the identified nodes, thus obtaining spongEffects scores; and (d) it uses the spongEffects scores to perform downstream machine learning tasks for classification and biomarker identification.

As an example of possible use cases, we show that spongEffects scores can be used to classify breast cancer subtypes with high accuracy. spongEffects modules uncover 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. We use our approach on two different breast cancer datasets, thus showing that the enrichment methods underlying spongEffects are robust to missing values and across different technologies (e.g. RNA-seq and microarray). Furthermore, the availability of two independet datasets allows us to test the generalization capabilities of predictive models built on spongEffects scores while controlling for overfitting. Finally, we show that our method has the potential to generate new hypotheses on the role of long non-coding RNAs, which may not be included in routine gene expression analyses.

MATERIALS 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 (18) as inputs for spongEffects. We downloaded log2(TPM + 0.001) transformed TCGA-BRCA expression data (RNAseq) from the XENA Browser (19). 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 (20) from cBioPortal (21) and used them as external testing datasets to validate our findings and control overfitting.

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 (22). The definitions of such centrality measures have been generalized for applications in weighted networks (23) to acknowledge that not all interactions are equally important. Here we use weighted degree centrality:

$$Degree_i = \sum_j^N x_{i,j} \quad [1]$$

$$Strength_i = \sum_j^N w_{i,j} \quad [2]$$

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

Equation [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 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). Equation [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 Equation [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,

cture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,-30)10 (-5,-30)(1,0)10 (0,-35)(0,1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,

"output" — 2022/6/22 — 5:14 — page 3 — #3
cture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,30)10 (-5,30)(1,0)10 (0,35)(0,-1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,30

