

# Scanning sample-specific miRNA regulation from bulk and single-cell RNA-sequencing data

Junpeng Zhang<sup>1,\*</sup>, Lin Liu<sup>2</sup>, Xuemei Wei<sup>1</sup>, Chunwen Zhao<sup>1</sup>, Yanbi Luo<sup>1</sup>, Jiuyong Li<sup>2</sup> and Thuc Duy Le<sup>2,\*</sup>

<sup>1</sup>School of Engineering, Dali University, Dali, China

<sup>2</sup>UniSA STEM, University of South Australia, Mawson Lakes, Australia

\*To whom correspondence should be addressed. Email: zhangjunpeng411@gmail.com

Correspondence may also be addressed to Thuc Duy Le. Email: thuc.le@unisa.edu.au

## Abstract

RNA-sequencing technology provides an effective tool for understanding miRNA regulation in complex human diseases, including cancers. A large number of computational methods have been developed to make use of bulk and single-cell RNA-sequencing data to identify miRNA regulations at the resolution of multiple samples (i.e. group of cells or tissues). However, due to the heterogeneity of individual samples, there is a strong need to infer miRNA regulation specific to individual samples to uncover miRNA regulation at single-sample resolution level. Here, we develop a framework, Scan, for scanning sample-specific miRNA regulation. Since a single network inference method or strategy cannot perform well for all types of new data, Scan incorporates 27 network inference methods and two strategies to infer tissue-specific or cell-specific miRNA regulation from bulk or single-cell RNA-

sequencing data. Results on bulk and single-cell RNA-sequencing data demonstrate the effectiveness of Scan in inferring sample-specific miRNA regulation. Moreover, we have found that incorporating priori information of miRNA targets can improve the accuracy of miRNA target prediction. In addition, Scan can contribute to the clustering cells/tissues and construction of cell/tissue correlation networks. Finally, the comparison results have shown that the performance of network inference methods is likely to be data-specific, and selecting optimal network inference methods is required for more accurate prediction of miRNA targets. We have made Scan freely available to the public to help infer sample-specific miRNA regulation for new data, benchmark new network inference methods and deepen the understanding of miRNA regulation at the resolution of individual samples.

## Introduction

As a type of regulatory non-coding RNA (ncRNA), microRNA (miRNA, 19-25 nt in length) can modulate the expression levels of thousands of coding RNAs, thereby playing a key role in biological processes, signaling pathways and pathological processes of human cancers<sup>1-3</sup>. Due to the importance of miRNAs in biological system, they have the potential to be biomarkers of human cancers, e.g. breast cancer<sup>4</sup>, leukemia<sup>5</sup> and prostate cancer<sup>6</sup> in clinical applications. In addition to the miRNA-directed regulation which results in inhibition or degradation of messenger RNAs (mRNAs), miRNAs can also act as mediators of the

crosstalk between other ncRNAs (e.g. long non-coding RNAs, circular RNAs and pseudogenes) and mRNAs<sup>7,8</sup>. The miRNA-mediated or miRNA-indirect regulation is also known as competing endogenous RNA (ceRNA) regulation, and has been demonstrated to be involved in the regulatory mechanism of human cancers<sup>9-11</sup>.

There are multiple types of biological networks, including gene regulatory network within a biological sample (cell or tissue). Among the different types of biological networks, gene regulatory network is regarded as an important feature of biological samples, and gives rise to the unique characteristics of each biological sample<sup>12</sup>. In the gene regulation field, miRNA regulation has attracted broad attentions on account of its potential clinical translation. Therefore, inferring and characterizing miRNA regulation is a central problem for revealing miRNA regulatory mechanisms in systems biology. Just like dynamic biological system in a biological sample (cell or tissue), miRNA regulation also tends to be dynamic. For example, by using bulk and single-cell RNA-sequencing data, existing studies<sup>13-17</sup> have found that the miRNA regulatory networks are condition-specific, tissue-specific and cell-specific. Moreover, it is widely known that biological samples are characterized by heterogeneity, and are shown to be unique. Thus, to understand the miRNA regulation specific to biological samples (cancer tissues or cells in particular), it is essential to investigate sample-specific miRNA regulation (i.e. the miRNA regulatory network for a specific sample).

At the single-sample level, previous methods including CeSpGRN<sup>12</sup>, CSN<sup>18</sup>, DEVC-net<sup>19</sup>, EdgeBiomarker<sup>20</sup>, SSN<sup>21</sup>, LIONESS<sup>22</sup>, c-CSN<sup>23</sup> and LocCSN<sup>24</sup> have been presented to identify sample-specific gene regulation (one gene regulatory network for one sample), and the identified gene regulatory networks with these methods are undirected networks. To infer causal regulatory networks (represented as directed networks) specific to individual samples, an inference method<sup>25</sup> using drift-diffusion processes is also proposed. However, the above computational methods<sup>12,18-25</sup> are designed for exploring sample-specific transcriptional regulation rather than sample-specific miRNA regulation. To study miRNA regulation at single-cell level, CSMiR<sup>16</sup> is introduced to infer cell-specific miRNA regulation from single-cell miRNA-mRNA co-sequencing data (where each single cell is considered as a sample). As stated in CSMiR, for single-cell miRNA-mRNA co-sequencing data with less than 100 single-cells, CSMiR needs to interpolate pseudo-cells by using a re-sampling technique. However, the original single-cell miRNA-mRNA co-sequencing data itself contains technical noise (e.g. high dropout rate), thus as a result, the enlarged data by adding pseudo-cells may aggravate the technical noise.

To model the dynamic regulatory processes of miRNAs at the single-sample level, in this work, we propose a Sample-specific miRNA regulation (Scan) framework to scan sample-specific miRNA regulation from bulk and single-cell RNA-sequencing data. Due to the fact that that a single network inference method or strategy does not always perform the best in all

types of new data, Scan incorporates 27 network inference methods and two strategies to infer tissue-specific or cell-specific miRNA regulation from bulk or single-cell RNA-sequencing data. Given bulk or single-cell RNA-sequencing data with or without priori information of miRNA-mRNA interactions, 27 well-used network inference methods in Scan can be used to calculate the strength between miRNAs and mRNAs. Based on the strength between miRNAs and mRNAs, Scan adapts two strategies: statistical perturbation<sup>21</sup> and linear interpolation<sup>22</sup> to infer sample-specific miRNA regulatory networks.

By applying Scan to two RNA-sequencing datasets, a bulk dataset from breast cancer tissues, and a single-cell dataset from chronic myelogenous leukemia cells, we demonstrate the effectiveness of Scan in terms of accuracy and efficiency of miRNA target prediction. Moreover, the performance comparison between with using and without using priori information indicates that the priori information of miRNA targets can improve the accuracy of miRNA target prediction. We also reveal that Scan can help to cluster samples and construct sample correlation network. The comparison with CSmiR shows that with using priori information of miRNA targets, Scan mostly has better or similar performance compared with CSmiR<sup>16</sup> in inferring cell-specific miRNA regulation from the chronic myelogenous leukemia dataset. Furthermore, without using priori information of miRNA targets, the performance of Scan is mostly better than or comparable to the performance of CSmiR in inferring tissue-specific miRNA regulation from the breast cancer dataset.

Altogether, Scan provides a useful way to identify sample-specific miRNA regulation in human cancers, and can help to elucidate miRNA regulation mechanisms at the resolution of individual samples.

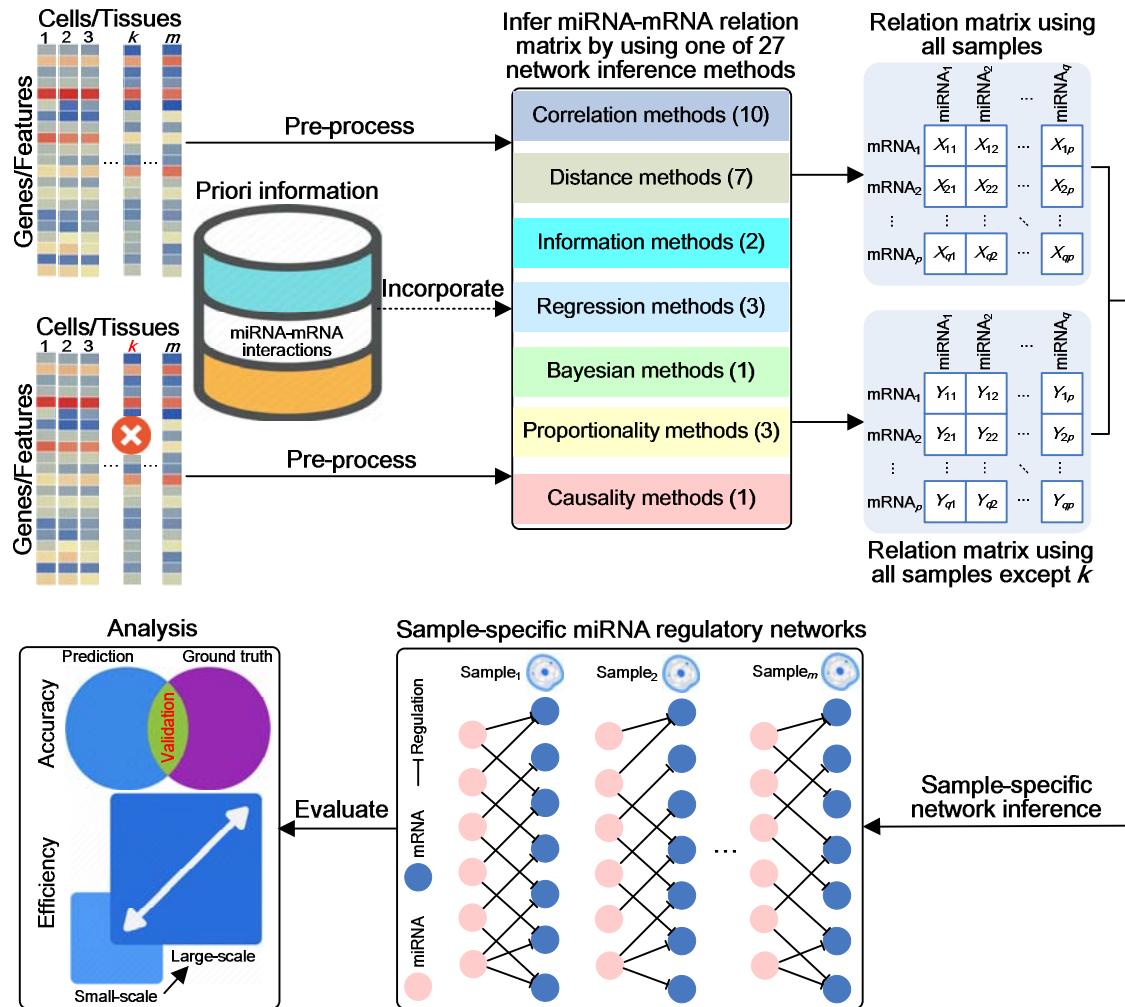
## Results

### **Sample-specific miRNA regulation (Scan)**

To investigate miRNA regulation at the resolution of single samples, we develop an approach called Scan to infer sample-specific miRNA regulation from bulk and single-cell RNA-sequencing data (**Fig. 1**). The bulk RNA-sequencing data of 690 breast cancer (BRCA) tissues is obtained from The Cancer Genome Atlas (TCGA)<sup>26</sup> project, and the single-cell RNA-sequencing data of 19 half K562 cells (the first human chronic myeloid leukemia cell line) is from the Gene Expression Omnibus (GEO)<sup>27</sup> database (Methods). The optional priori information of miRNA-mRNA interactions is from two well-known databases: TargetScan v8.0<sup>28</sup> and ENCORI<sup>29</sup> (Methods).

Given bulk or single-cell RNA-sequencing data with or without priori information of miRNA-mRNA interactions, Scan apply one of 27 network inference methods (Methods) spanning seven types (including Correlation, Distance, Information, Regression, Bayesian, Proportionality and Causality) to infer miRNA-mRNA relation matrix. For each sample (cell or tissue), Scan needs to infer two miRNA-mRNA relation matrices (one for all samples and

the other for all samples except the sample of interest). Based on the two identified miRNA-mRNA relation matrices, Scan further adapts one strategy (statistical perturbation or linear interpolation) to infer miRNA regulatory network specific to the sample of interest (Methods).



**Fig. 1 Schematic illustration of Scan.** Given bulk or single-cell RNA-sequencing data with or without priori information of miRNA-mRNA interactions, Scan applies one of 27 network inference methods covering seven types (Correlation, Distance, Information, Regression, Bayesian, Proportionality and Causality) to construct miRNA-mRNA relation matrix. By using one of 27 network inference methods, Scan constructs two miRNA-mRNA relation matrices (one for all samples and the other for all samples except the  $k$ -th sample). Then for sample (cell or tissue)  $k$ , Scan conducts sample-specific network inference from the two constructed miRNA-mRNA relation matrices to infer a miRNA regulatory network specific to the sample  $k$ . In total, Scan can identify  $m$  sample-specific miRNA regulatory networks across  $m$  samples (one network for one sample).

