AARHUS UNIVERSITY

Bachelor Projekt

Use of a Deep Generative Decoder to infer mRNA differential expression and thereby miRNA expression

Students
Anna Sofie Hvid Christensen
Studienr. 202107879

Elisa Lægsgaard
Studienr. 202104652

Supervisor
Jakob Skou Pedersen

june 2024

Aarhus University
Department of Mathmatics
Data Science

Contents

1. Abstract

This project aims to improve our ability to estimate micro RNA (miRNA) expression levels from measured messenger RNA (mRNA) expression levels. Each miRNA has the ability to down regulate a number of mRNA targets. Methods such as miReact exploit this relationship, by using available mRNA expression data to infer miRNA activity. Accurate estimation of mRNA expression changes is thus critical for accurate inference. We evaluate if miReact performance can be improved by using a newly published method for inferring mRNA expression changes based on a generative AI model called a Deep Generative Decoder (DGD). We apply the DGD model to the thousands of cancer samples from the Cancer Genome Atlas (TCGA) dataset, where both mRNA and miRNA expression data is available. We benchmark the performance of miReact when using traditional methods for inferring cancer sample specific mRNA expression changes versus using the DGD-based approach. The project involves both studying, implementing, and applying generative AI in a big cancer genomics data setting. The inferred miRNA expression levels based on the DGD approach show a strong correlation with measured expression for many miRNAs. When analyzing liver specific miR-122 and tumor related let-7a, we obtain inferred expression levels consistent with literature. By investigating the distributions of inferred miRNA expression levels, results demonstrate a distinct ability to infer how activity of miRNA changes in cancer. These results demonstrate the potential of using DGD in combination with miReact to infer changes in miRNA activity due to cancer.

2. Introduction

2.1 The impact of microRNA on gene expression

Micro RNA (miRNA) are small non-coding RNAs that play a crucial role in the regulation of gene expression. Disruptions in miRNAs are often linked to diseases such as cancer. To understand their role, one must know a core principle in molecular biology: gene expression. Gene expression describes the intricate process through which the information stored in DNA is expressed as proteins in the cell.[1]
The process of gene expression is often shortly explained in two steps:

1. **Transcription**: The conversion of the DNA sequence of a gene into messenger RNA (mRNA).

2. **Translation**: The conversion of these mRNAs into proteins.

The genetic information stored within DNA on how to build different proteins are copied into small, portable pieces of RNA called mRNA. These mRNAs travel from within the cell's nucleus out to the cells cytoplasm, where they are read by ribosomes that translate them into specific proteins [8].
However, different cells have different functions and thereby need different combinations of proteins to function correctly. The specialization between cells for different tissue types is achieved by multiple regulatory processes. One of the key regulators of gene expression is miRNA[8].

Much like mRNA, miRNAs also stores small pieces of genetic information transcribed from the DNA strand. However, miRNAs are non-coding since they do not code for any protein. The first transcript has a hairpin structure (see Figure 1.a), which is then processed by the