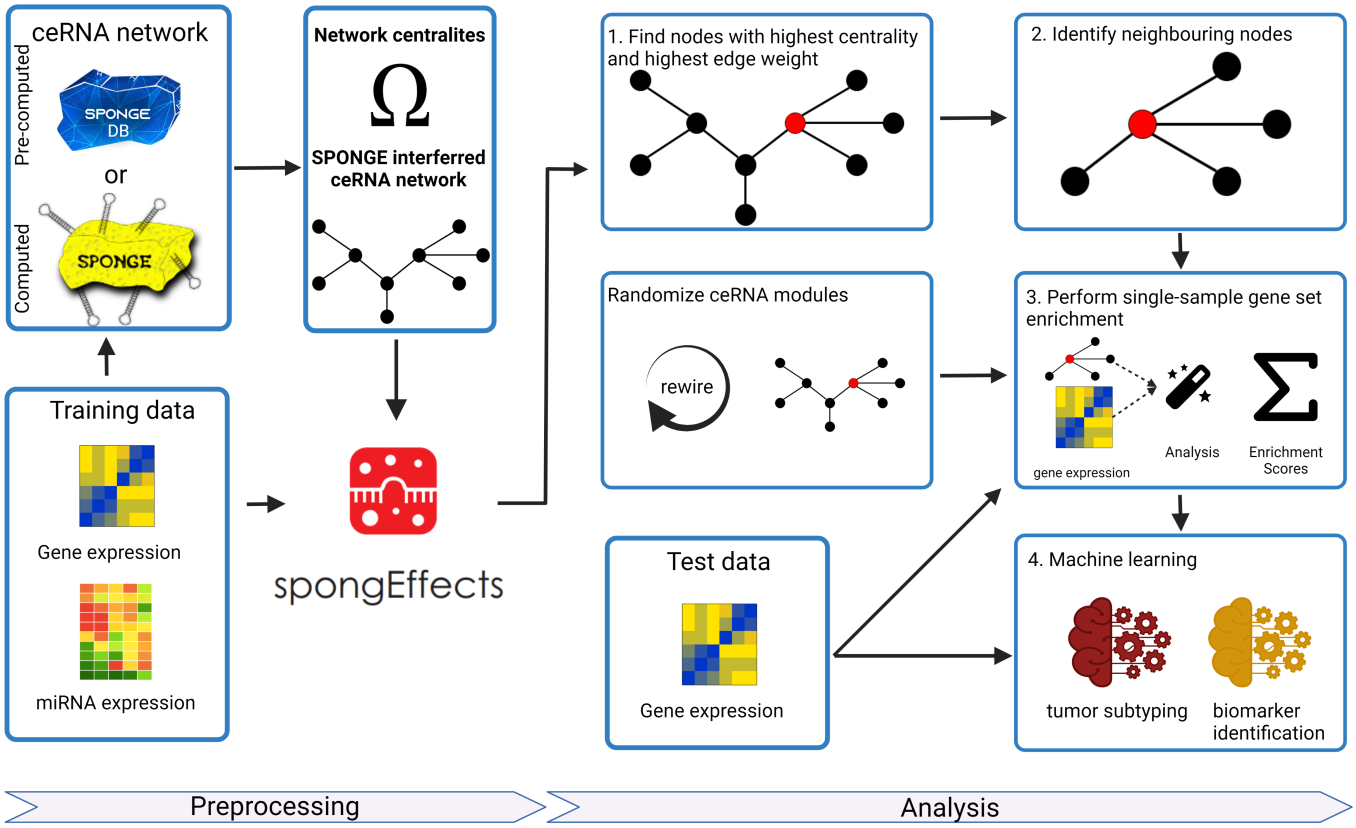

Figure 1. The workflow of spongEffects. spongEffects accepts a gene expression matrix and a ceRNA network as input. Subsequently, it a) filters the network and calculates weighted centrality scores to identify important nodes, b) identifies first neighbors, c) runs single-sample gene set enrichment,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

a low degree is favorable" (24). spongEffects uses weighted degree centrality as implemented in the R package tnet (version 3.0.16) (24) 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). The mscor extends the definition of sensitivity correlation by considering the effect of multiple miRNAs for the computation of the partial correlation (see List et al. (17)). We define ceRNA modules as subnetworks consisting of ceRNA genes with highest weighted centrality scores (central node) and their first-degree neighbors.

spongEffects scores

We implemented three different gene set enrichment approaches to aggregate the information associated with genes belonging to the newly defined spongEffects modules into unique enrichment scores, which we call spongEffects scores. These methods belong to the class of unsupervised and single-sample assessment tools (25), in that they do not require initial knowledge of existing phenotype groups, often unavailable, and they yield patient-specific scores for each module. In this work, we treat the genes in a module as a gene set. The approaches include single sample Gene Set Enrichment Analysis (ssGSEA) and Gene Set Variation analysis (GSVA) algorithms as implemented in the GSVA package (version 1.34.0) (25), and Overall Expression (OE) (26). Importantly, these approaches allow the calculation of the spongEffects scores even when some of the ceRNAs (e.g., circRNAs or lncRNAs) in the modules are not present in the input expression matrix, and result in module-by-sample score matrices that can be used for downstream analyses. When comparing results between GSVA, ssGSEA and OE (Suppl. Fig. 1) we observed no major differences in the results. 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.

Random Forest for subtype classification

We used Random Forest as implemented in the caret R package (version 6.0.90) (27) to classify tumor samples in the respective subtypes, using spongEffects scores calculated on the TCGA training dataset as input. We optimized the number of features, i.e. spongEffects modules, randomly sampled at each split (mtry parameter in the caret implementation) via repeated (3x) 10-fold cross-validation. We implemented a stratified approach in order to preserve the proportion of samples in each TCGA subtype at each iteration of the cross-validation procedure. Given the multi-classification problem at hand, we defined the optimal parameter as the one resulting in the highest subset accuracy (28) during cross-validation. After cross-validation, we used the learnt parameter to train a new Random Forest model on the complete TCGA training

cture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,-30)10 (-5,-30)(1,0)10 (0,-35)(0,1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,

"output" — 2022/6/22 — 5:14 — page 4 — #4
cture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,30)10 (-5,30)(1,0)10 (0,35)(0,-1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,30

set. The model was further tested to predict subtype labels from spongEffects scores calculated on the METABRIC expression data. METABRIC data were not part of the training procedure and were thus considered as a valid external testing dataset. The approach resulted in two performance metrics, subset accuracies for the training and testing sets, that were used to evaluate and compare Random Forest models. To identify modules driving the predictions, we calculated the Gini indexex as implemented in the randomForest package (version 4.6.14) (29).

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. An AIMe report has been generated at https://aime.report/UHkSyD (30).

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 spongEffects scores for functional interpretation and machine learning, we exemplarily consider the task of breast cancer subtyping.

spongEffects modules predict breast cancer subtypes

Breast cancer, the leading cause of cancer-related deaths in women (31), is a heterogeneous disease characterized by five subtypes (Luminal A (LumA), Luminal B (LumB), Basal, HER2-positive (Her2), and Normal) with different genetic prognostic profiles (32). Mechanisms of miRNA dysregulation play a role in breast cancer (33), offering the chance to validate miRNA mature strands as potential biomarkers (34). Here, we use spongEffects to study differences in miRNA regulation at the patient-specific level based on data from The Cancer Genome Consortium (35) (TCGA-BRCA, number of samples (n) = 1,063) and the Molecular Taxonomy of Breast Cancer International Consortium (20) (METABRIC cohorts 1 and 2, n = 1,905). In TCGA and METABRIC 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). 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 independent external validation set to control for potential overfitting (see Supplementary Table 1 for number of samples per subtype for TCGA and METABRIC).

spongEffects modules were calculated based on the TCGA-BRCA ceRNA network publicly available in SPONGEdb (18). The original network, containing ~3×10^7 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 (17). 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 spongEffects modules, motivated by the importance these molecules have in ceRNA networks, their potential as biomarkers (13), and their relevance in breast cancer subtypes (36). The 750 lncRNAs with the highest weighted centrality measures were selected and used as central nodes to define the spongEffects modules. We retain modules containing 10 to 200 genes after removing genes missing in the expression matrices, as previously suggested (25).