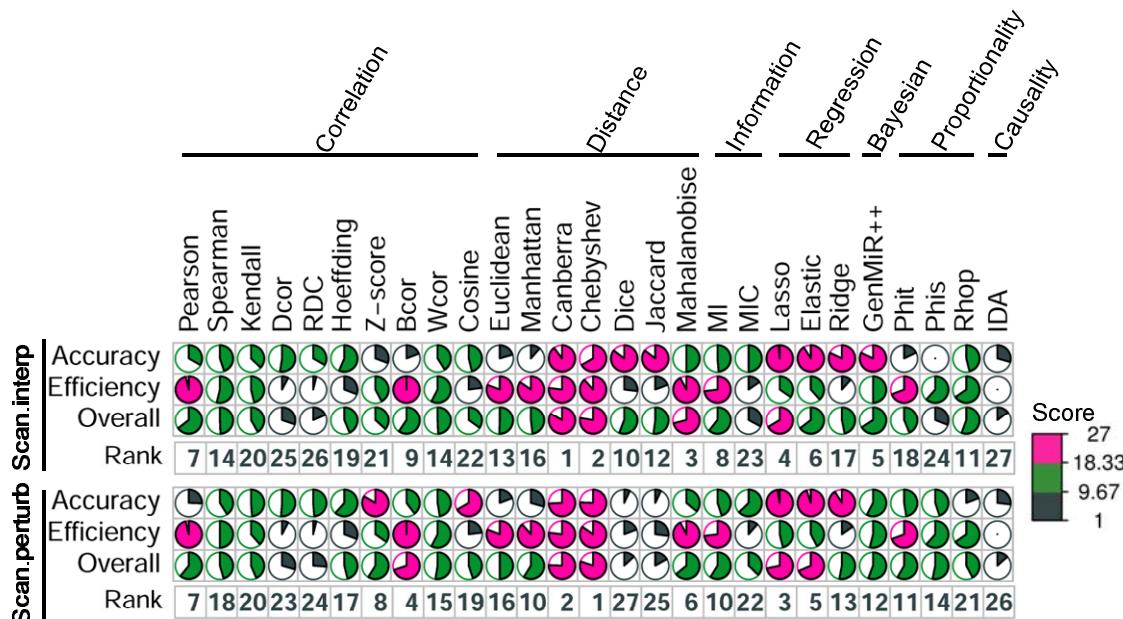
## Cell-specific miRNA regulation in K562

Using the single-cell miRNA-mRNA co-sequencing data in K562, we apply Scan to infer cell-specific miRNA regulation across 19 half K562 cells. Scan firstly uses 27 network inference methods to infer miRNA-mRNA relation matrices. For each cell, each network inference method infers two miRNA-mRNA relation matrices (one for all cells and the other for all cells except the cell of interest). Based on the identified miRNA-mRNA relation matrices, Scan further applies two strategies Scan.interp (using linear interpolation strategy) and Scan.perturb (using statistical perturbation strategy) to construct cell-specific miRNA regulatory networks across 19 half K562 cells. To systematically compare the performance of different combinations of network inference methods and strategies used, we have analyzed a total of 1026 networks: 27 (network inference methods) times 2 (strategies) times 19 (cells). We have found that using different network inference methods results in different number of cell-specific miRNA-mRNA interactions, and Dcor<sup>30</sup> has the largest number of miRNA-mRNA interactions predicted (**Supplementary Fig. 1a and 2a in Supplementary File 1**). Furthermore, when priori information of miRNA targets is used, more cell-specific miRNA regulatory networks are found to have their node degree obey a power law distribution and thus are regarded as scale-free networks (**Supplementary Fig. 1b and 2b in Supplementary File 1**). Interestingly, with or without using priori information of miRNA targets, the node degree of the identified cell-specific miRNA regulatory networks by Dcor does not follow a

power law distribution. In terms of the predicted miRNA-mRNA interactions, the similarity of the majority of cell pairs is less than 0.50 (**Supplementary Fig. 1c** and **2c** in **Supplementary File 1**), indicating the heterogeneity of miRNA regulation for each K562 cell.

When Scan.interp is used for inferring cell-specific miRNA regulation, using Lasso<sup>31</sup> for strength calculation gives the best final performance in terms of accuracy and using Bcor<sup>32</sup> leads to the highest efficiency (**Fig. 2**). Overall, using Canberra<sup>33</sup> has the largest overall rank score when using Scan.interp for inferring cell-specific miRNA regulation. When Scan.perturb is used for inferring cell-specific miRNA regulation, using Lasso also has the best final performance in terms of accuracy whereas using Bcor also leads to the highest efficiency (**Fig. 2**). Overall, using Chebyshev<sup>34</sup> shows the best overall performance when using Scan.perturb for inferring cell-specific miRNA regulation. This result indicates that in terms of accuracy and efficiency, Lasso and Bcor are consistent in the K562 dataset regardless of which strategy is used for inferring cell-specific miRNA regulation. Although the best network inference method (in terms of overall rank score) is different between Scan.interp and Scan.perturb, the network inference methods which lead to the best final result are all distance based network inference methods. Furthermore, in terms of efficiency, the rank scores obtained by using 27 network inference methods between Scan.interp and Scan.perturb are consistent (*p*-value less than 2.22E-16 with Cohen's kappa). However, in

terms of accuracy and overall, the rank scores obtained by using 27 network inference methods between Scan.interp and Scan.perturb are not consistent ( $p$ -values are 0.53 and 0.69 with Cohen's kappa, respectively). This result suggests that the performance of the network inference methods with the K562 dataset is closely related to the selected strategy for sample-specific miRNA regulatory networks.



**Fig. 2 Performance of each network inference method in K562 using Scan.interp and Scan.perturb.** Scores for each network inference method are calculated to evaluate the performance of predicting cell-specific miRNA regulation in terms of accuracy, efficiency and overall. Higher scores and smaller ranks denote better performance. The rank is visualized as a clock where clockwise direction indicates descending order.

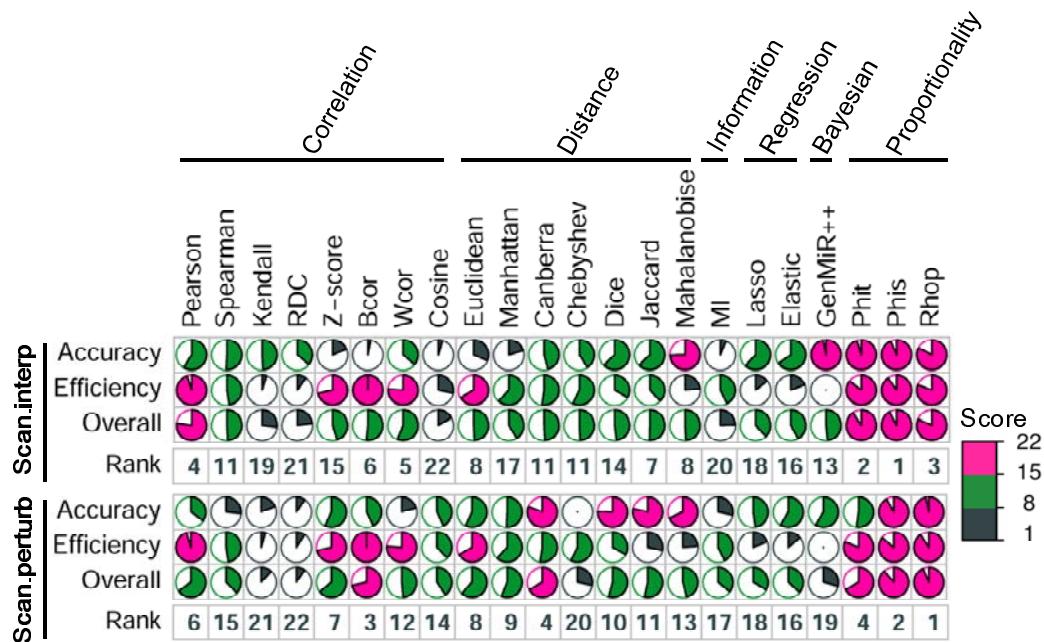
### Tissue-specific miRNA regulation in BRCA

Using the bulk miRNA and mRNA expression data in BRCA, we apply Scan to infer tissue-specific miRNA regulation across 690 BRCA tissues. Firstly, Scan uses 22 network inference methods (five methods are discarded due to long runtime or large memory usage) to infer miRNA-mRNA relation matrices. For each tissue, each network inference method infers two

miRNA-mRNA relation matrices (one for all tissues and the other for all tissues except the tissue of interest). For Scan.interp and Scan.perturb, Scan generates tissue-specific miRNA regulatory networks across 690 BRCA tissues based on the identified miRNA-mRNA relation matrices. To systematically evaluate the performance of different combinations of network inference methods and strategies used, we have analyzed a total of 30,360 networks: 22 (network inference methods) times 2 (strategies) times 690 (tissues). We have discovered that using different network inference methods leads to different number of tissue-specific miRNA-mRNA interactions (**Supplementary Fig. 3a** and **4a** in **Supplementary File 1**). Moreover, when priori information of miRNA targets is used, more tissue-specific miRNA regulatory networks are found to have their node degree follow a power law distribution and thus are regarded as scale-free networks (**Supplementary Fig. 3b** and **4b** in **Supplementary File 1**). In terms of the predicted miRNA-mRNA interactions, the similarity of most of tissue pairs is also less than 0.50 (**Supplementary Fig. 3c** and **4c** in **Supplementary File 1**), revealing the heterogeneity of miRNA regulation for each BRCA tissue.

When Scan.interp is used for inferring tissue-specific miRNA regulation, using GenMiR++<sup>35</sup> for strength calculation gives the best final results in terms of accuracy and whereas using Bcor<sup>32</sup> has the highest efficiency (**Fig. 3**). Overall, using Phis<sup>36</sup> has the largest overall rank score when using Scan.interp for inferring tissue-specific miRNA regulation. When Scan.perturb is used for inferring tissue-specific miRNA regulation, using Rhop<sup>36</sup>

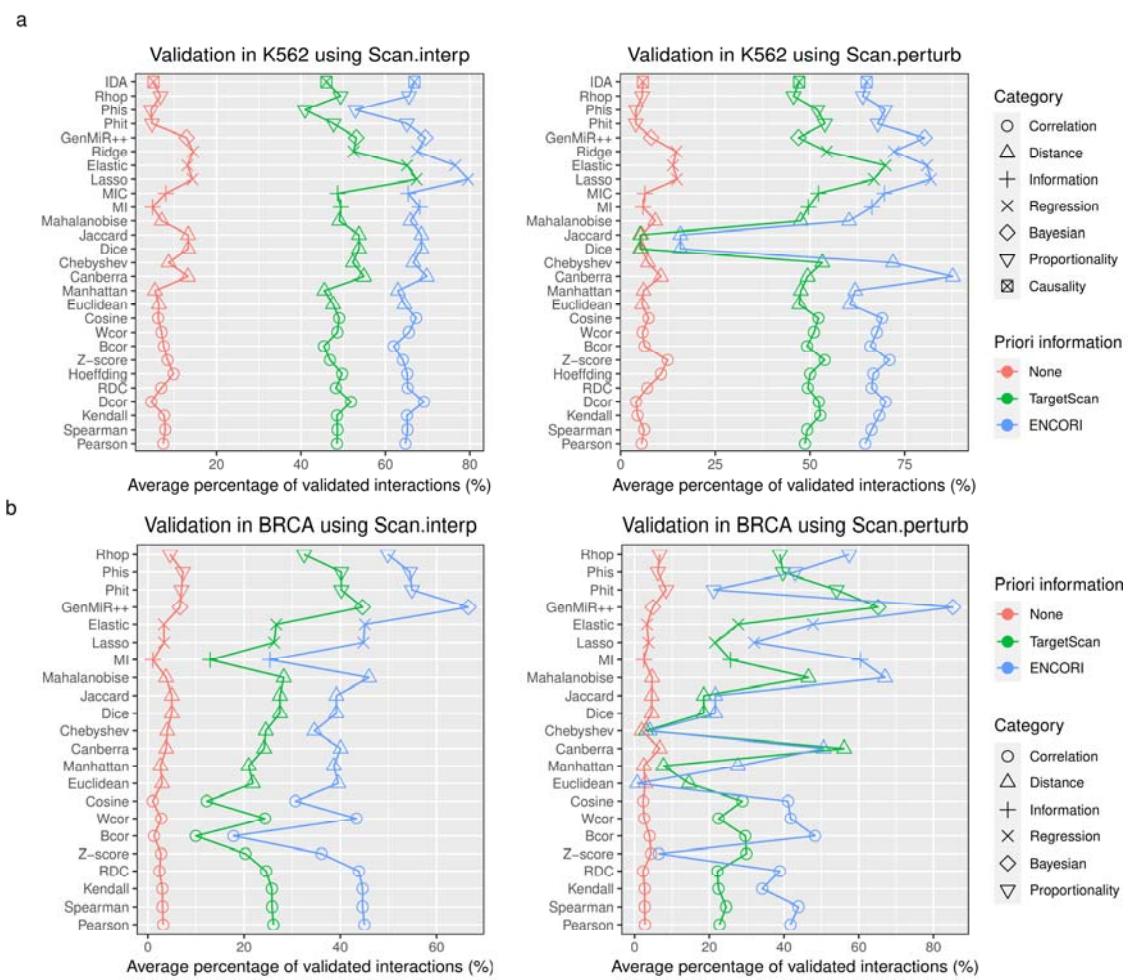
gives the best final results in terms of accuracy whereas using Bcor also has the highest efficiency (**Fig. 3**). Overall, using Rhop<sup>36</sup> displays the best performance when using Scan.perturb for inferring tissue-specific miRNA regulation. This result suggests that in terms of efficiency, Bcor is consistent in the BRCA dataset regardless of which strategy is used for inferring tissue-specific miRNA regulation. Although the best network inference method (in terms of overall rank score) is different between Scan.interp and Scan.perturb, the network inference methods which lead to the best final result are all proportionality based network inference methods. Moreover, in terms of efficiency, the rank scores obtained by using 22 network inference methods between Scan.interp and Scan.perturb are consistent (*p*-value less than 2.22E-16 with Cohen's kappa). However, in terms of accuracy and overall, the rank scores obtained by using 22 network inference methods between Scan.interp and Scan.perturb are not consistent (*p*-values are 0.58 and 0.75 with Cohen's kappa, respectively). This result implies that the performance of the network inference methods with the BRCA dataset is also closely associated with the selected strategy.