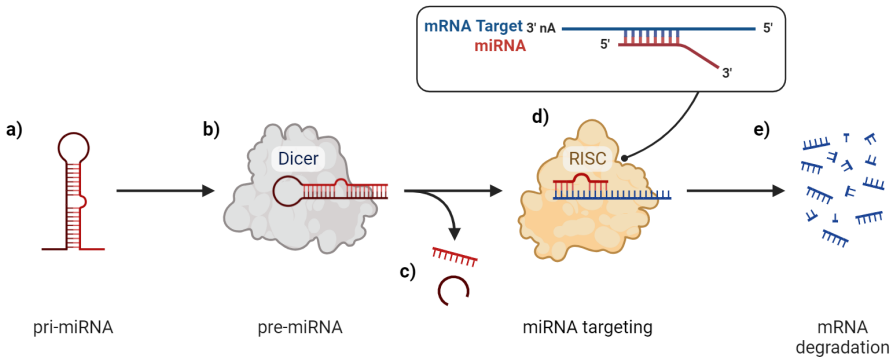


Figure 1: miRNA processing. **a)** The first transcript, primary miRNA transcript (pri-miRNA), cleaved by the Drosha enzyme to create **b)** the precurser miRNA (pre-miRNA), which is then transported to the cell cytoplasm, where it is cleaved by Dicer to generate double stranded miRNA. **c)** Passenger strand is discarded, whilst **d)** guide strand forms the miRISC with RISC, which the miRNA guide strand guides to find target mRNA, triggering the RISC to **e)** inhibit translation or degrade the mRNA [11]. *Adapted from "miRNA Processing" by BioRender.com [5].*

enzyme Drosha and exported into the cell cytoplasm. Here, it is further processed by the enzyme Dicer, and the terminal loop is removed leaving a double stranded miRNA (see Figure 1.b). This double-stranded miRNA is incorporated into the RNA-induced-silencing-complex (RISC) which separates the strands of the miRNA and uses one as a guide and discards the other (see Figure 1.d). The two sides of the strand are not identical and have different target sites leading to 5p and 3p versions of miRNA dependent on which side is used in the RISC complex. The miRNA-RISC complex binds to complementary target sequences usually found within the mRNAs 3'untranslated region (3'UTR). To bind to the 3'UTR region of a mRNA, the *seed* region of the miRNA, typically defined as nucleotides at positions 2 to 7 or 2 to 8 at the beginning of the miRNA, must be perfectly base-paired with the mRNA for efficient targeting. The miRNAs will bind to these target sites, leading to either translation repression and/or degradation of the mRNA through cleavage, mediated by an endonuclease in RISC [8, 11]. The miRNA process in gene-silencing is depicted in figure 1.

Some miRNAs are tissue specific and mainly only found in that one tissue type. An example of a highly tissue-specific miRNA is miR-122, which is highly expressed in liver. It constitutes about 70% of the total miRNA expression found in liver. It is therefore often used as a biomarker to identify liver samples[7]. Others have a basic function and can be found expressed in most cells. They can also have target sites on multiple different mRNAs, with a given miRNA potentially affecting the regulation of up to 400 target genes. Overall the expression of at least 60% of all human protein-encoding genes is regulated in part by miRNAs[11].

2.2 miRNAs for cancer diagnosis and prognosis

It has been proven that there is a strong link between the development, progression, and metastasis of cancer cells and the dysregulation of miRNA expression[15].
This is caused by the up or downregulation of different oncomiRs (miRNAs associated with cancer). Since miRNAs directly affect the up or downregulation of mRNAs and the protein product of these mRNAs, it is easy to connect miRNAs to cancer if the direct target of the miRNA is an oncogene, a gene with the potential to cause cancer, or a tumor suppressor[9].
An example of this is the miRNA family let-7, which directly regulates RAS oncogenes; a family of genes that encode proteins involved in cell signaling pathways regulating cell growth and division. An up-regulation of the RAS oncogenes have often been shown to have a strong presence in cancer samples, with the most extreme case being the identification of K-RAS mutations in

90% of pancreatic cancer samples [19].
Similar to the tissue-specific miRNAs discussed in section 2.1 that can serve as biomarkers for certain tissues, miRNAs with a strong association to either oncogenes or tumor suppressors can act as biomarkers for cancer. For example, an abnormal amount of let-7 activity could be used as a biomarker for cancer.
Given that miRNA expression levels is directly tied to mRNAs connected to the development and progression of cancer, being able to classify whether miRNAs are significantly up or down-regulated could be used for diagnosis and prognosis[9].

2.3 miReact

miRNA cannot currently be studied at the same scale as mRNA, making the study of miRNA expression in cases such as cancer difficult. However, by utilizing mRNA expressions produced by high-throughput RNA sequencing, miRNA can be derived through miRNA-mRNA interaction models.