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 (26).

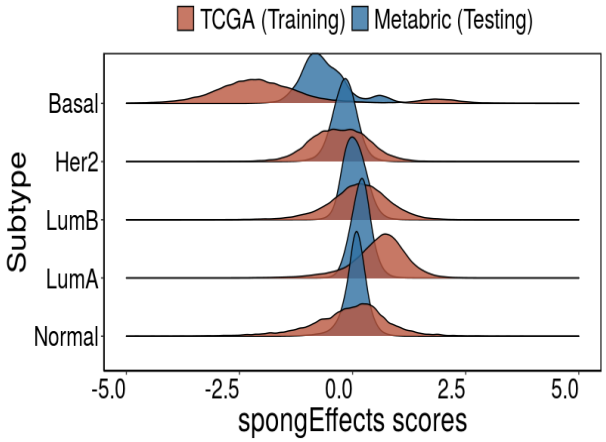

Figure 2. Distribution of the spongEffects scores over all modules in the TCGA-BRCA (training) and METABRIC (testing) dataset. With the exception of Basal breast cancer, the scores for the various breast cancer subtypes are approximately normally distributed. For the Basal subtype, the bimodality of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

Fig. 2 shows that spongeEffects scores over all found modules differ markedly between subtypes. We note that the distribution of the basal subtypes diverges from a normal-like distribution suggesting a hidden subgroup of patients. After disecting this distribution using the R package mixtools (version 1.2.0) (37), we find that the smaller group is enriched in stroma-high tumors (Fig. 2 and Suppl. Fig. 2.a), as calculated via ESTIMATE (38), and in extracellular matrix (ECM)-related genes, which have been associated with Basal invasion programs under the regulation of KRT14 related genes (39) (Suppl. Fig. 2.b). Analogous studies have identified similar Basal subgroups from different data types and cohorts (40). 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 the relatively small influence of the number of initial lncRNAs (and thus

cture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,-30)10 (-5,-30)(1,0)10 (0,-35)(0,1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,

"output" — 2022/6/22 — 5:14 — page 5 — #5
cture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,30)10 (-5,30)(1,0)10 (0,35)(0,-1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,30

modules) considered on the final results (Suppl. Fig. 3). To assess whether spongEffects scores offer additional value over the expression level of the ceRNAs alone, we compared our results to the predictive performane of randomly defined modules and a baseline model trained on the expression values (not on spongEffects scores) of the 25 lncRNAs common to both datasets. The subset accuracy metric (28) 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).

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 (41) and non-small cell lung cancer (42), and to downregulate p53 levels in colorectal cancer (43). In contrast, DNM3OS has been shown to promote tumor progression and SNAIL-mediated EMT in gastric cancers (44) 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 (45). Modules such as LINC00461 show strong enrichment in Basal samples. LINC00461 has been shown to promote migration and invasion in breast cancer (46, 47) and to play a role in progression in other cancer types such as hepatocellular carcinoma (48) and renal cell carcinoma (49), or in mediating radiosensitivity in lung adenocarcinoma (50). 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 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.

Predictive spongEffects modules can identify potential biomarkers in breast cancer

Here, as an example, we describe in more detail the CACNA1G-AS1 module, and its miRNAs, which have been experimentally validated. Specifically, we investigate its role in the Basal subtype, one of the most aggressive breast cancer subtypes for which effective targeted therapies are still missing (51, 52).

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 (53), while genes such as MYB and ZBTB16 are tumor suppressors whose promoter regions are often hypermethylated in Basal breast cancers (54, 55).

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 had 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).

CONCLUSION

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

cture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,-30)10 (-5,-30)(1,0)10 (0,-35)(0,1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,

"output" — 2022/6/22 — 5:14 — page 6 — #6
cture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,30)10 (-5,30)(1,0)10 (0,35)(0,-1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,30