**Fig. 3 Performance of each network inference method in BRCA using Scan.interp and Scan.perturb.** Scores for each network inference method are calculated to evaluate the performance of predicting tissue-specific miRNA regulation in terms of accuracy, efficiency and overall. Higher scores and smaller ranks represent better performance. The rank is visualized as a clock where clockwise direction indicates descending order.

### Priori information can improve the accuracy of miRNA target prediction

To understand whether priori information of miRNA targets can improve the accuracy of miRNA target prediction, we compare the average percentage of validated miRNA targets predicted by Scan with and without using priori information of miRNA targets. We find that incorporating priori information of miRNA targets generally results in a larger average percentage of validated miRNA targets in both K562 and BRCA datasets (Fig. 4). Moreover, the higher the confidence of the priori information is, generally the larger the average percentage of validated miRNA targets in both K562 and BRCA datasets is (Fig. 4). This finding demonstrates that the priori information of miRNA targets can improve the accuracy of miRNA target prediction.



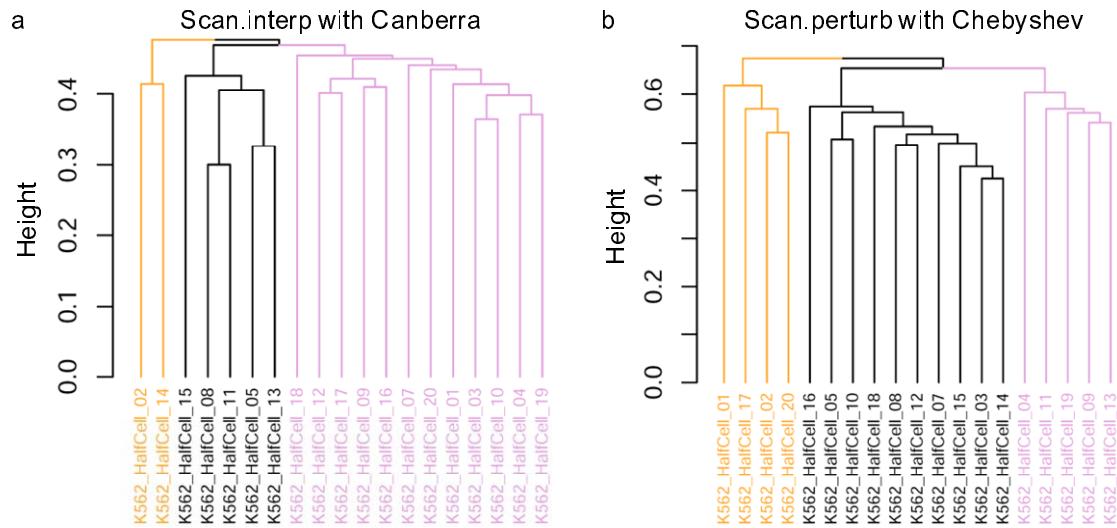
**Fig. 4 Validation of sample-specific miRNA regulations with and without using priori information.** a Validation in the K562 dataset using Scan.interp and Scan.perturb. b Validation in the BRCA dataset using Scan.interp and Scan.perturb.

### Clusters of K562 cells and BRCA tissues

With the K562 dataset, when Scan.interp and Scan.perturb are used together with Canberra and Chebyshev respectively, the best result in terms of overall rank score is achieved. For the BRCA dataset, when Scan.interp and Scan.perturb are used together with Phis and Rhop respectively, the best result in terms of overall rank score is obtained. As a result, we select the optimal combinations, Scan.interp\_Canberra (Scan.interp with Canberra) and Scan.perturb\_Chebyshev (Scan.perturb with Chebyshev) for the K562 dataset to infer cell-

specific miRNA regulatory networks, and Scan.interp\_Phis (Scan.interp with Phis) and Scan.perturb\_Rhop (Scan.perturb with Rhop) for the BRCA dataset to identify tissue-specific miRNA regulatory networks. We use the identified cell-specific (or tissue-specific) miRNA regulatory networks without using priori information to construct cell-cell (or tissue-tissue) similarity matrices. Based on the constructed cell-cell and tissue-tissue similarity matrices, we further conduct hierarchical clustering analysis to cluster K562 cells and BRCA tissues, respectively (Methods).

With the K562 dataset, we have three distinct clusters of K562 cells (**Fig. 5a and 5b**). With the BRCA dataset, we obtain five clear clusters of BRCA tissues (**Supplementary Fig. 5a and 5b in Supplementary File 1, Supplementary Data 1**). It is noted that the different clustering results of K562 cells or BRCA tissues are explained by using different network inference methods and different strategies. Since gene regulatory networks (e.g. miRNA regulatory networks) are regarded as stable forms to characterize a cell or a tissue<sup>37,38</sup>, our clustering results based on the identified cell-specific or tissue-specific miRNA regulatory networks may help to identify novel cell or tissue subtypes.



**Fig. 5 Clusters of K562 cells.** **a** Clusters of K562 cells, and Scan.interp with Canberra is used to identify cell-specific miRNA regulatory networks. **b** Clusters of K562 cells, and Scan.perturb with Chebyshev is used to infer cell-specific miRNA regulatory networks. Each color denotes a cluster.

### Correlation networks of K562 cells and BRCA tissues

Similar to the clustering of K562 cells and BRCA tissues, we also select the four optimal combinations (Scan.interp\_Canberra, Scan.perturb\_Chebyshev, Scan.interp\_Phis, Scan.perturb\_Rhop) without using priori information to infer cell-specific and tissue-specific miRNA regulatory networks respectively for the two datasets. We use the inferred cell-specific (or tissue-specific) miRNA regulatory networks to generate cell-cell (or tissue-tissue) similarity matrices. Based on the generated cell-cell and tissue-tissue similarity matrices, we further construct correlation networks of K562 cells and BRCA tissues respectively (Methods).

We have constructed two correlation networks for the K562 cells and two correlation networks for the BRCA tissues (**Supplementary Data 2**). Based on the identified cell-specific miRNA regulatory networks by Scan.interp\_Canberra and Scan.perturb\_Chebyshev,

the numbers of cell-cell correlation pairs are 171 and 22 respectively for the K562 dataset, Based on the identified tissue-specific miRNA regulatory networks by Scan.interp\_PhIs and Scan.perturb\_Rhop, the numbers of tissue-tissue correlation pairs are 51,769 and 8805 respectively for BRCA dataset. Network topological analysis reveals that the node degree of the identified cell-cell correlation network (10 nodes and 22 edges) by Scan.perturb\_Chebyshev obeys a power law distribution (*p*-value more than 0.99 with Kolmogorov-Smirnov test), and the node degrees of two identified tissue-tissue correlation networks by Scan.interp\_PhIs and Scan.perturb\_Rhop also follow power law distributions (*p*-values more than 0.85 with Kolmogorov-Smirnov test). Additionally, the identified cell-cell correlation network (19 nodes and 171 edges) by Scan.interp\_Canberra is a fully connected network. This result indicates that the identified correlation networks of cells or tissues tend to be scale-free or fully connected networks.

### **Dynamic and conservative miRNA regulation across K562 cells and BRCA tissues**

We also choose the four optimal combinations (Scan.interp\_Canberra, Scan.perturb\_Chebyshev, Scan.interp\_PhIs, Scan.perturb\_Rhop) without using priori information to infer miRNA regulatory networks specific to K562 cells and BRCA tissues. Based on the identified cell-specific or tissue-specific miRNA regulatory networks, we further investigate the dynamic and conservative miRNA regulation across K562 cells and BRCA tissues (Methods).

With both K562 and BRCA dataset, the numbers of dynamic and conservative miRNA-mRNA interactions identified tend to be varying (**Supplementary Data 3**). With the K562 dataset, the number of dynamic miRNA-mRNA interactions identified by Scan.interp\_Canberra (13786) is smaller than that identified by Scan.perturb\_Chebyshev (77627), but the number of conservative miRNA-mRNA interactions identified by Scan.interp\_Canberra (6490) is larger than that identified by Scan.perturb\_Chebyshev (3213). With the BRCA dataset, the numbers of dynamic and conservative miRNA-mRNA interactions identified by Scan.interp\_Phis (82850 and 29053, respectively) are larger than those identified by Scan.perturb\_Rhop (5549 and 36, respectively). Moreover, for Scan.interp\_Canberra, Scan.perturb\_Chebyshev, Scan.interp\_Phis and Scan.perturb\_Rhop, the number of miRNAs involved in the dynamic miRNA regulation (45, 212, 163 and 98, respectively) is larger than or equal to that of the conservative miRNA regulation (8, 212, 37 and 25, respectively).

With the K562 dataset, the node degrees of the identified dynamic and conservative miRNA regulatory networks by Scan.interp\_Canberra do not follow power law distributions (*p*-values are 2.41E-05 and 3.77E-04 with Kolmogorov-Smirnov test, respectively), but the node degrees of the identified dynamic and conservative miRNA regulatory networks by Scan.perturb\_Chebyshev obey power law distributions (*p*-values are 1.00 and 0.99 with Kolmogorov-Smirnov test, respectively). With the BRCA dataset, the node degrees of the

identified dynamic miRNA regulatory networks by Scan.interp\_PhIs and the identified conservative miRNA regulatory networks by Scan.perturb\_Rhop follow power law distributions (*p*-values are 0.21 and 1.00 with Kolmogorov-Smirnov test, respectively). However, with the BRCA dataset, the node degrees of the identified conservative miRNA regulatory networks by Scan.interp\_PhIs and the identified dynamic miRNA regulatory networks by Scan.perturb\_Rhop do not obey power law distributions (*p*-values are 2.77E-33 and 9.47E-05 with Kolmogorov-Smirnov test, respectively).

In addition, the numbers of dynamic and conservative hub miRNAs by Scan.interp and Scan.perturb are also different (**Supplementary Data 4**). With the K562 dataset, the number of dynamic and conservative hub miRNAs identified by Scan.interp\_Canberra (16 and 1, respectively) is larger than those identified by Scan.perturb\_Chebyshev (0 and 0, respectively). With the BRCA dataset, the numbers of dynamic and conservative hub miRNAs identified by Scan.interp\_PhIs (44 and 15, respectively) are larger than those identified by Scan.perturb\_Rhop (39 and 0, respectively). Furthermore, with the K562 dataset, 4 out of 16 dynamic hub miRNAs and 1 out of 1 conservative hub miRNA identified by Scan.interp\_Canberra are confirmed to be CML-related miRNAs. With the BRCA dataset, 38 out of 44 dynamic hub miRNAs and 15 out of 15 conservative hub miRNAs identified by Scan.interp\_PhIs are confirmed to be BRCA-related miRNAs, and 39 out of 39 dynamic hub miRNAs identified by Scan.perturb\_Rhop are validated to be BRCA-related miRNAs.

Altogether, the above results suggest that the findings of the dynamic and conservative miRNA regulation across K562 cells and BRCA tissues depend on the selected network inference methods and strategies.

### Performance comparison

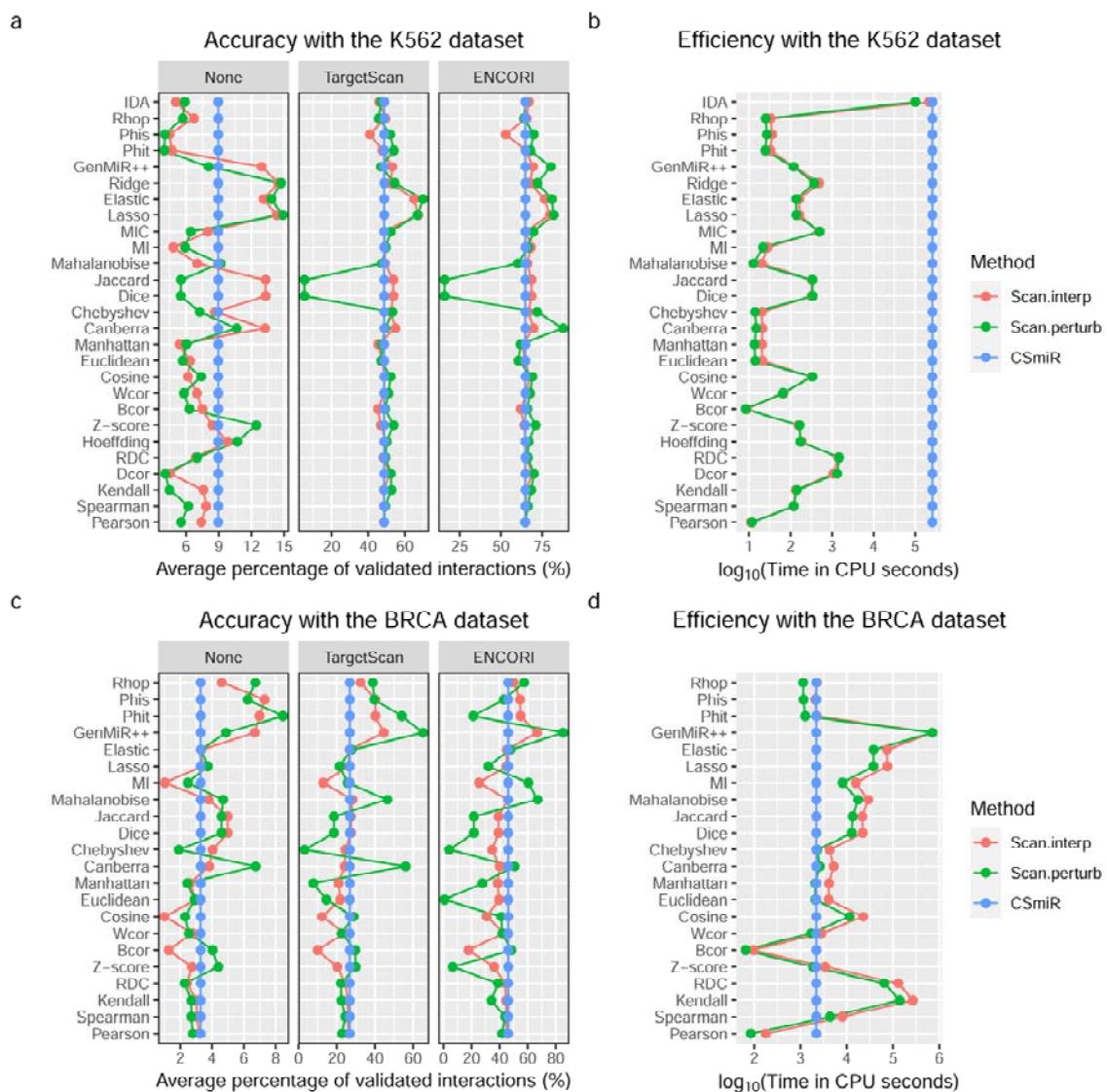
To evaluate the effectiveness of Scan, we compare the performance of Scan (Scan.interp and Scan.perturb) with CSmiR<sup>16</sup> on the K562 and BRCA datasets. Here, CSmiR is the first method to explore miRNA regulation at a single-cell resolution level, so we use it as the baseline for comparison. With the K562 dataset, without using priori information of miRNA targets, Scan.interp with 8 network inference methods and Scan.perturb with 7 network inference methods perform better than CSmiR in terms of accuracy (**Fig. 6a**). After integrating priori information of miRNA targets from TargetScan, Scan.interp with 14 network inference methods and Scan.perturb with 18 network inference methods perform better than CSmiR in terms of accuracy, and after integrating priori information of miRNA targets from ENCORL, Scan.interp with 21 network inference methods and Scan.perturb with 20 network inference methods perform better than CSmiR in terms of accuracy (**Fig. 6a**). This result shows that in terms of accuracy, Scan mostly performs better than or comparable to CSmiR with the K562 dataset when using priori information of miRNA targets. Furthermore, in terms of efficiency, both Scan.interp and Scan.perturb with all of the network inference methods perform better than CSmiR (**Fig. 6b**).