The paper "miRNA activity inferred from single cell mRNA expression" by Nielsen and Pedersen introduces the statistical tool *miReact* to infer bulk miRNA activity scores from bulk mRNA expression profiles. miReact exploits the fact that miRNAs bind to well-defined sequence motifs in their target mRNA, and by analyzing the abundance of these motifs and the association with expression level, the activity of the corresponding miRNA can be inferred.

2.3.1 Inferring miRNA activity from bulk mRNA expression profiles

miReact utilizes that the presence of a specific miRNA's target site within the 3' UTR indicates interaction between the mRNA and miRNA pair. As a result, the relative expression of the mRNA is dependent on the activity of the miRNA, or in reverse, miRNA activity is strongly related to the expression level of it's mRNA target. By identifying the presence of known 7-mer target sites of miRNAs in the 3' UTR of mRNAs, and combining this information with mRNA expression, miReact conducts motif enrichment analysis to estimate miRNA activity. Specifically, motif enrichment analysis involves ranking the 3' UTR sequences based on their relative expression levels in a given sample, and statistically assessing whether miRNA target sites tend to cluster along the ranked sequence list. The statistical significance of clustering is then used as the activity score [12, 13]

miRNA binding model
The bases from position 2 to 8 from the 5'end is used to define the seed site of a given miRNA, and the 7-mer complementary sequence is used to define the binding site (target site).

Sequence specific p-values
For a bulk mRNA expression profiling dataset, the 3'UTR sequences of each mRNA and the 7-mer seed site of a given set of miRNA is collected. To identify the presence of the 7-mers within each 3'UTR sequence, the statistical tool *Regmex* [13] is used. Regmex identifies motifs - defined by regular expressions - in DNA sequences and calculates sequence specific p-values (SSPs) for observing the motif by embedded Markov models. In a ranked sequence list, it is possible that motifs will cluster in one end of the ranking due to sequences having variable lengths, and hence making motifs more or less probable. By using a Markov model that depends on both sequence lengths and base compositions to compute sequence specific/dependent p-values, sequence length bias should be countered[13].

Relative expression
Sequences are ranked by relative expression. The idea behind this approach is to evaluate whether mRNAs with miRNA target sites are downregulated compared to other mRNAs. Expression levels are made comparable between mRNAs by transforming expression into fold-changes based on a control sample. In the this setting, however, we do not have control samples available. Fold-changes are instead computed relative to the median mRNA expression level

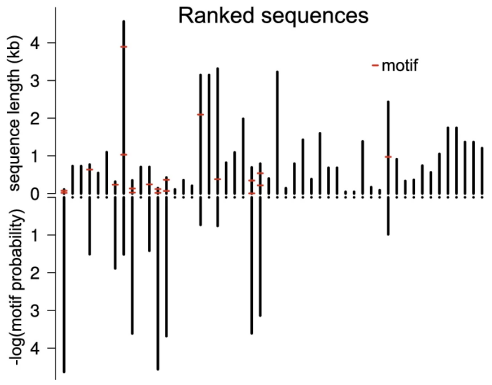
across all samples [12]. For a given sample j the relative expression of the mRNA i, denoted $\Delta e_{i,j}$ is:

$$\Delta e_{i,j} = e_{i,j} - \tilde{e}_i \tag{1}$$

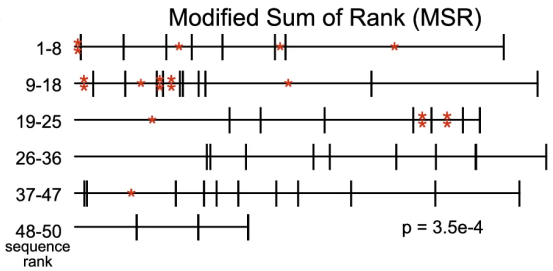
Where $\tilde{e}_i$ denotes the median mRNA expression level across all samples for mRNA i [12].