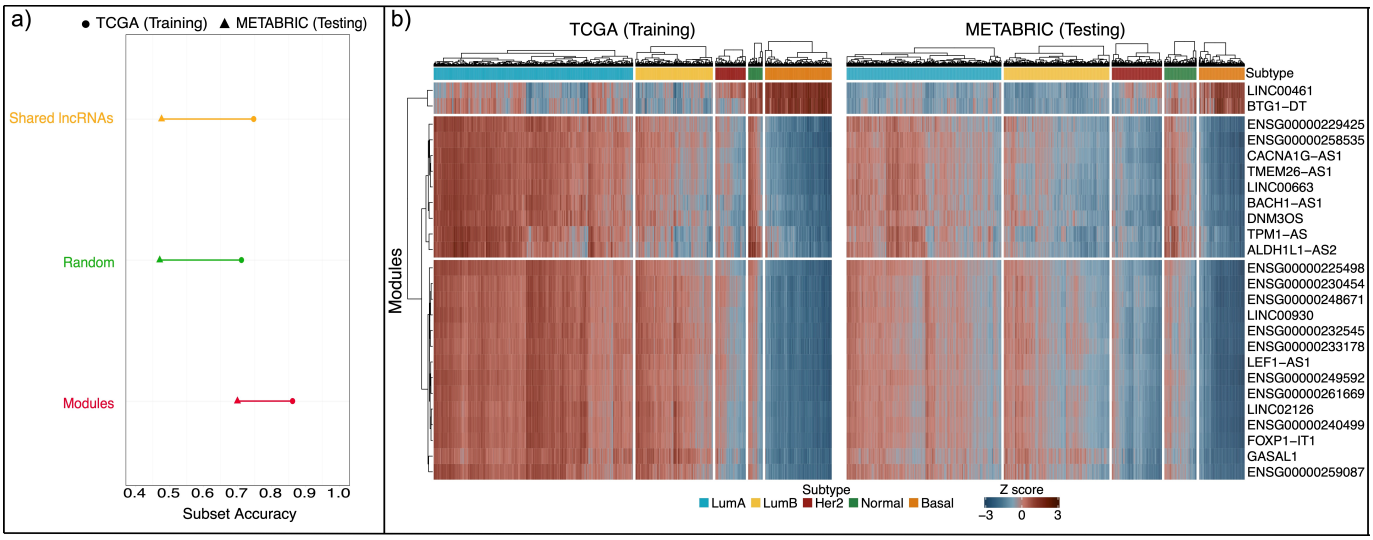

Figure 3. Overview of Results. Overview of performance metrics for the Random Forest models calibrated on spongEffects modules (red), randomly defined modules (green), and lncRNAs common to both the TCGA (training) and METABRIC (testing) datasets (yellow). Subset accuracy (the proportion of samples that 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. .

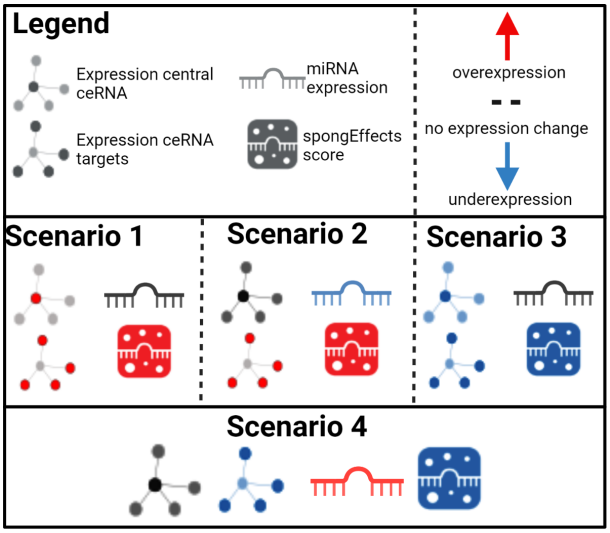

Figure 4. Understanding increments and reductions in spongEffects scores. Increments in spongEffects scores can result from: Scenario 1) an increase in the expression of a module's central ceRNA while miRNA expression levels remain constant, leading the to upregulation of the module's target ceRNAs, 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.

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 Section 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 (56). 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 (7). 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 (57, 58). We foresee two interesting directions for further research. First, recent studies on the inference of regulons (57, 59) 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 (60, 61, 62), paving the way to the design of methods with the potential of uncovering new insights into the complexity of tumor biology.

cture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,-30)10 (-5,-30)(1,0)10 (0,-35)(0,1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,

"output" — 2022/6/22 — 5:14 — page 7 — #7
cture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,30)10 (-5,30)(1,0)10 (0,35)(0,-1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,30)10 (-5,30)(1,0)10 (0,35)(0,-1)30 pictu

DATA AVAILABILITY

SPONGE and spongEffects are available via bioconductor and github:
https://bioconductor.org/packages/devel/bioc/html/SPONGE.html
https://github.com/biomedbigdata/SPONGE

SPONGEdb is available at:
http://sponge.biomedical-big-data.de/
spongEffects scores calculated for TCGA datasets: https://doi.org/10.6084/m9.figshare.19328885.v1

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.

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.

Conflict of interest statement. None declared.