With the BRCA dataset, without using priori information of miRNA targets, Scan.interp with 11 network inference methods and Scan.perturb with 12 network inference methods perform better than CSmiR in terms of accuracy (**Fig. 6c**). After integrating priori information of miRNA targets from TargetScan, Scan.interp with 7 network inference methods and Scan.perturb with 10 network inference methods perform better than CSmiR in terms of accuracy, and after integrating priori information of miRNA targets from ENCORL, Scan.interp with 4 network inference methods and Scan.perturb with 7 network inference methods perform better than CSmiR in terms of accuracy (**Fig. 6c**). This result shows that in terms of accuracy, Scan mostly performs better than or comparable to CSmiR with the BRCA dataset when without using priori information of miRNA targets. In addition, in terms of efficiency, Scan.interp with 3 network inference methods and Scan.perturb with 9 network inference methods perform better than CSmiR (**Fig. 6d**).

The above performance comparison suggests that in the small-scale K562 dataset, Scan with priori information generally performs better than or comparable to CSmiR in terms of accuracy and efficiency. In the large-scale BRCA dataset, Scan without using priori information generally performs better than or comparable to CSmiR in terms of accuracy, but in terms of efficiency, only a few instances of Scan (Scan with a few network inference methods) perform better than CSmiR. The inconsistent result of performance comparison

with the two datasets implies that the performance of Scan is more likely to be data-specific,

and it is necessary to select optimal network inference methods for new data.



**Fig. 6 Performance comparison in inferring sample-specific miRNA regulation between Scan.interp, Scan.perturb and CSmiR. a** Accuracy comparison with the K562 dataset. **b** Efficiency comparison with the K562 dataset. **c** Accuracy comparison with the BRCA dataset. **d** Efficiency comparison with the BRCA dataset.

## Discussion

It is well-established that miRNAs are important regulators of gene expression, and their

dysregulations can lead to the occurrence and development of complex human diseases,

including cancers. Yet, the research on miRNA regulation at the resolution of single samples (cells or tissues) is still limited. Here, we present the Scan framework, and further show the effectiveness of it in inferring sample-specific miRNA regulation. By applying Scan into bulk and single-cell RNA-sequencing data, we have discovered that adding priori information of miRNA targets can improve the accuracy of miRNA target prediction. By instantiating the Scan framework with 27 network inference methods, we have found that the performance of Scan instantiated with different network inference methods exhibits to be data-specific. In addition, the identified sample-specific miRNA regulatory network by Scan can be used for downstream analysis, e.g. clustering samples and constructing sample correlation network. The freely available framework Scan provides a useful method for exploring miRNA regulation at the resolution of single samples.

For constructing miRNA-mRNA relation matrix, Scan includes 27 network inference methods spanning seven types (Correlation, Distance, Information, Regression, Bayesian, Proportionality and Causality). It is noted that other types of computational methods (e.g. deep learning<sup>39</sup>, probabilistic modeling<sup>40</sup>) can be also plugged into Scan for generating miRNA-mRNA relation matrix. To improve the generalization ability of Scan, it is our plan to add more types of network inference methods in future.

When applying Scan to a new data, user can choose one or multiple network inference methods according to the rank score in terms of accuracy or efficiency or overall. For

example, for the K562 dataset, we can select network inference method Canberra to be used with Scan.interp and Chebyshev with Scan.perturb in terms of overall rank score. But for the BRCA dataset, we can select network inference method Phis to be used with Scan.interp and Rhop with Scan.perturb in terms of overall rank score.

We provide two strategies, Scan.interp (linear interpolation strategy) and Scan.perturb (statistical perturbation strategy) to infer miRNA regulation specific to single cells or tissues. Generally, using the same network inference method, Scan.interp can generate larger cell-specific or tissue-specific miRNA regulatory networks than those by Scan.perturb. This can be explained by the fact that Scan.interp infers cell-specific miRNA regulatory networks with differential values (e.g. differential correlation, distance or regression) between all samples and all samples except sample  $k$ , but Scan.perturb identifies differential miRNA regulatory networks between the networks of using all samples and the networks of using all samples except sample  $k$  (see details in Methods). Moreover, the runtime of Scan.interp and Scan.perturb with the K562 and BRCA datasets is similar. For new data, we suggest use both Scan.interp and Scan.perturb for comprehensively exploring miRNA regulation specific to samples.

To infer sample-specific miRNA regulation, we have applied Scan into bulk and single-cell RNA-sequencing data across tumor cells or tissues. Certainly, Scan is also applicable in bulk and single-cell RNA-sequencing data across healthy cells or tissues. In future, it will be

a meaningful direction to reveal dynamic and conservative miRNA regulation between matched tumor and healthy cells or tissues. Moreover, with the advancement of spatial RNA-sequencing technology<sup>41</sup>, Scan is also a potential method to explore the heterogeneity of miRNA regulation between different spatial positions from spatial RNA-sequencing data.

When Scan is applied to large-scale transcriptomics data, the process of inferring miRNA-mRNA relation matrix using network inference methods will be computationally intensive. This is a common issue of existing computational methods, including Scan. To alleviate the issue, users can allocate more CPU cores to identify miRNA-mRNA relation matrix in parallel. Furthermore, users can also use priori information of samples to divide samples into several subtypes, and then focus on investigating sample-specific miRNA regulation across cells or tissues of each subtype. Finally, selecting a fast network inference method (e.g. Pearson and Bcor) is also a feasible way to quickly infer sample-specific miRNA regulation from large-scale transcriptomics data.

In addition to studying miRNA regulation at single-sample resolved level, Scan also can be used for other types of gene regulation specific to individual samples, e.g. transcriptional regulation, long non-coding RNA regulation, circular RNA regulation and PIWI-interacting RNA regulation.

Since miRNAs play important roles in tumor microenvironments, it is important for us to understand the regulation of miRNAs in tumor cells or tissues. In-depth investigation of

miRNA regulation specific to individual tumor cells or tissues will help to figure out the heterogeneity of tumor microenvironments and discover novel subtypes of tumor cells or tissues.

## Methods

### Bulk and single-cell RNA-sequencing data

From TCGA<sup>26</sup>, we have obtained the expression data of 894 miRNAs and 19,068 mRNAs in 690 matching breast cancer tissues. Since epithelial-mesenchymal transition (EMT) is closely related to the development, progression and metastasis of breast cancer<sup>42-45</sup>, we obtain a list of 315 EMT signatures<sup>46</sup> (**Supplementary Data 5**) and further divide the matched 690 breast cancer tissues into four EMT types (epithelial, intermediate epithelial, intermediate mesenchymal and mesenchymal) by using GSVA R package<sup>47</sup>. As a result, the numbers of breast cancer tissues belonging to epithelial, intermediate epithelial, intermediate mesenchymal and mesenchymal types are 491, 107, 46 and 46, respectively (**Supplementary Data 6**). By using the *limma-trend* approach in limma R package<sup>48</sup>, we have identified 163 miRNAs and 5801 mRNAs that are differentially expressed between epithelial and mesenchymal type (adjusted *p*-value < 0.01, fold change > 1.5) (**Supplementary Data 7**). Here, the *p*-values are adjusted by the Benjamini–Hochberg (BH) method<sup>49</sup>. Therefore, the

input bulk RNA-sequencing data used in our case study includes the expression data of 163 miRNAs and 5801 mRNAs in 690 breast cancer tissues.

The K562 single-cell RNA-sequencing data (the accession number is GSE114071 in GEO database<sup>27</sup>) includes the expression data of 2822 miRNAs and 21,704 mRNAs in 19 half K562 cells (the first human chronic myeloid leukemia cell line). In the single-cell RNA-sequencing data, we calculate the average expression values of the duplicate miRNAs or mRNAs as their ultimate expression values. As a feature selection step, we discard the miRNAs or mRNAs with constant expression values, and we are interested in the high expression miRNAs or mRNAs with mean values greater than the median of mean values of all non-constant expression miRNAs or mRNAs. Moreover, the reserved miRNA data and mRNA expression data are then log-transformed. As a result, the input single-cell RNA-sequencing data used in our case study includes the expression data of 212 miRNAs and 7680 mRNAs in 19 half K562 cells.

### **Priori information of miRNA targets**

To improve the prediction of sample-specific miRNA regulation, the miRNA target information in TargetScan v8.0<sup>28</sup> and ENCORI<sup>29</sup> (the pilot version is starBase) is used as priori information for Scan. From TargetScan, a list of 235,109 predicted miRNA-mRNA interactions has been obtained. To identify miRNA-mRNA interactions, ENCORI provides seven prediction tools (PITA<sup>50</sup>, RNA22<sup>51</sup>, miRmap<sup>52</sup>, DIANA-microT<sup>53</sup>, miRanda<sup>54</sup>, PicTar<sup>55</sup>

and TargetScan<sup>28</sup>) for users. To obtain high-confidence miRNA-mRNA interactions, we only retain the miRNA-mRNA interactions predicted by at least five prediction algorithms. In total, a list of 55,343 high-confidence miRNA-mRNA interactions are obtained from ENCORI.

### Network inference methods

We select 27 network inference methods (**Supplementary Data 8**) for constructing miRNA-mRNA relation matrix: Pearson<sup>56</sup>, Spearman<sup>57</sup>, Kendall<sup>58</sup>, Distance correlation (Dcor)<sup>30</sup>, Random Dependence Coefficient (RDC)<sup>59</sup>, Hoeffding's D statistics (Hoeffding)<sup>60</sup>, Z-score<sup>61</sup>, Biweight midcorrelation (Bcor)<sup>32</sup>, Weighted rank correlation (Wcor)<sup>62</sup>, Cosine<sup>63</sup>, Euclidean<sup>64</sup>, Manhattan<sup>65</sup>, Canberra<sup>33</sup>, Chebyshev<sup>34</sup>, Dice<sup>66</sup>, Jaccard<sup>67</sup>, Mahalanobise<sup>68</sup>, Mutual Information (MI)<sup>69</sup>, Maximal Information Coefficient (MIC)<sup>70</sup>, Lasso<sup>31</sup>, Elastic<sup>31</sup>, Ridge<sup>31</sup>, GenMiR++<sup>35</sup>,  $\phi$ (Phit)<sup>36</sup>,  $\phi_s$ (Phis)<sup>36</sup>,  $\rho_p$ (Rhop)<sup>36</sup>, and Intervention calculus when the Directed acyclic graph is Absent (IDA)<sup>71</sup>. These methods can be divided into seven types: Correlation, Distance, Information, Regression, Bayesian, Proportionality and Causality. Ten methods (Pearson, Spearman, Kendall, Dcor, RDC, Hoeffding, Z-score, Bcor, Wcor and Cosine) belong to the Correlation type, seven methods (Euclidean, Manhattan, Canberra, Chebyshev, Dice, Jaccard and Mahalanobise) belong to the Distance type, two methods (MI and MIC) are of the Information type, three methods (Lasso, Elastic and Ridge) belong to the Regression type, three methods (Phit, Phis and Rhop) belong to the Proportionality type, and GenMiR++ and IDA are of the Bayesian and Causality types, respectively.