Motif rank correlation p-value

For each sample, 3'UTR sequences are ordered by the relative expression level, with lowly expressed sequences at the start of the ranked list and highly expressed sequences at the end. Each 3'UTR sequence is annotated with the presence of miRNA binding sites [12]. Regmex is used to perform a statistical evaluation of whether motifs tend to cluster in any highly- or lowly ranked regions of the ordered list. Regmex implements a modified rank sum test as statistical test for motif clustering along the sequence list, using a null-hypothesis that no clustering occurs. The modified sum of ranks method is based on using a rank sum test to determine a rank bias in motif containing sequences [13]. The modified sum of ranks test does not *sum* ranks, but instead uses a normalized sum of scores based on the computed sequence specific p-values, since these eliminate biases of sequence length and sequence composition. The normalized score is constructed by scaling the ranks of the sequences proportionally to the sequence specific probabilities. An example of 50 ranked sequences and their log-transformed SSPs is illustrated in figure 2a [13].



(a) A ranked sequence list with motif occurrences marked with red. log-transformed sequence specific p-values are included below.



(b) Visualization of the MSR method. Lines represent sequences with lengths proportional to the probability of observing motifs one or more times. Motif is marked by an asterisk [13].

Figure 2: Modified Sum of Ranks Test. Modified figure from Nielsen et al.. [13]

Motif occurrences are thought of as a Poisson process, where the "time axis" is a finite interval representing the ranked sequence list. The "time axis" is composed of consecutive sub-intervals, each representing a 3'UTR sequence, with a length equal to the normalized scores. The sub-intervals are ordered according to their relative expression. This is illustrated in figure 2b, which visualizes the so-called "time axis" divided into sub-intervals representing the ranked sequences. The constructed "time axis" can then be thought of as an interval where motif events will randomly occur along the interval. Under the assumption or null-hypothesis that no clustering should occur across the sequence list, the motif distribution along the "time axis" is expected to be even when using the normalized scores. Or equivalently, the probability of motif events is proportional to the length of a given sub-interval. When a motif occurs in a sequence, the motif is then assigned a rank score based on its position along the time axis. As a rank score the midpoint of the sub-interval or sequence containing the motif (as illustrated in figure 2b) is used. This is also called the *motif occurrence score*[13]. A test statistic is then constructed using the weighted average of the motif occurrence scores (where each motif occurrence score for a sequence is weighted by the number motif occurrences) for a given motif. Under the null hypothesis, motif occurrence scores are expected to be uniformly distributed across the entire time axis, and hence the weighted average and test statistic becomes normally

distributed. The rank correlation p-value (RCP) can then be computed as as the probability of observing extreme values of the weighted average[13].

miRNA activity score The RCP is computed for each miRNA target site motif for each sample. The activity score is then computed as [12]:

$$\log 10(\text{RCP}) * sign(\text{statistics})$$

Note, the test statistic becomes a normal null distribution with mean 0, due to the expected value of the weighted average being subtracted when constructing the test statistic [13]. This means that large negative observed test statistic values indicate a clustering of motifs or a "bias of ranking" towards the end of the ranked list with low relative expression values, while large positive values indicate a clustering towards relatively high expression levels. By scaling the RCP value by the sign of the statistic, negative activity scores should then indicate a downregulation of miRNA activity, whilst positive scores indicate upregulation. Activity scores of 0 should indicate no clustering of motifs towards either end of the ranked sequence list.

2.3.2 Published findings

The paper by Nielsen and Pedersen[12] demonstrates that miReact miRNA activity scores inferred from bulk mRNA expression profiles correlate with bulk miRNA expression levels. In the paper about ten thousand samples of expression profiling data from The Cancer Genome Atlas (TCGA) are utilized to infer activity, and for about half of the samples miRNA expression profiling data is available. The paper further looks at examples of tissue-specific miRNAs such as the liver-specific miR-122. The achieved miR-122 activity scores were significant in liver and further elevated liver cancer samples compared to samples with other tissues of origin, which corresponds with the observed miR-122 expression levels observed across the same samples. However, many miRNAs did not show a high correlation between activity and expression, which may be due to several reasons, as noted in the paper. The low correlation can be due to miRNA activity not resulting in direct degradation of mRNA but instead inhibiting translation without reducing mRNA expression, or due to the choice of binding-site modelling, which can be inaccurate as outlined in the paper. Lastly, miReact does not have control samples in the applied setting, and as substitute the median gene expression level across all samples is used, which can cause inaccuracies[12].