REFERENCES

1. Lola Rahib, Mackenzie R Wehner, Lynn M Matrisian, and Kevin T Nead. Estimated projection of US cancer incidence and death to 2040. *JAMA Netw Open*, 4(4):e214708, April 2021.
2. Hongyu Liu, Cheng Lei, Qin He, Zou Pan, Desheng Xiao, and Yongguang Tao. Nuclear functions of mammalian MicroRNAs in gene regulation, immunity and cancer. *Mol. Cancer*, 17(1):64, February 2018.
3. Reena V Kartha and Subbaya Subramanian. Competing endogenous RNAs (ceRNAs): new entrants to the intricacies of gene regulation. *Front. Genet.*, 5:8, January 2014.
4. Robin C Friedman, Kyle Kai-How Farh, Christopher B Burge, and David P Bartel. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.*, 19(1):92–105, January 2009.
5. Qinghua Jiang, Yadong Wang, Yangyang Hao, Liran Juan, Mingxiang Teng, Xinjun Zhang, Meimei Li, Guohua Wang, and Yunlong Liu. miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res.*, 37(Database issue):D98–104, January 2009.
6. D P Bartel. MicroRNAs: target recognition and regulatory functions. *Cell*, 2009.
7. Sean E McGeary, Kathy S Lin, Charlie Y Shi, Thy M Pham, Namita Bisaria, Gina M Kelley, and David P Bartel. The biochemical basis of microRNA targeting efficacy. *Science*, 366(6472), December 2019.
8. Leonardo Salmena, Laura Poliseno, Yvonne Tay, Lev Kats, and Pier Paolo Pandolfi. A ceRNA hypothesis: the rosetta stone of a hidden RNA language? *Cell*, 146(3):353–358, August 2011.
9. John S Tsang, Margaret S Ebert, and Alexander van Oudenaarden. Genome-wide dissection of microRNA functions and cotargeting networks using gene set signatures. *Mol. Cell*, 38(1):140–153, April 2010.
10. Yvonne Tay, John Rinn, and Pier Paolo Pandolfi. The multilayered complexity of ceRNA crosstalk and competition. *Nature*, 505(7483):344–352, January 2014.
11. Iain M Dykes and Costanza Emanueli. Transcriptional and post-transcriptional gene regulation by long non-coding RNA. *Genomics Proteomics Bioinformatics*, 15(3):177–186, June 2017.
12. Vikram Agarwal, George W Bell, Jin-Wu Nam, and David P Bartel. Predicting effective microRNA target sites in mammalian mRNAs. *Elife*, 4, August 2015.
13. Peng Wang, Qiuyan Guo, Yue Qi, Yangyang Hao, Yue Gao, Hui Zhi, Yuanfu Zhang, Yue Sun, Yakun Zhang, Mengyu Xin, Yunpeng Zhang, Shangwei Ning, and Xia Li. 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, January 2022.
14. Ander Muniategui, Jon Pey, Francisco J Planes, and Angel Rubio. Joint analysis of miRNA and mRNA expression data. *Brief. Bioinform.*, 14(3):263–278, May 2013.
15. Sarvenaz Choobdar, Mehmet E Ahsen, Jake Crawford, Mattia Tomasoni, Tao Fang, David Lamparter, Junyuan Lin, Benjamin Hescott, Xiaozhe Hu, Johnathan Mercer, Ted Natoli, Rajiv Narayan, DREAM Module Identification Challenge Consortium, Aravind Subramanian, Jitao D Zhang, Gustavo Stolovitzky, Zoltán Kutalik, Kasper Lage, Donna K Slonim, Julio Saez-Rodriguez, Lenore J Cowen, Sven Bergmann, and Daniel Marbach. Assessment of network module identification across complex diseases. *Nat. Methods*, 16(9):843–852, September 2019.
16. Junpeng Zhang, Lin Liu, Taosheng Xu, Wu Zhang, Jiuyong Li, Nini Rao, and Thuc Duy Le. Time to infer miRNA sponge modules. *Wiley Interdiscip. Rev. RNA*, page e1686, August 2021.
17. Markus List, Azim Dehghani Amirabad, Dennis Kostka, and Marcel H Schulz. Large-scale inference of competing endogenous RNA networks with sparse partial correlation. *Bioinformatics*, 35(14):i596–i604, July 2019.
18. Markus Hoffmann, Elisabeth Pachl, Michael Hartung, Veronika Stiegler, Jan Baumbach, Marcel H Schulz, and Markus List. SPONGEdb: a pan-cancer resource for competing endogenous RNA interactions. *NAR Cancer*, 3(1), 2021.
19. Mary J Goldman, Brian Craft, Mim Hastie, Kristupas Repečka, Fran McDade, Akhil Kamath, Ayan Banerjee, Yunhai Luo, Dave Rogers, Angela N Brooks, Jingchun Zhu, and David Haussler. Visualizing and interpreting cancer genomics data via the xena platform. *Nat. Biotechnol.*, 38(6):675–678, June 2020.
20. Christina Curtis, Sohrab P Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M Rueda, Mark J Dunning, Doug Speed, Andy G Lynch, Shamith Samarajiwa, Yinyin Yuan, Stefan Gräf, Gavin Ha, Gholamreza Haffari, Ali Bashashati, Roslin Russell, Steven McKinney, METABRIC Group, Anita Langerød, Andrew Green, Elena Provenzano, Gordon Wishart, Sarah Pinder, Peter Watson, Florian Markowetz, Leigh Murphy, Ian Ellis, Arnie Purushotham, Anne-Lise Børresen-Dale, James D Brenton, Simon Tavaré, Carlos Caldas, and Samuel Aparicio. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, April 2012.
21. Jianjiong Gao, Bülent Arman Aksoy, Ugur Dogrusoz, Gideon Dresdner, Benjamin Gross, S Onur Sumer, Yichao Sun, Anders Jacobsen, Rileen Sinha, Erik Larsson, Ethan Cerami, Chris Sander, and Nikolaus Schultz. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.*, 6(269):l1, April 2013.
22. Xionglei He and Jianzhi Zhang. Why do hubs tend to be essential in protein networks? *PLoS Genet.*, 2(6):e88, June 2006.
23. A Barrat, M Barthélemy, R Pastor-Satorras, and A Vespignani. The architecture of complex weighted networks. *Proc. Natl. Acad. Sci. U. S. A.*, 101(11):3747–3752, March 2004.
24. Tore Opsahl. *Structure and evolution of weighted networks*. PhD thesis, Queen Mary, University of London, 2009.
25. Sonja Hänzelmann, Robert Castelo, and Justin Guinney. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*, 14:7, January 2013.
26. Livnat Jerby-Arnon, Parin Shah, Michael S Cuoco, Christopher Rodman, Mei-Ju Su, Johannes C Melms, Rachel Leeson, Abhay Kanodia, Shaolin Mei, Jia-Ren Lin, Shu Wang, Bokang Rabasha, David Liu, Gao Zhang, Claire Margolais, Orr Ashenberg, Patrick A Ott, Elizabeth I