## Sample-specific network inference

Given bulk or single-cell RNA-sequencing data with  $m$  samples, with or without priori information of miRNA-mRNA interactions, Scan constructs two miRNA-mRNA relation matrices (one for all samples and the other for all samples except the  $k$ -th sample of interest,  $k \in [1, m]$ ) by using one of the network inference methods. The two constructed miRNA-mRNA relation matrices (of  $p$  miRNAs and  $q$  mRNAs) for all samples and all samples except the  $k$ -th sample (denoted as  $X^{(k)}$  and  $Y^{(k)}$  respectively) are as follows.

$$X^{(k)} = \begin{pmatrix} X_{11}^{(k)} & X_{12}^{(k)} & \cdots & X_{1p}^{(k)} \\ X_{21}^{(k)} & X_{22}^{(k)} & \cdots & X_{2p}^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ X_{q1}^{(k)} & X_{q2}^{(k)} & \cdots & X_{qp}^{(k)} \end{pmatrix} = \left( X_{ij}^{(k)} \right) \in \mathbb{Q}^{q \times p}$$

$$Y^{(k)} = \begin{pmatrix} Y_{11}^{(k)} & Y_{12}^{(k)} & \cdots & Y_{1p}^{(k)} \\ Y_{21}^{(k)} & Y_{22}^{(k)} & \cdots & Y_{2p}^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{q1}^{(k)} & Y_{q2}^{(k)} & \cdots & Y_{qp}^{(k)} \end{pmatrix} = \left( Y_{ij}^{(k)} \right) \in \mathbb{Q}^{q \times p}.$$

where  $X_{ij}^{(k)}$  and  $Y_{ij}^{(k)}$  represent the connection between mRNA  $i$  and miRNA  $j$ .

For the network inference methods which are of the Correlation, Information, Regression, Bayesian, Proportionality and Causality types, larger absolute values of  $X_{ij}^{(k)}$  and  $Y_{ij}^{(k)}$  indicate higher connections between mRNA  $i$  and miRNA  $j$ . In contrast, for Distance based network inference methods, larger absolute values of  $X_{ij}^{(k)}$  and  $Y_{ij}^{(k)}$  indicate weaker connections between mRNA  $i$  and miRNA  $j$ . To keep consistency between the constructed miRNA-mRNA relation matrices by different types of network inference methods, we

transform the two miRNA-mRNA relation matrices ( $X^{(k)}$  and  $Y^{(k)}$ ) generated by Distance based methods into  $X'^{(k)}$  and  $Y'^{(k)}$  as follows:

$$X'^{(k)} = \left( \frac{1}{X_{ij}^{(k)} + \text{eps}} \right) \in \mathbb{D}^{q \times p}$$

$$Y'^{(k)} = \left( \frac{1}{Y_{ij}^{(k)} + \text{eps}} \right) \in \mathbb{D}^{q \times p}$$

where  $\text{eps}$  refers to the precision of floating numbers (the value is 2.22E-16 in default).

After transformation, a larger absolute value in a relation matrix generated by any type of the network inference methods indicates a higher connection between a miRNA and a mRNA.

#### *Linear interpolation strategy*

Following the use of a network inference method of the Correlation, Information, Regression, Bayesian, Proportionality or Causality type to obtain the two constructed miRNA-mRNA relation matrices ( $X^{(k)}$  and  $Y^{(k)}$ ), Scan applies a linear interpolation strategy<sup>22</sup> to estimate the miRNA-mRNA relation matrix  $Z^{(k)}$  specific to sample  $k$  as follows:

$$Z^{(k)} = mX^{(k)} - (m-1)Y^{(k)} = \begin{pmatrix} Z_{11}^{(k)} & Z_{12}^{(k)} & \cdots & Z_{1p}^{(k)} \\ Z_{21}^{(k)} & Z_{22}^{(k)} & \cdots & Z_{2p}^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{q1}^{(k)} & Z_{q2}^{(k)} & \cdots & Z_{qp}^{(k)} \end{pmatrix} = (Z_{ij}^{(k)}) \in \mathbb{D}^{q \times p}$$

where  $m$  denotes the number of samples in bulk or single-cell RNA-sequencing data.

When a Distance based network inference method is used to obtain the two constructed miRNA-mRNA relation matrices ( $X^{(k)}$  and  $Y^{(k)}$ ), the miRNA-mRNA relation matrix  $Z^{(k)}$  specific to sample  $k$  is estimated as:

$$Z^{(k)} = mX^{(k)} - (m-1)Y^{(k)} = \begin{pmatrix} Z_{11}^{(k)} & Z_{12}^{(k)} & \cdots & Z_{1p}^{(k)} \\ Z_{21}^{(k)} & Z_{22}^{(k)} & \cdots & Z_{2p}^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{q1}^{(k)} & Z_{q2}^{(k)} & \cdots & Z_{qp}^{(k)} \end{pmatrix} = (Z_{ij}^{(k)}) \in \mathbb{R}^{q \times p}$$

$Z_{ij}^{(k)}$  is further normalized as:

$$Z_{ij}^{(k)} = \frac{Z_{ij}^{(k)} - \mu^{(k)}}{\sigma^{(k)}}$$

where  $\mu^{(k)}$  and  $\sigma^{(k)}$  denote the mean value and standard deviation of  $Z^{(k)}$ , and the normalized miRNA-mRNA relation matrix  $Z^{(k)}$  is:

$$Z^{(k)} = \begin{pmatrix} Z_{11}^{(k)} & Z_{12}^{(k)} & \cdots & Z_{1p}^{(k)} \\ Z_{21}^{(k)} & Z_{22}^{(k)} & \cdots & Z_{2p}^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{q1}^{(k)} & Z_{q2}^{(k)} & \cdots & Z_{qp}^{(k)} \end{pmatrix} = (Z_{ij}^{(k)}) \in \mathbb{R}^{q \times p}$$

Each value of  $Z_{ij}^{(k)}$  corresponds to a significance  $p$ -value. The  $p$ -value is calculated as follows:

$$p_{ij}^{(k)} = 1 - pnorm(|Z_{ij}^{(k)}|)$$

where  $|Z_{ij}^{(k)}|$  is the absolute value of  $Z_{ij}^{(k)}$ , and the  $pnorm$  function is used to calculate the probability a random value from the standard normal distribution being less than  $|Z_{ij}^{(k)}|$ .

Smaller value of  $p_{ij}^{(k)}$  indicates that miRNA  $j$  is more likely to interact with mRNA  $i$  in sample  $k$ . If the value of  $p_{ij}^{(k)}$  is less than a cutoff (e.g. 0.05), miRNA  $j$  is considered to interact with mRNA  $i$  in sample  $k$ .

#### *Statistical perturbation strategy*

For the Correlation, Distance, Information, Regression, Bayesian, Proportionality and Causality methods, each value of  $X^{(k)}$  ( $X'^{(k)}$ ) and  $Y^{(k)}$  ( $Y'^{(k)}$ ) corresponds to a significance  $p$ -value. The corresponding  $p$ -value matrices of  $X^{(k)}$  ( $X'^{(k)}$ ) and  $Y^{(k)}$  ( $Y'^{(k)}$ ) are denoted as  $p_x^{(k)}$  and  $p_y^{(k)}$ , respectively.

$$p_x^{(k)} = \begin{pmatrix} x_{11}^{(k)} & x_{12}^{(k)} & \cdots & x_{1p}^{(k)} \\ x_{21}^{(k)} & x_{22}^{(k)} & \cdots & x_{2p}^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{q1}^{(k)} & x_{q2}^{(k)} & \cdots & x_{qp}^{(k)} \end{pmatrix} = (x_{ij}^{(k)}) \in \square^{q \times p}$$

$$p_y^{(k)} = \begin{pmatrix} y_{11}^{(k)} & y_{12}^{(k)} & \cdots & y_{1p}^{(k)} \\ y_{21}^{(k)} & y_{22}^{(k)} & \cdots & y_{2p}^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ y_{q1}^{(k)} & y_{q2}^{(k)} & \cdots & y_{qp}^{(k)} \end{pmatrix} = (y_{ij}^{(k)}) \in \square^{q \times p}$$

Smaller value of  $x_{ij}^{(k)}$  and  $y_{ij}^{(k)}$  indicates that miRNA  $j$  is more likely to interact with mRNA  $i$  in all samples and all samples except  $k$ , respectively. Given a  $p$ -value cutoff (e.g. 0.05), we can obtain two zero-one matrices  $ZO_x^{(k)}$  and  $ZO_y^{(k)}$  in all samples and all samples except  $k$ . Scan further uses a statistical perturbation strategy<sup>21</sup> to calculate the miRNA-mRNA zero-one matrix  $ZO^{(k)}$  for sample  $k$  as follows:

$$ZO^{(k)} = xor(ZO_x^{(k)}, ZO_y^{(k)})$$

where  $xor$  is the XOR logical function. One value of  $ZO_{ij}^{(k)}$  indicates that miRNA  $j$  interacts with mRNA  $i$  in sample  $k$ , and zero value of  $ZO_{ij}^{(k)}$  represents that miRNA  $j$  doesn't interact with mRNA  $i$  in sample  $k$ .

In this work, the cutoff of significant  $p$ -value for both linear interpolation and statistical perturbation strategies is set to be 0.05 in default. By applying linear interpolation or statistical perturbation strategy, Scan can identify  $m$  sample-specific miRNA regulatory networks across  $m$  samples. Each sample-specific miRNA regulatory network is a directed graph where the direction of an edge is from a miRNA to a mRNA.

### Degree distribution analysis

The degree of a node in a sample-specific miRNA regulatory network is the number of connections with other nodes, and the degree distribution denotes the probability distribution of node degrees over the sample-specific miRNA regulatory network. If the node degree of a sample-specific miRNA regulatory network obeys a power law distribution, the network is considered as a scale-free network. In this work, we use the R package `igraph`<sup>72</sup> to calculate the degree distribution of the identified sample-specific miRNA regulatory networks. The Kolmogorov-Smirnov (KS) test<sup>73</sup> is used to determine whether the node degree of a sample-specific miRNA regulatory network obeys a power law distribution. If the  $p$ -value of the KS test is smaller than a cutoff (e.g. 0.05), the node degree of a sample-specific miRNA

regulatory network does not obey a power law distribution, suggesting that the sample-specific miRNA regulatory network is not a scale-free network, and vice versa.

### Clustering analysis

We use sample-specific miRNA regulatory networks to compute the similarities between samples. In the context of sample-specific miRNA regulatory networks, the similarity between samples  $a$  and  $b$  is calculated in the following.

$$sim_{ab} = \frac{|Net_a \cap Net_b|}{\min(|Net_a|, |Net_b|)}$$

where  $Net_a$  and  $Net_b$  represent the miRNA regulatory networks specific to samples  $a$  and  $b$ , respectively,  $|Net_a \cap Net_b|$  is the number of common miRNA-mRNA interactions between  $Net_a$  and  $Net_b$ , and  $\min(|Net_a|, |Net_b|)$  denotes the smaller number of miRNA-mRNA interactions between  $Net_a$  and  $Net_b$ .

For  $m$  samples, the sample-sample similarity matrix  $SM$  (a symmetric matrix) is:

$$SM = \begin{pmatrix} sim_{11} & sim_{11} & \cdots & sim_{1m} \\ sim_{21} & sim_{22} & \cdots & sim_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ sim_{m1} & sim_{m2} & \cdots & sim_{mm} \end{pmatrix} = (sim_{ab}) \in \mathbb{R}^{m \times m}$$

Based on the sample-sample similarity matrix  $SM$ , we further conduct clustering analysis of samples, e.g. hierarchical clustering analysis.

### Sample correlation network construction

In addition to the clustering analysis of samples, the sample-sample similarity matrix is also used for sample correlation network construction<sup>74</sup>. For each sample pair, a higher similarity value indicates that the pair of samples is more correlated with each other. We use an empirical similarity cutoff (e.g. 0.50) to infer whether two samples are correlated or not. In other words, if the similarity value between samples  $a$  and  $b$  is larger than the cutoff, samples  $a$  and  $b$  are correlated with each other. After assembling the correlated sample pairs, we can construct a sample correlation network.

### **Dynamic and conservative analysis**

The miRNA-mRNA interaction existing in only one cell or tissue are defined as a dynamic miRNA-mRNA interaction, whereas the miRNA-mRNA interaction existing in at least a half of cells or tissues are defined as a conservative miRNA-mRNA interaction. Given the identified sample-specific miRNA regulatory networks, we can obtain dynamic and conservative miRNA-mRNA interactions which form dynamic and conservative miRNA regulatory networks, respectively.

### **Hub miRNA identification**

Hub miRNAs are defined as highly connected miRNAs in a dynamic and conservative miRNA regulatory network. In this work, we use the cumulative probability of Poisson distribution to evaluate whether a miRNA is a hub:

$$p(d \geq z) = 1 - \sum_{i=0}^{z-1} \frac{\lambda^i e^{-\lambda}}{i!}$$

where  $\lambda = np$ ,  $p = \frac{t}{C}$ ,  $n$  is the number of genes (including miRNAs and mRNAs),  $t$  is the number of miRNA-mRNA interactions in a dynamic or conservative miRNA regulatory network, and  $C$  is the number of all possible miRNA-mRNA interaction pairs. Smaller  $p$ -value of a miRNA shows that the miRNA is more likely to be a hub. Here, the cutoff of the  $p$ -value is set to 0.05.

### Validation analysis

To understand whether the identified dynamic and conservative hub miRNAs are closely associated with CML and BRCA, we get a list of CML-related and BRCA-related miRNAs from RNADisease v4.0<sup>75</sup>. As a result, we have obtained 203 CML-related and 1493 BRCA-related miRNAs (**Supplementary Data 9**). If a dynamic or conservative hub miRNA is overlapped with the list of CML-related or BRCA-related miRNAs, the dynamic or conservative hub miRNA is regarded as a CML-related or BRCA-related miRNA.

### Comparison metrics

We use two metrics (accuracy and efficiency) to compare different instances of Scan using different network inference methods, and compare Scan with other methods for inferring sample-specific miRNA regulation. For accuracy, the ground truth of miRNA-mRNA interactions are acquired from miRTarBase v9.0<sup>76</sup> and TarBase v8.0<sup>77</sup> for validation. If a method has a larger percentage of validated miRNA-mRNA interactions, the method will have higher accuracy. For efficiency, we compare the runtime of different methods in the

same bulk or single-cell RNA-sequencing data. If a method takes less runtime in the same bulk or single-cell RNA-sequencing data, the method will have better efficiency.