2.4 Aim and hypothesis

The objective of this paper is to improve the performance of the statistical tool miReact. As described, we do not traditionally have control samples available in the applied setting, which can potentially have an impact on the inference accuracy of miReact. To address this limitation, we propose to apply and integrate a Deep Generative Decoder (DGD) [18], which eliminates the need for healthy controls when inferring miRNA activity for cancer samples. We hypothesize that this integration can improve the accuracy of miReact results. Subsequently, we will evaluate the performance of the integrated model against the original miReact performance.

2.5 The deep generative decoder

The paper "N-of-one differential gene expression without control samples using a deep generative model" by Prada-Luengo et al. presents a Deep Generative Decoder (DGD)[20] trained on mRNA expression data from healthy normal samples. By training on normal samples, the model is capable of identifying the closest normal representation for any given disease sample. This extends the analysis of gene expression to individual cancer samples without the need for control samples[18].
Utilizing DGD in place of the global median in miReact could provide a more accurate measure of whether the gene expression of a sample is abnormal (see sec. 2.3.1).

The model is based on the Deep Generative Decoder as described bySchuster and Krogh but trained by Prada-Luengo et al. on normal mRNA expression data from Genotype-Tissue Expression (GTEx) containing about 20.000 bulk samples of expression profiling from 31 different human tissues and 948 individuals.[18, 14]

2.5.1 The model structure of DGD

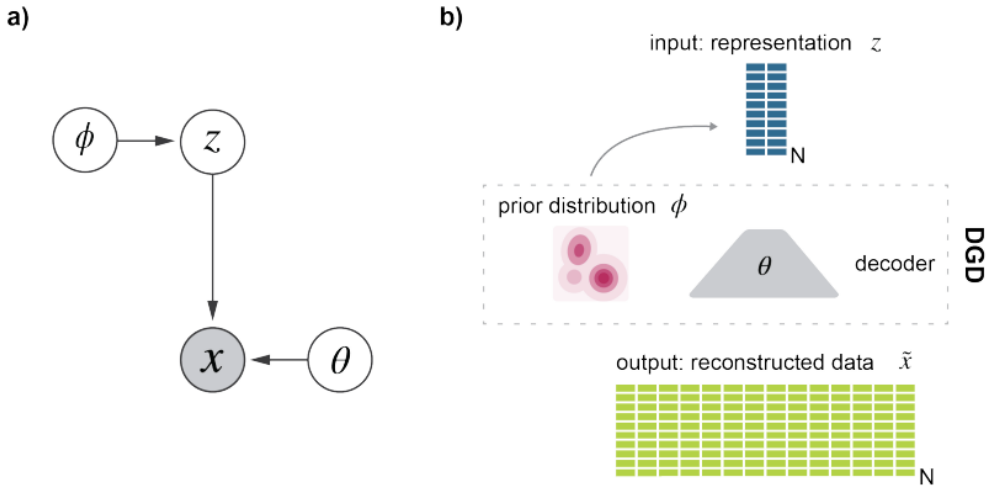


Figure 3: a) Graphical model and b) schematic of the deep generative decoder. Figure from Schuster and Krogh[20].

DGD is presented as a probabilistic formulation of a generative decoder using maximum a posteriori (MAP) estimation of all parameters. The DGD model consists of three parts: representations (Z), decoder (θ) and prior distribution (ϕ), as visualised in Figure.3[20].

In the bulkDGD model the distribution over latent space is represented by a Gaussian mixture model (GMM) with K components. The priors over the GMM parameters, ϕ, are trained to best capture the underlying structure of the data. The number of components, therefore, needs to be adjusted based on the subtlety of the structures that the model aims to capture[18].