cture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,-30)10 (-5,-30)(1,0)10 (0,-35)(0,1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,-30)10 (-5,-30)(1,0)10 (0,-35)(0,1)30 p

"output" — 2022/6/22 — 5:14 — page 8 — #8
cture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,30)10 (-5,30)(1,0)10 (0,35)(0,-1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,30)10 (-5,30)(1,0)10 (0,35)(0,-1)30 pictu

Buchbinder, Rizwan Haq, F Stephen Hodi, Genevieve M Boland, Ryan J Sullivan, Dennie T Frederick, Benchun Miao, Tabea Moll, Keith T Flaherty, Meenhard Herlyn, Russell W Jenkins, Rohit Thummalapalli, Monika S Kowalczyk, Israel Cañadas, Bastian Schilling, Adam N R Cartwright, Adrienne M Luoma, Shruti Malu, Patrick Hwu, Chantale Bernatchez, Marie-Andrée Forget, David A Barbie, Alex K Shalek, Itay Tirosh, Peter K Sorger, Kai Wucherpfennig, Eliezer M Van Allen, Dirk Schadendorf, Bruce E Johnson, Asaf Rotem, Orit Rozenblatt-Rosen, Levi A Garraway, Charles H Yoon, Benjamin Izar, and Aviv Regev. A cancer cell program promotes T cell exclusion and resistance to checkpoint blockade. *Cell*, 175(4):984–997.e24, November 2018.