In terms of accuracy and efficiency, we use an overall rank score<sup>78</sup> to evaluate the performance of each method. For the method  $i$ , the overall rank score  $ors_i$  is calculated as:

$$ors_i = \frac{1}{2}(rs_{i1} + rs_{i2})$$

where  $rs_{i1}$  and  $rs_{i2}$  denote the rank scores of method  $i$  in terms of accuracy and efficiency, respectively.

A method with higher accuracy or better efficiency will obtain a larger rank score. A method with a larger overall rank score is regarded as a better or practical method.

### **Method execution**

Each execution of Scan and CSmiR on bulk or single-cell RNA-sequencing data is performed in a separate task. Each task is allocated 32 CPU cores of Intel(R) Xeon(R) Platinum 8375C CPU at 2.90 GHz, and one R session is opened for each task. The network inference methods of Scan with runtime more than 10 days or with memory usage more than 256 GB are discarded.

### **Data availability**

All accession codes, unique identifiers, and web links for publicly available datasets are described in the paper. All data supporting the findings of the current study are listed in Supplementary Data and our GitHub website (<https://github.com/zhangjunpeng411/Scan/>).

## Code availability

Scan is released under the GPL-3.0 License, and is available at <https://github.com/zhangjunpeng411/Scan/>.

## References

1. Bartel, D. P. MicroRNAs: target recognition and regulatory functions. *Cell* **136**, 215–233 (2009).
2. Krol, J., Loedige, I. & Filipowicz, W. The widespread regulation of microRNA biogenesis, function and decay. *Nat Rev Genet* **11**, 597–610 (2010).
3. Peng, Y. & Croce, C. M. The role of microRNAs in human cancer. *Signal Transduct Target Ther* **1**, 15004 (2016).
4. Hamam, R. *et al.* Circulating microRNAs in breast cancer: novel diagnostic and prognostic biomarkers. *Cell Death Dis* **8**, e3045 (2017).
5. Yendamuri, S. & Calin, G. A. The role of microRNA in human leukemia: a review. *Leukemia* **23**, 1257–1263 (2009).
6. Coradduzza, D. *et al.* miRNAs as molecular biomarkers for prostate cancer. *J Mol Diagn* **24**, 1171–1180 (2022).
7. Salmena, L., Poliseno, L., Tay, Y., Kats, L. & Pandolfi, P. P. A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell* **146**, 353–358 (2011).
8. Tay, Y., Rinn, J. & Pandolfi, P. P. The multilayered complexity of ceRNA crosstalk and competition. *Nature* **505**, 344–352 (2014).

9. Karreth, F. A. & Pandolfi, P. P. ceRNA cross-talk in cancer: when ce-bling rivalries go awry. *Cancer Discov* **3**, 1113–1121 (2013).
10. Thomson, D. W. & Dinger, M. E. Endogenous microRNA sponges: evidence and controversy. *Nat Rev Genet* **17**, 272–283 (2016).
11. Wang, Y. *et al.* The emerging function and mechanism of ceRNAs in cancer. *Trends Genet* **32**, 211–224 (2016).
12. Zhang, Z., Han, J., Song, L. & Zhang, X. Inferring cell-specific gene regulatory networks from single cell gene expression data. *bioRxiv* 2022.03.03.482887 (2022) doi:10.1101/2022.03.03.482887.
13. Le, H. S. & Bar-Joseph, Z. Integrating sequence, expression and interaction data to determine condition-specific miRNA regulation. *Bioinformatics* **29**, i89-97 (2013).
14. Zhang, J. *et al.* Inferring condition-specific miRNA activity from matched miRNA and mRNA expression data. *Bioinformatics* **30**, 3070–3077 (2014).
15. Yoon, S. *et al.* Biclustering analysis of transcriptome big data identifies condition-specific microRNA targets. *Nucleic Acids Res* **47**, e53 (2019).
16. Zhang, J. *et al.* Exploring cell-specific miRNA regulation with single-cell miRNA-mRNA co-sequencing data. *BMC Bioinformatics* **22**, 578 (2021).
17. Ludwig, N. *et al.* Distribution of miRNA expression across human tissues. *Nucleic Acids Res* **44**, 3865–3877 (2016).

18. Dai, H., Li, L., Zeng, T. & Chen, L. Cell-specific network constructed by single-cell RNA sequencing data. *Nucleic Acids Res* **47**, e62 (2019).
19. Yu, X., Zeng, T., Wang, X., Li, G. & Chen, L. Unravelling personalized dysfunctional gene network of complex diseases based on differential network model. *J Transl Med* **13**, 189 (2015).
20. Zhang, W., Zeng, T., Liu, X. & Chen, L. Diagnosing phenotypes of single-sample individuals by edge biomarkers. *J Mol Cell Biol* **7**, 231–241 (2015).
21. Liu, X., Wang, Y., Ji, H., Aihara, K. & Chen, L. Personalized characterization of diseases using sample-specific networks. *Nucleic Acids Res* **44**, e164 (2016).
22. Kuijjer, M. L., Tung, M. G., Yuan, G., Quackenbush, J. & Glass, K. Estimating sample-specific regulatory networks. *iScience* **14**, 226–240 (2019).
23. Li, L., Dai, H., Fang, Z. & Chen, L. c-CSN: Single-cell RNA sequencing data analysis by conditional cell-specific network. *Genomics Proteomics Bioinformatics* **19**, 319–329 (2021).
24. Wang, X., Choi, D. & Roeder, K. Constructing local cell-specific networks from single-cell data. *Proc Natl Acad Sci U S A* **118**, (2021).
25. Zhang, S. Y. & Stumpf, M. P. H. Inferring cell-specific causal regulatory networks from drift and diffusion. in *The 2022 ICML Workshop on Computational Biology* (2022).
26. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45**, 1113–1120 (2013).

27. Clough, E. & Barrett, T. The Gene Expression Omnibus database. *Methods Mol Biol* **1418**, 93–110 (2016).
28. Agarwal, V., Bell, G. W., Nam, J. W. & Bartel, D. P. Predicting effective microRNA target sites in mammalian mRNAs. *Elife* **4**, e05005 (2015).
29. Li, J. H., Liu, S., Zhou, H., Qu, L. H. & Yang, J. H. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res* **42**, D92–97 (2014).
30. Székely, G. J., Rizzo, M. L. & Bakirov, N. K. Measuring and testing dependence by correlation of distances. *The Annals of Statistics* **35**, 2769–2794 (2007).
31. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* **33**, 1–22 (2010).
32. Wilcox, R. *Introduction to robust estimation and hypothesis testing*. (Academic Press, 2017). doi:10.1016/B978-0-12-804733-0.00008-1.
33. Lance, G. N. & Williams, W. T. Computer programs for hierarchical polythetic classification (“similarity analyses”). *The Computer Journal* **9**, 60–64 (1966).
34. Cantrell, C. D. *Modern mathematical methods for physicists and engineers*. (Cambridge University Press, 2000). doi:10.1017/9780511811487.
35. Huang, J. C. *et al.* Using expression profiling data to identify human microRNA targets. *Nat Methods* **4**, 1045–1049 (2007).

36. Quinn, T. P., Richardson, M. F., Lovell, D. & Crowley, T. M. *propr: An R-package for identifying proportionally abundant features using compositional data analysis.* *Sci Rep* **7**, 16252 (2017).
37. Liu, X. *et al.* Quantifying critical states of complex diseases using single-sample dynamic network biomarkers. *PLoS Comput Biol* **13**, e1005633 (2017).
38. Yang, B. *et al.* Dynamic network biomarker indicates pulmonary metastasis at the tipping point of hepatocellular carcinoma. *Nat Commun* **9**, 678 (2018).
39. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
40. Larrañaga, P., Karshenas, H., Bielza, C. & Santana, R. A review on probabilistic graphical models in evolutionary computation. *Journal of Heuristics* **18**, 795–819 (2012).
41. Marx, V. Method of the Year: spatially resolved transcriptomics. *Nat Methods* **18**, 9–14 (2021).
42. Felipe Lima, J., Nofech-Mozes, S., Bayani, J. & Bartlett, J. M. S. EMT in breast carcinoma-a review. *J Clin Med* **5**, (2016).
43. Ye, X. *et al.* Upholding a role for EMT in breast cancer metastasis. *Nature* **547**, E1–E3 (2017).
44. Neelakantan, D. *et al.* EMT cells increase breast cancer metastasis via paracrine GLI activation in neighbouring tumour cells. *Nat Commun* **8**, 15773 (2017).
45. Kröger, C. *et al.* Acquisition of a hybrid E/M state is essential for tumorigenicity of basal breast cancer cells. *Proc Natl Acad Sci U S A* **116**, 7353–7362 (2019).

46. Tan, T. Z. *et al.* Epithelial-mesenchymal transition spectrum quantification and its efficacy in deciphering survival and drug responses of cancer patients. *EMBO Mol Med* **6**, 1279–1293 (2014).
47. Hänelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14**, 7 (2013).
48. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47 (2015).
49. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300 (1995).
50. Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U. & Segal, E. The role of site accessibility in microRNA target recognition. *Nat Genet* **39**, 1278–1284 (2007).
51. Miranda, K. C. *et al.* A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell* **126**, 1203–1217 (2006).
52. Vejnar, C. E. & Zdobnov, E. M. MiRmap: comprehensive prediction of microRNA target repression strength. *Nucleic Acids Res* **40**, 11673–11683 (2012).
53. Paraskevopoulou, M. D. *et al.* DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. *Nucleic Acids Res* **41**, W169-173 (2013).
54. Enright, A. J. *et al.* MicroRNA targets in Drosophila. *Genome Biol* **5**, R1 (2003).
55. Krek, A. *et al.* Combinatorial microRNA target predictions. *Nat Genet* **37**, 495–500 (2005).

56. Pearson, K. Notes on the history of correlation. *Biometrika* **13**, 25–45 (1920).
57. Spearman, C. ‘General intelligence,’ objectively determined and measured. *The American Journal of Psychology* **15**, 201–292 (1904).
58. Kendall, M. G. A new measure of rank correlation. *Biometrika* **30**, 81–93 (1938).
59. Lopez-Paz, D., Hennig, P. & Schölkopf, B. The randomized dependence coefficient. in *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1* 1–9 (Curran Associates Inc., 2013).
60. Hoeffding, W. A non-parametric test of independence. *The Annals of Mathematical Statistics* **19**, 546–557 (1948).
61. Prill, R. J. *et al.* Towards a rigorous assessment of systems biology models: the DREAM3 challenges. *PLoS One* **5**, e9202–e9202 (2010).
62. Zar, J. *Biostatistical analysis*. (Prentice-Hall/Pearson, 2010).
63. Deza, E. & Deza, M.-M. *Dictionary of distances*. (Elsevier, 2006). doi:10.1016/B978-044452087-6/50006-8.
64. Deza, M. M. & Deza, E. Encyclopedia of distances. in *Encyclopedia of Distances* (eds. Deza, E. & Deza, M. M.) 1–583 (Springer Berlin Heidelberg, 2009). doi:10.1007/978-3-642-00234-2\_1.
65. Craw, S. Manhattan distance. in *Encyclopedia of Machine Learning* (eds. Sammut, C. & Webb, G. I.) 639–639 (Springer US, 2010). doi:10.1007/978-0-387-30164-8\_506.

66. Dice, L. R. Measures of the amount of ecologic association between species. *Ecology* **26**, 297–302 (1945).
67. Duda, R., Hart, P. & G. Stork, D. Pattern classification. in *Wiley Interscience* vol. xx (2001).
68. Mahalanobis, P. C. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)* **2**, 49–55 (1936).
69. Kraskov, A., Stögbauer, H. & Grassberger, P. Estimating mutual information. *Phys. Rev. E* **69**, 066138 (2004).
70. Reshef, D. N. *et al.* Detecting novel associations in large data sets. *Science* **334**, 1518–1524 (2011).
71. Maathuis, M. H., Kalisch, M. & Bühlmann, P. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics* **37**, 3133–3164 (2009).
72. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal, Complex Systems* **1695**, 1–9 (2006).
73. Massey, F. J. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association* **46**, 68–78 (1951).
74. Zhang, J. *et al.* miRspongeR 2.0: an enhanced R package for exploring miRNA sponge regulation. *Bioinformatics Advances* **2**, vbac063 (2022).
75. Chen, J. *et al.* RNADisease v4.0: an updated resource of RNA-associated diseases, providing RNA-disease analysis, enrichment and prediction. *Nucleic Acids Res* **51**, D1397–D1404 (2023).

76. Huang, H. *et al.* miRTarBase update 2022: an informative resource for experimentally validated miRNA-target interactions. *Nucleic Acids Res* **50**, D222–D230 (2022).
77. Karagkouni, D. *et al.* DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA-gene interactions. *Nucleic Acids Res* **46**, D239–D245 (2018).
78. Zhang, J. *et al.* miRSM: an R package to infer and analyse miRNA sponge modules in heterogeneous data. *RNA Biol* **18**, 2308–2320 (2021).