The data samples, X, are mapped to a latent space, Z, of shape (N, m), where N is the number of samples and m is the dimension of the mean vector, μ, of the K GMM components. In the case of bulkDGD, the latent space is 50 (m) X 45 dimensional space[18].

The representations are fed into the decoder, which can be any type of neural network as long as it takes input dimensions, m, and outputs dimension, n, being the original number of data features. The neural network parameters (θ) are trained to take representations (z) as input and outputs $f_\theta(z)$.

2.5.2 A DGD trained on healthy samples

The pretrained DGD model (bulkDGD) is trained by Prada-Luengo et al. to generate controls for bulk RNA-seq studies, where controls are often absent because obtaining them would typically require operations on healthy individuals.
Training the DGD parameters focuses on identifying the tissue from which the sample is taken, describing data in a representational space, and then finding the most probable mRNA expression profile for these representations.
The neural network used in bulkDGD is a fully-connected feed-forward neural network with two hidden layers.

In the case of bulkDGD the outputs $f_\theta(z)$ are the parameters for a negative binomial over count values for every gene. Scaled up by the means of the samples, these should resemble the normal expression values of the given sample.

2.5.3 Training process of the pretrained DGD model

Parameters are initialized to zero for training and run through the training process described in algorithm 1 from [20]. During each epoch, the representations are updated along with the parameters of the decoder and the GMM.

Algorithm 1 TRAINING

Initialize parameters for representations $\mathbf{z^i}$, decoder and GMM
for epoch in n_{epochs} **do**
  for $\mathbf{x}^i, i$ in training data **do**
    $\mathbf{z}^i = \mathbf{Z}_i$
    $\mathbf{y}^i = \text{model}(\mathbf{z}^i)$
    total_loss = $L_{reconstruction}(\mathbf{y}^i, \mathbf{x}^i) + L_{GMM}(\mathbf{z}^i)$
    Backpropagation
    Optimizer step for model and GMM
  end for
  Optimizer step for Representation
end for

Reconstruction loss The Reconstruction loss represents $P(\mathbf{x}|\mathbf{z})$ and should reflect the models ability to predict gene expression levels based on the latent representations[20].
This loss is calculated between the expression count input and the rescaled bulkDGD output. The reconstruction loss corresponds to the negative log-probability, which is calculated for each of the negative binomial outputs, and then summed for the total reconstruction loss[17]. Hence, it is a measure for how well the model can generate outputs that resembles input data after running it through the model. The lower the value, the better the model performs when training.

GMM loss The GMM loss computes the negative log-probability density of $\mathbf{z}^i$ being drawn from the Gaussian mixture model.

Total loss The total loss is then used for finding the best representation (minimum loss) for each sample across all representations and components.

Backpropagation Backpropagation updates the parameters using separate optimizers for the decoder, GMM, and representations to be able to utilize different learning rates for different parameters.

2.5.4 Published findings

The paper found that most GMM components were represented by only one tissue, except for a few that were split between multiple tissues. Tissues could be represented by multiple components, most notably the brain, which was distributed across 8 of the 45 components[18]. When testing it on the TCGA-cancer samples it assigned the samples to the correct type more than 80% of the time for 11/14 tested tissuetypes.

2.5.5 Using the pretrained model for prediction on new datapoints

When applying the model to new data, the GMM and decoder parameters are fixed and only the representations are optimized to find the best representation for each new sample. This is

to prevent the model from being trained to output tumor samples. In the paper it is suggested to train the representations from 10-50 epochs, with 50 being recommended for differential expression analysis. Hence, 50 epochs is used for this papers analysis[18].
The longer the model optimizes the representations, the more it should begin to acquire traits comparable to the original tumor sample. This is because we initialize the representations from the component means. As training progresses, the model gradually diverges from these initial means in an effort to minimize reconstruction loss. Consequently, an extended training period allows the model to progressively deviate further from the mean, thereby enhancing its ability to capture nuanced characteristics that varies between individuals.
The process of predicting on a new sample by optimizing the representation is outlined in Algorithm 2 below:

Algorithm 2 PRETRAINED MODEL

epochs = 50
Initialize parameters for representations $\mathbf{z^i}$ from component means
for epoch in epochs **do**
  for $\mathbf{x}^i, i$ in training data **do**
    $\mathbf{z}^i = \mathbf{Z}_i$
    $\mathbf{y}^i = \text{model}(\mathbf{z}^i)$ # frozen parameters
    $\text{total\_loss} = L_{reconstruction}(\mathbf{y}^i, \mathbf{x}^i) + L_{GMM}(\mathbf{z}^i)$
  end for
  Optimizer step for Representation
end for

2.5.6 miReact inferring relative expression using DGD

The utilization of bulkDGD could potentially enhance the miReact analytical framework, particularly in addressing the current limitations in the measure for "normal" expression levels when calculating fold changes. By incorporating bulkDGD, the hopes are to achieve a more accurate and tissue-specific baseline for comparison. This approach would find measures of a relative miRNA activity score that is compared against normal tissue-specific expression profiles, rather than a general activity score for the sample.
The pseudo normal samples produced by bulkDGD would in theory retain some of the biological patterns that are unique to people, making it not only a comparable cell tissue-wise but also would convey what a normal sample expression pattern would look like for the individual. This would be an even better choice than the few normal controls that are currently used, since they are often taken from comparable individuals but not the actual patient.
The hypothesis is that the incorporating of the DGD model into miReact would give a more accurate analysis for deviations from the normative state of the cell. Thus, improving analysis for whether the miRNAs have been up of down regulated, leading to new biological insights and potential classification for whether or not we are looking at a cancer/disease sample. It could also help in other insights for potential therapeutic targets for cancer cells.

3. Data and methods

Using a dataset consisting of tumor and normal samples, we employ a DGD model to generate healthy controls for each sample and compute the relative expression (fold changes) between each sample and its control. Additionally, for the same dataset, we compute the relative expression using the global median as a control surrogate, as described in section \ref{sec}.

This serves as a baseline method for comparison. For each method of computing fold changes, we use miReact to infer activity for all samples. Finally, we will compare the inferred activity with observed miRNA expression profiling data for the corresponding samples. An overview of the process illustrated in Figure 4.