27. Max Kuhn. Building predictive models inRUsing thecaretpackage. *J. Stat. Softw.*, 28(5), 2008.
28. Nadia Ghamrawi and Andrew McCallum. 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, October 2005. Association for Computing Machinery.
29. Andy Liaw, Matthew Wiener, and Others. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.
30. Julian Matschinske, Nicolas Alcaraz, Arriel Benis, Martin Golebiewski, Dominik G Grimm, Lukas Heumos, Tim Kacprowski, Olga Lazareva, Markus List, Zakaria Louadi, Josch K Pauling, Nico Pfeifer, Richard Röttger, Veit Schwämmle, Gregor Sturm, Alberto Traverso, Kristel Van Steen, Martiela Vaz de Freitas, Gerda Cristal Villalba Silva, Leonard Wee, Nina K Wenke, Massimiliano Zanin, Olga Zolotareva, Jan Baumbach, and David B Blumenthal. The AIMe registry for artificial intelligence in biomedical research. *Nat. Methods*, 18(10):1128–1131, October 2021.
31. Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, 2021.
32. C M Perou, T Sørlie, M B Eisen, M van de Rijn, S S Jeffrey, C A Rees, J R Pollack, D T Ross, H Johnsen, L A Akslen, O Fluge, A Pergamenschikov, C Williams, S X Zhu, P E Lønning, A L Børresen-Dale, P O Brown, and D Botstein. Molecular portraits of human breast tumours. *Nature*, 406(6797):747–752, August 2000.
33. Laoighse Mulrane, Sharon F McGee, William M Gallagher, and Darran P O'Connor. miRNA dysregulation in breast cancer. *Cancer Res.*, 73(22):6554–6562, November 2013.
34. Rimi Hamam, Dana Hamam, Khalid A Alsaleh, Moustapha Kassem, Waleed Zaher, Musaad Alfayez, Abdullah Aldahmash, and Nehad M Alajez. Circulating microRNAs in breast cancer: novel diagnostic and prognostic biomarkers. *Cell Death Dis.*, 8(9):e3045, September 2017.
35. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, October 2012.
36. Xiaoping Su, Gabriel G Malouf, Yunxin Chen, Jianping Zhang, Hui Yao, Vicente Valero, John N Weinstein, Jean-Philippe Spano, Funda Meric-Bernstam, David Khayat, and Francisco J Esteva. Comprehensive analysis of long non-coding RNAs in human breast cancer clinical subtypes. *Oncotarget*, 5(20):9864–9876, October 2014.
37. Tatiana Benaglia, Didier Chauveau, David R Hunter, and Derek S Young. mixtools: An R package for analyzing mixture models. *J. Stat. Softw.*, 32:1–29, 2010.
38. Kosuke Yoshihara, Maria Shahmoradgoli, Emmanuel Martínez, Rahulsimham Vegesna, Hoon Kim, Wandaliz Torres-Garcia, Victor Treviño, Hui Shen, Peter W Laird, Douglas A Levine, Scott L Carter, Gad Getz, Katherine Stemke-Hale, Gordon B Mills, and Roel G W Verhaak. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.*, 4:2612, 2013.
39. Christopher J Hanley, Elodie Henriet, Orit Katarina Sirka, Gareth J Thomas, and Andrew J Ewald. Tumor-Resident stromal cells promote breast cancer invasion through regulation of the basal phenotype. *Mol. Cancer Res.*, 18(11):1615–1622, November 2020.
40. Karama Asleh, Gian Luca Negri, Sandra E Spencer Miko, Shane Colborne, Christopher S Hughes, Xiu Q Wang, Dongxia Gao, C Blake Gilks, Stephen K L Chia, Torsten O Nielsen, and Gregg B Morin. Proteomic analysis of archival breast cancer clinical specimens identifies biological subtypes with distinct survival outcomes. *Nat. Commun.*, 13(1):896, February 2022.
41. Jinghui Yang, Chunsheng Li, Hang Li, and Changyong E. LncRNA CACNA1G-AS1 facilitates hepatocellular carcinoma progression through the miR-2392/C1orf61 pathway. *J. Cell. Physiol.*, 234(10):18415–18422, August 2019.
42. P-F Yu, A-R Kang, L-J Jing, and Y-M Wang. 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, February 2018.
43. L-J Wei, D-M Bai, Z-Y Wang, and B-C Liu. Upregulated lncRNA CACNA1G-AS1 aggravates the progression of colorectal cancer by downregulating p53. *Eur. Rev. Med. Pharmacol. Sci.*, 24(1):130–136, January 2020.
44. Shuchang Wang, Bo Ni, Zizhen Zhang, Chaojie Wang, Lulu Wo, Cixiang Zhou, Qian Zhao, and Enhao Zhao. 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, March 2019.
45. Guo-Wei Huang, Ying-Li Zhang, Lian-Di Liao, En-Min Li, and Li-Yan Xu. 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, September 2017.
46. Hongfang Zhang, Yuhui Hua, Zhenzhen Jiang, Jing Yue, Ming Shi, Xiaoli Zhen, Xiaoyan Zhang, Ling Yang, Rongjing Zhou, and Shixiu Wu. 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, March 2019.
47. Lifeng Dong, Junbin Qian, Fangfang Chen, Yangfan Fan, and Jingpei Long. LINC00461 promotes cell migration and invasion in breast cancer through mir-30a-5p/integrin $\beta 3$ axis. *J. Cell. Biochem.*, 120(4):4851–4862, April 2019.
48. Degang Ji, Yue Wang, Hang Li, Baozhen Sun, and Xiao Luo. Long non-coding RNA LINC00461/miR-149-5p/LRIG2 axis regulates hepatocellular carcinoma progression. *Biochem. Biophys. Res. Commun.*, 512(2):176–181, April 2019.
49. Lan Chen, Tao Zhang, Shanshan Zhang, Jinxing Huang, Genxi Zhang, Kaizhou Xie, Jinyu Wang, Haiqing Wu, and Guojun Dai. Identification of long Non-Coding RNA-Associated competing endogenous RNA network in the differentiation of chicken preadipocytes. *Genes*, 10(10), October 2019.
50. Jiqiu Hou, Yanjun Wang, Hongmei Zhang, Yuxin Hu, Xiuqin Xin, and Xiaodan Li. Silencing of LINC00461 enhances radiosensitivity of lung adenocarcinoma cells by down-regulating HOXA10 via microRNA-195. *J. Cell. Mol. Med.*, 24(5):2879–2890, March 2020.
51. William D Foulkes, Ian E Smith, and Jorge S Reis-Filho. Triple-negative breast cancer. *N. Engl. J. Med.*, 363(20):1938–1948, November 2010.
52. Brian D Lehmann, Joshua A Bauer, Xi Chen, Melinda E Sanders, A Bapsi Chakravarthy, Yu Shyr, and Jennifer A Pietenpol. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J. Clin. Invest.*, 121(7):2750–2767, July 2011.
53. Charu Kothari, Alisson Clemenceau, Geneviève Ouellette, Kaoutar Ennour-Idrissi, Annick Michaud, René C-Gaudreault, Caroline Diorio, and Francine Durocher. TBC1D9: An important modulator of tumorigenesis in breast cancer. *Cancers*, 13(14), July 2021.
54. Jin He, Mingjun Wu, Lei Xiong, Yijia Gong, Renjie Yu, Weiyan Peng, Lili Li, Li Li, Shaorong Tian, Yan Wang, Qian Tao, and Tingxiu Xiang. BTB/POZ zinc finger protein ZBTB16 inhibits breast cancer proliferation and metastasis through upregulating ZBTB28 and antagonizing BCL6/ZBTB27. *Clin. Epigenetics*, 12(1):82, June 2020.
55. J Devon Roll, Ashley G Rivenbark, Rupninder Sandhu, Joel S Parker, Wendell D Jones, Lisa A Carey, Chad A Livasy, and William B Coleman. 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, December 2013.
56. Luisa Statello, Chun-Jie Guo, Ling-Ling Chen, and Maite Huarte. Gene regulation by long non-coding RNAs and its biological functions. *Nat. Rev. Mol. Cell Biol.*, 22(2):96–118, February 2021.
57. Mauro A A Castro, Ines de Santiago, Thomas M Campbell, Courtney Vaughn, Theresa E Hickey, Edith Ross, Wayne D Tilley, Florian Markowetz, Bruce A J Ponder, and Kerstin B Meyer. Regulators of genetic risk of breast cancer identified by integrative network analysis. *Nat. Genet.*, 48(1):12–21, January 2016.
58. Vinicius S Chagas, Clarice S Groeneveld, Kelin G Oliveira, Sheyla Trefflich, Rodrigo C de Almeida, Bruce A J Ponder, Kerstin B Meyer, Steven J M Jones, A Gordon Robertson, and Mauro A A Castro. RTNduals: an R/Bioconductor package for analysis of co-regulation and inference of dual regulons. *Bioinformatics*, 35(24):5357–5358, December 2019.