## Acknowledgements

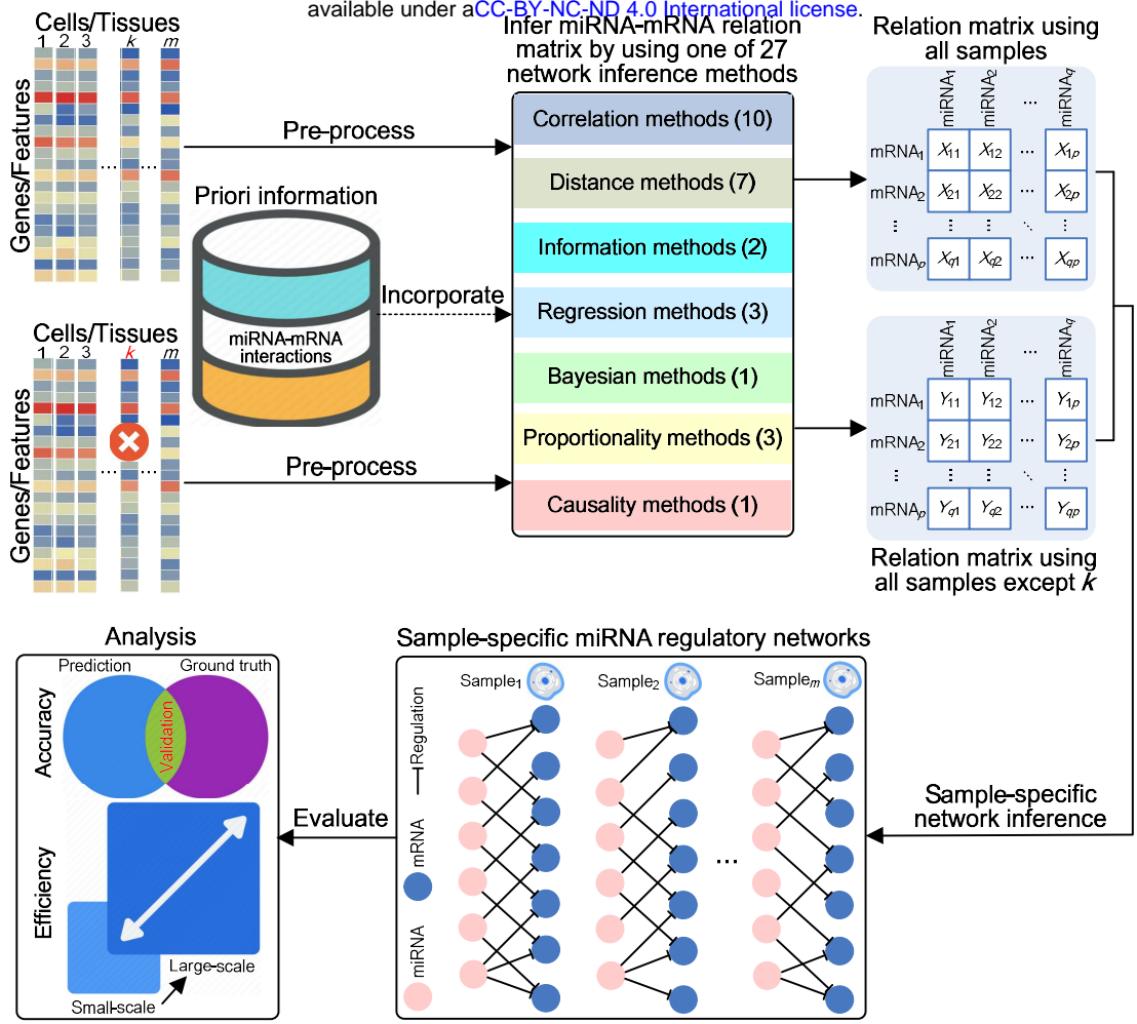
This work has been supported by the National Natural Science Foundation of China (61963001), the Yunnan Fundamental Research Projects (202001AT070024, 202101BA070001-221), and the ARC DECRA (DE200100200).

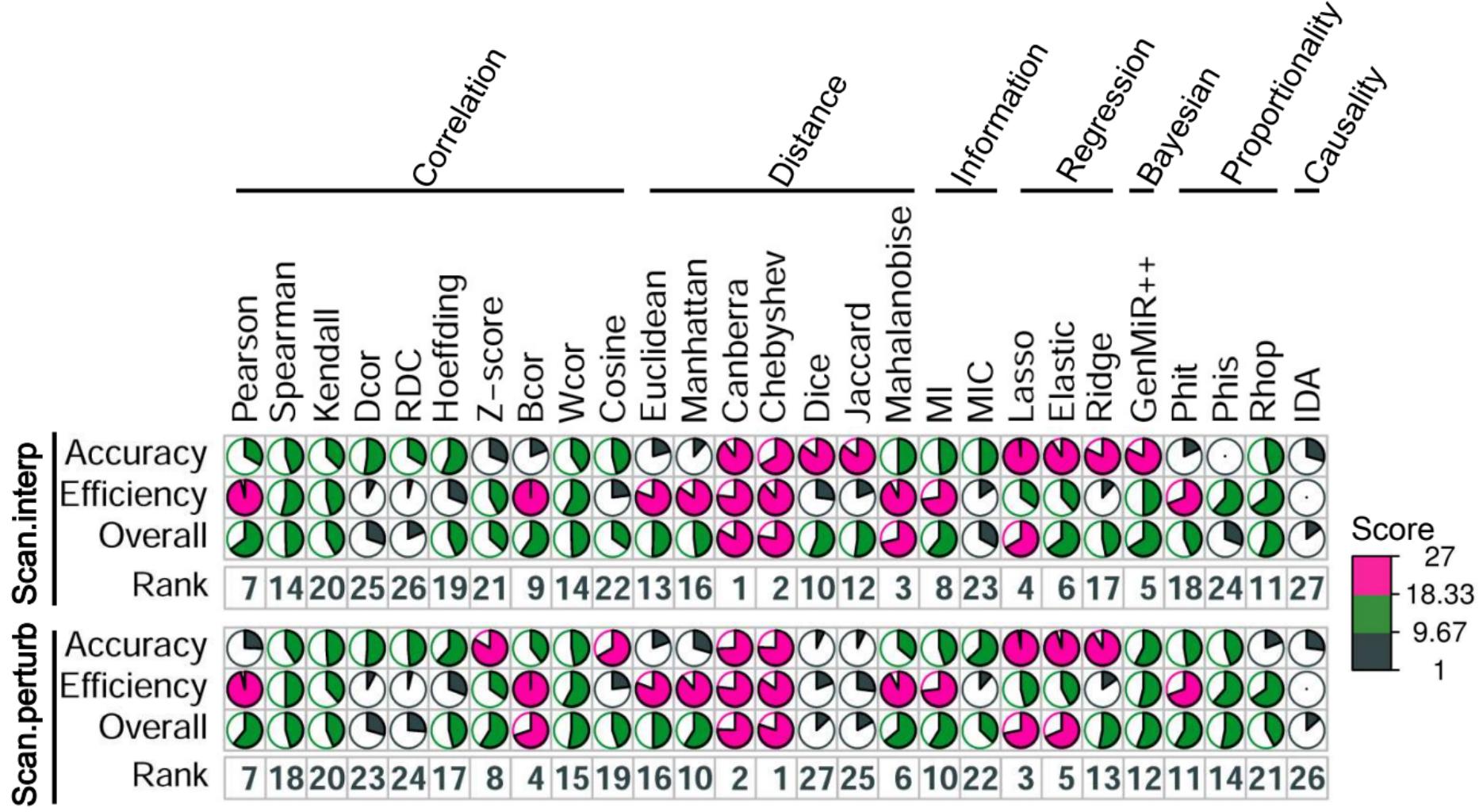
## Author contributions

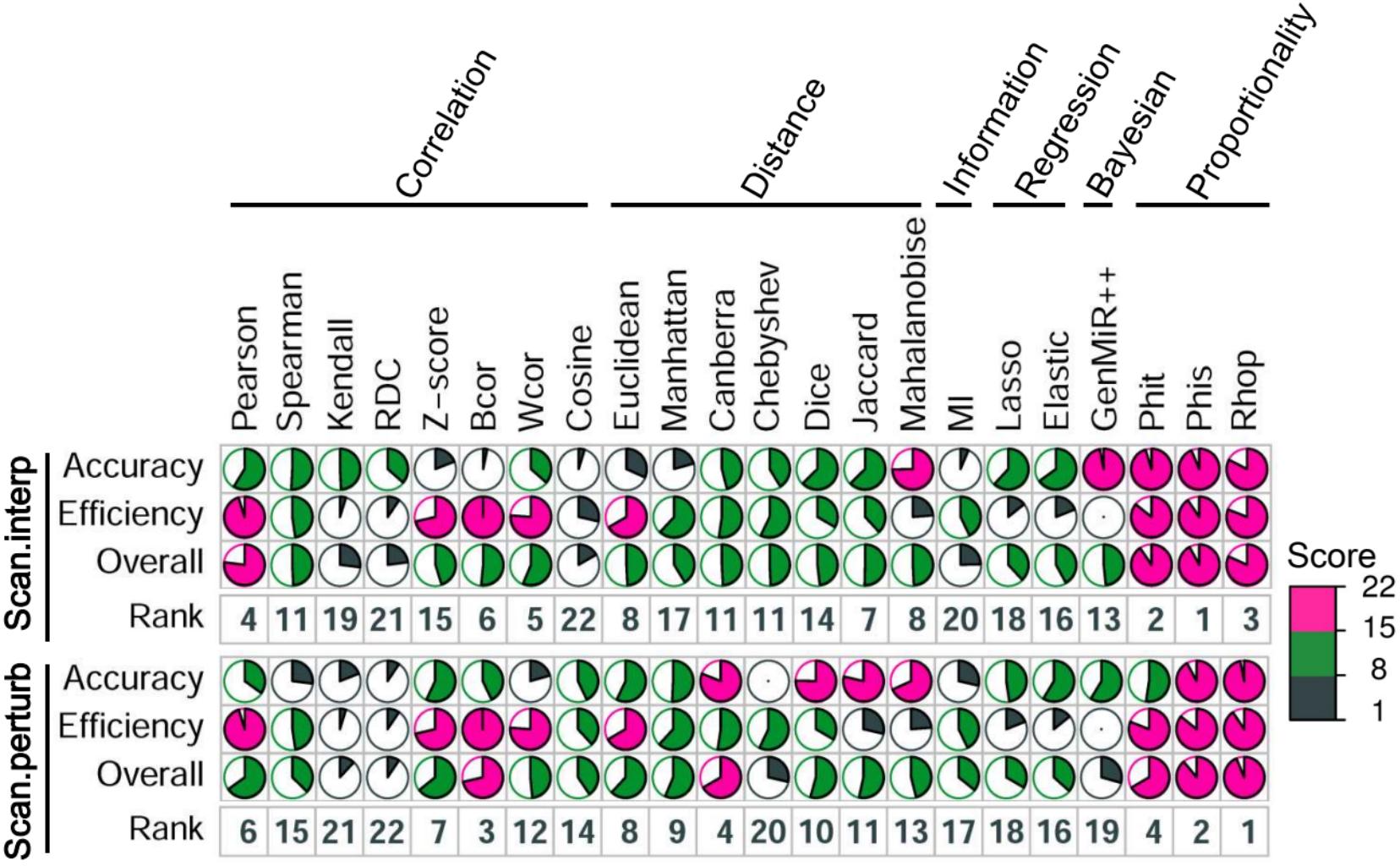
J.Z. and T.D.L. conceived the idea of this work. L.L. and J.L. refined the idea. J.Z. designed and performed the experiments. X.W., C.Z. and Y.L. participated in the design of the study and performed the statistical analysis. J.Z., L.L., T.D.L. and J.L. drafted the manuscript. All authors revised the manuscript. All authors read and approved the final manuscript.

## Competing interests

The authors declare no competing interests.

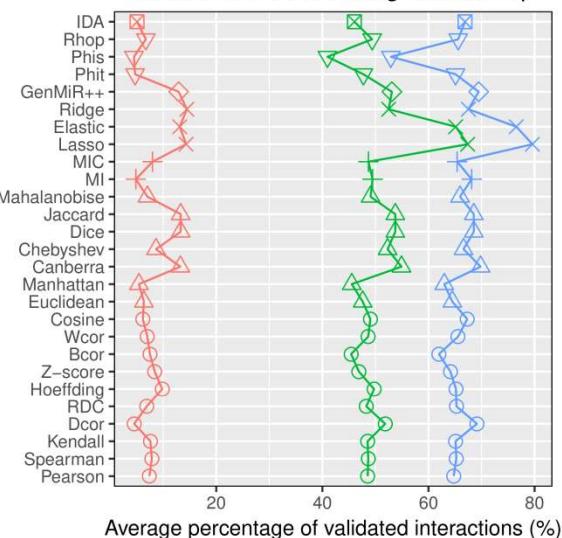




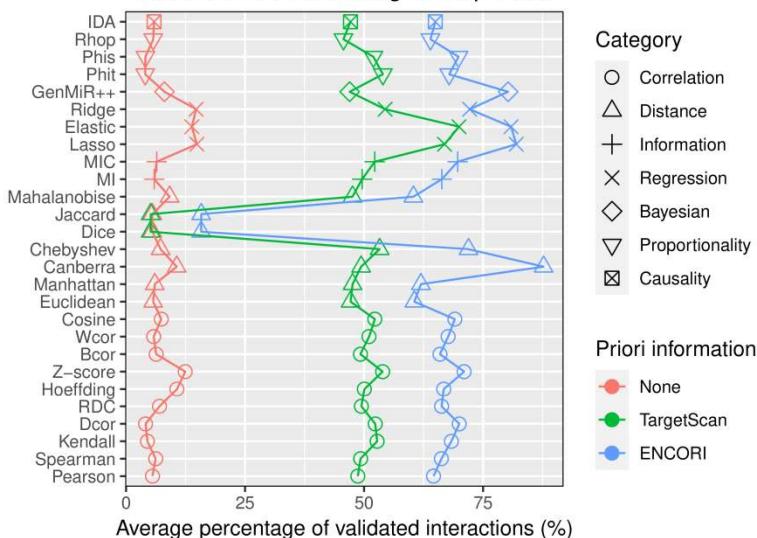


a

## Validation in K562 using Scan.interp

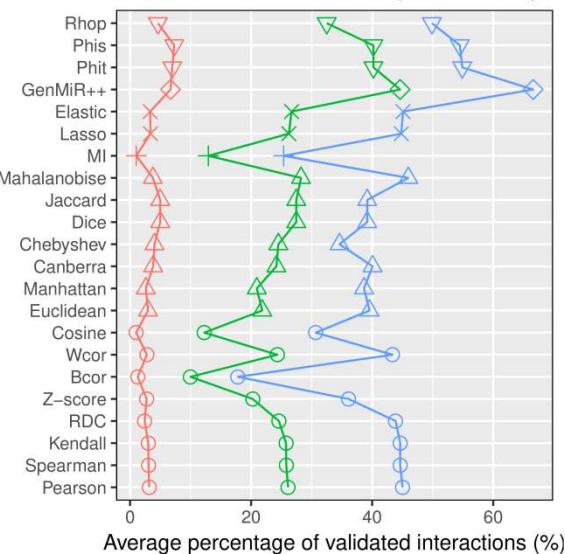


## Validation in K562 using Scan.perturb

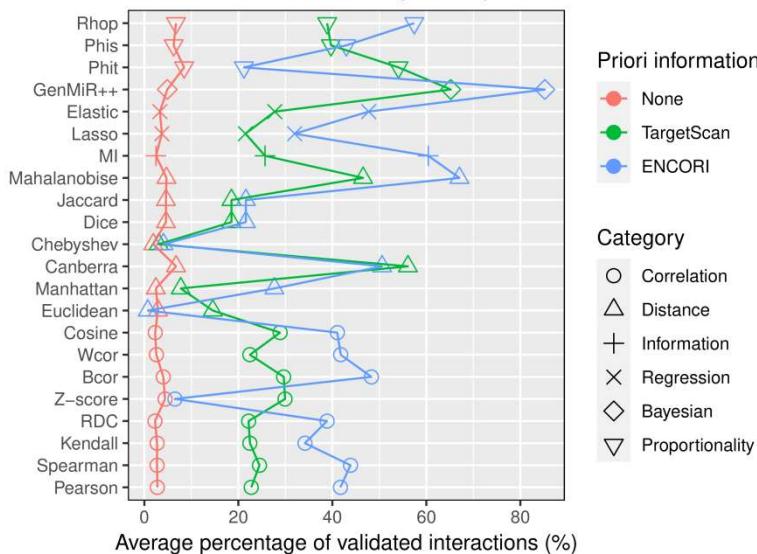


b

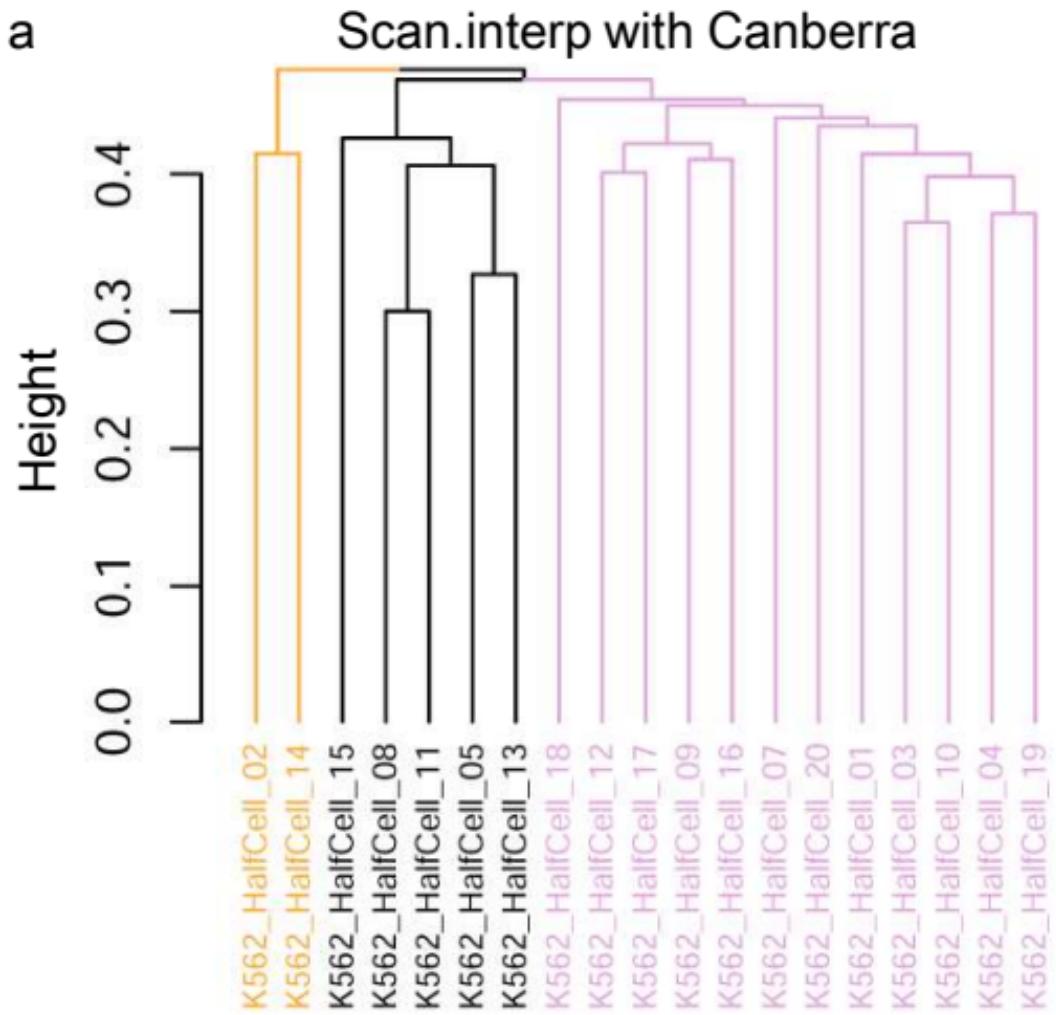
## Validation in BRCA using Scan.interp



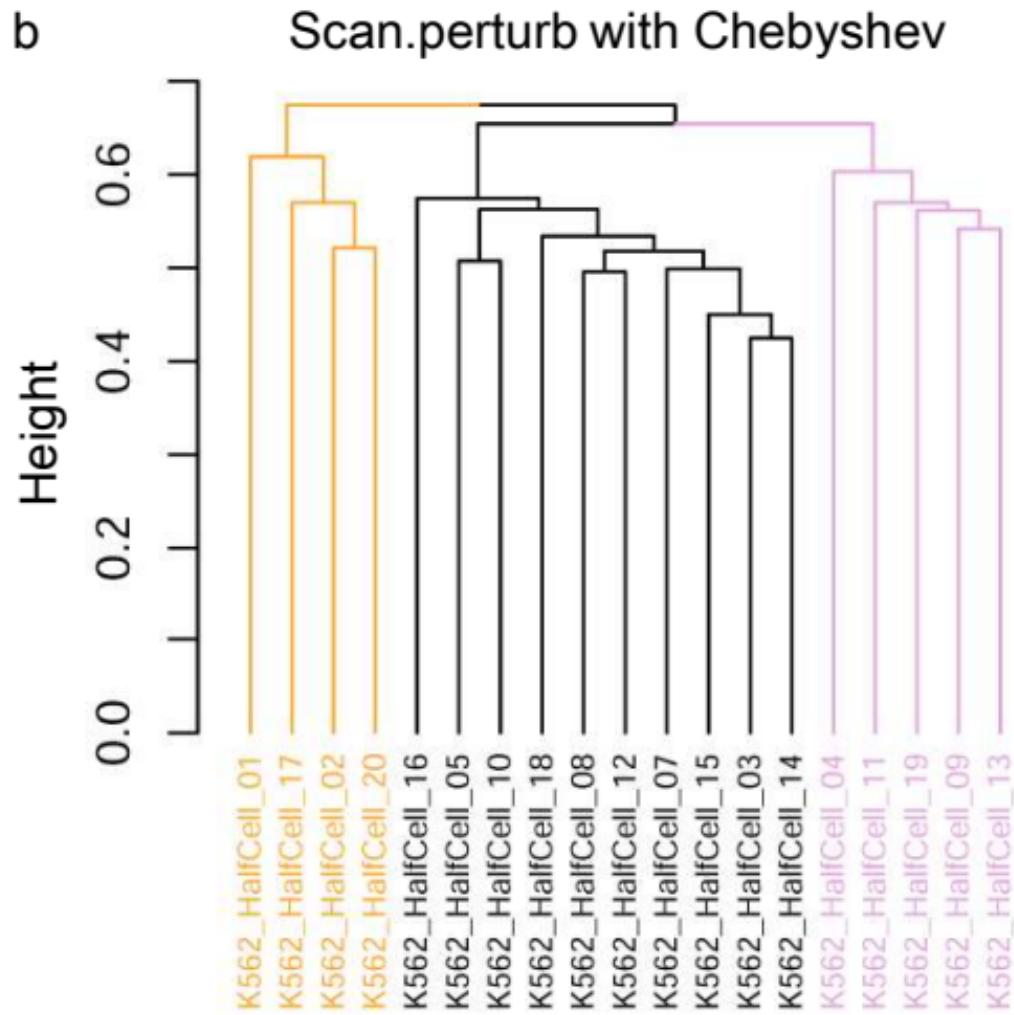
## Validation in BRCA using Scan.perturb



a

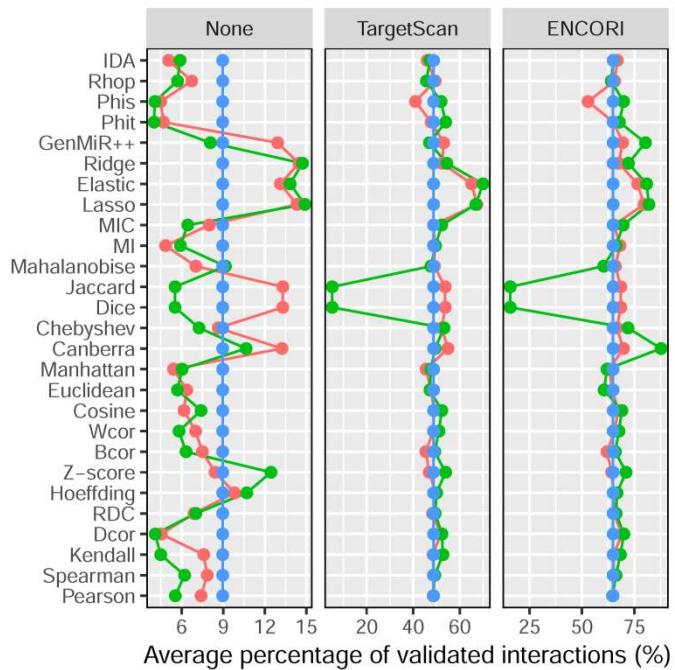


b



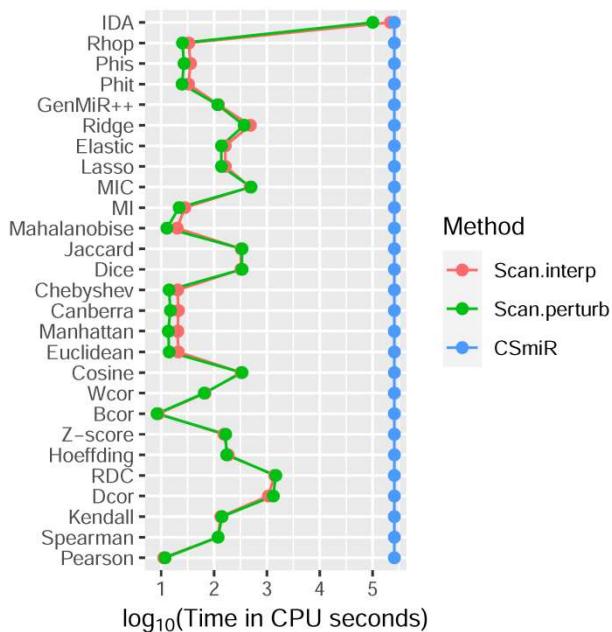
a

## Accuracy with the K562 dataset



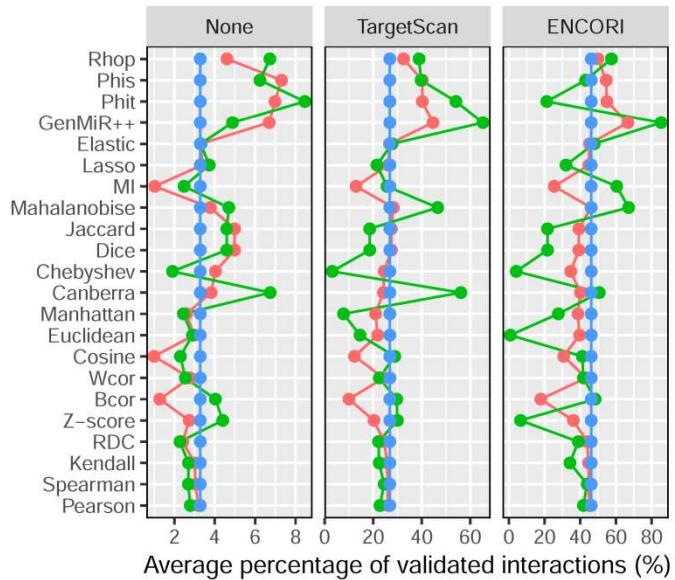
b

## Efficiency with the K562 dataset



c

## Accuracy with the BRCA dataset



d

## Efficiency with the BRCA dataset