cture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,-30)10 (-5,-30)(1,0)10 (0,-35)(0,1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,-30)10 (-5,-30)(1,0)10 (0,-35)(0,1)30 p

"output" — 2022/6/22 — 5:14 — page 9 — #9
cture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,30)10 (-5,30)(1,0)10 (0,35)(0,-1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,30

59. Michael N C Fletcher, Mauro A A Castro, Xin Wang, Ines de Santiago, Martin O'Reilly, Suet-Feung Chin, Oscar M Rueda, Carlos Caldas, Bruce A J Ponder, Florian Markowetz, and Kerstin B Meyer. Master regulators of FGFR2 signalling and breast cancer risk. *Nat. Commun.*, 4:2464, 2013.
60. Bram Van de Sande, Christopher Flerin, Kristofer Davie, Maxime De Waegeneer, Gert Hulselmans, Sara Aibar, Ruth Seurinck, Wouter Saelens, Robrecht Cannoodt, Quentin Rouchon, Toni Verbeiren, Dries De Maeyer, Joke Reumers, Yvan Saeys, and Stein Aerts. A scalable SCENIC workflow for single-cell gene regulatory network analysis. *Nat. Protoc.*, 15(7):2247–2276, July 2020.
61. Anjun Ma, Cankun Wang, Yuzhou Chang, Faith H Brennan, Adam McDermaid, Bingqiang Liu, Chi Zhang, Phillip G Popovich, and Qin Ma. IRIS3: integrated cell-type-specific regulon inference server from single-cell RNA-Seq. *Nucleic Acids Res.*, 48(W1):W275–W286, July 2020.
62. Claudia Skok Gibbs, Christopher A Jackson, Giuseppe-Antonio Saldi, Andreas Tjärnberg, Aashna Shah, Aaron Watters, Nicholas De Veaux, Konstantine Tchourine, Ren Yi, Tymor Hamamsy, Dayanne M Castro, Nicholas Carriero, Bram L Gorissen, David Gresham, Emily R Miraldi, and Richard Bonneau. High performance single-cell gene regulatory network inference at scale: The inferelator 3.0. February 2022.

cture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,-30)10 (-5,-30)(1,0)10 (0,-35)(0,1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,