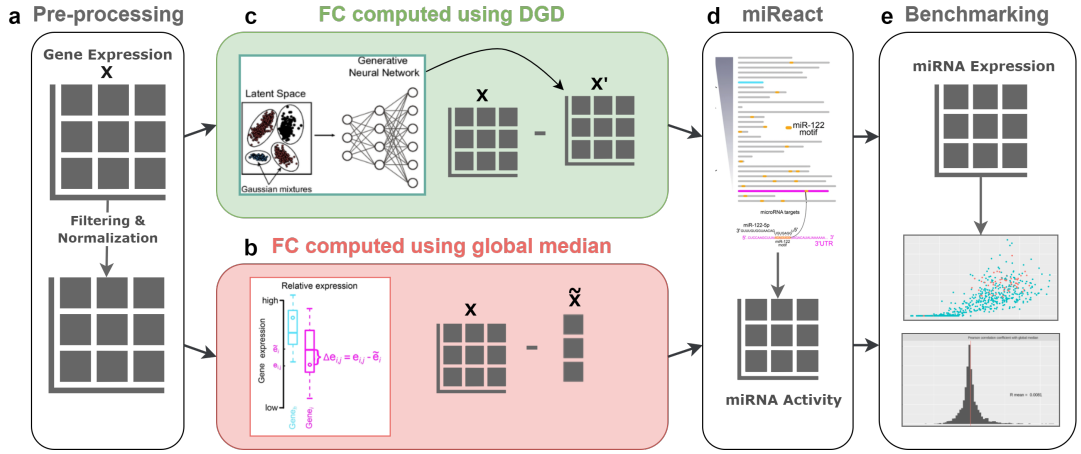


Figure 4: Study workflow overview. **a.** Data pre-processing: The expression matrix (raw counts) from TCGA is filtered and normalized before being used in miReact. **b.** Conventional method; fold changes are computed as described in [12] using global median across samples (denoted as $\tilde{X}$) as surrogate for controls. **c.** Expression matrix (note, the raw expression counts) is put through DGD to generate control samples X', which are then filtered and normalized before being used to compute fold changes. **d.** Fold changes computed by both methods are then each processed by miReact. **e.** Lastly, inferred activity is then benchmarked by comparing to miRNA expression profiling data. A more detailed overview can be found in Appendix B. Figures are modified from [20] and [12].

3.1 Data

We apply the models to cancer and normal sample from TCGA [6].
From the TCGA dataset, we obtained mRNA expression data, miRNA expression data, and annotation files.

mRNA expression values
Consisting of 16,126 genes and 10,682 samples, all from different donors.

Sample annotation
Containing information about the samples. Most notably:

- *Cancer type*
- *Primary site*
- *Tissue type* - Normal or Tumor

The distribution of samples over these annotations are visualized in Figure 5.

miRNA expression values
Consists of 2,450 miRNAs and 10,676 samples, matching the ones in the mRNA expression data.

miRNA annotation
Containing information about the miRNAs. Most notably:

- *Name*
- *Target* - RNA base sequence
- *Tissue type* - Normal or Tumor

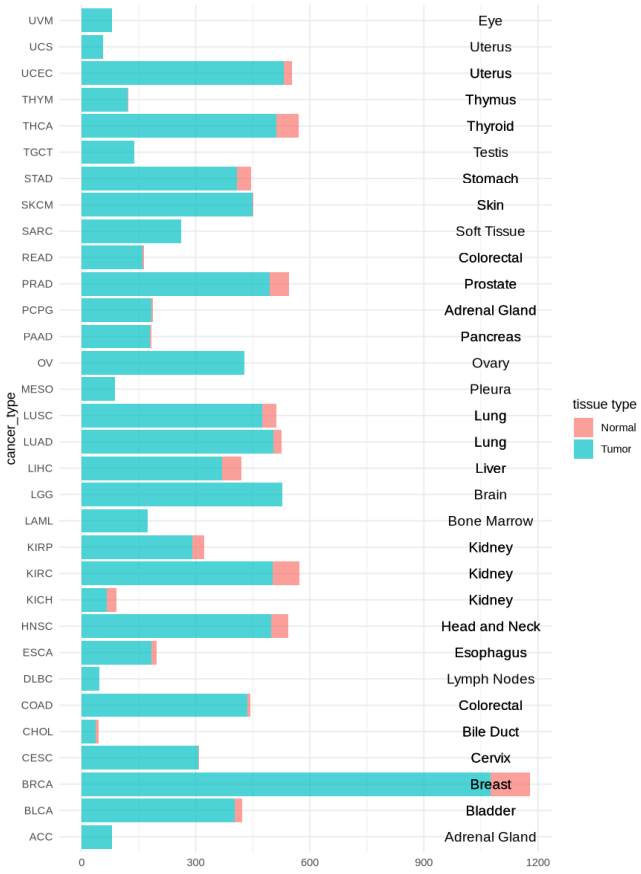


Figure 5: Bar plot visualizing the counts of different cancer types in the sample annotation data, categorized by their primary sites, which are labeled on the right side of the plot.

3.2 Generating Control Samples Using DGD

To generate normal control samples for the TCGA datset, we use a pre-trained DGD model by Prada-Luengo et al., as described in section 2.5. The model is implemented in the Python package bulkDGD, where pre-trained model parameters and optimization schemes for finding best representations are available (see section 8 for available code and resources).

The pre-trained DGD model is trained on bulk expression profiling data consisting of counts; hence, control samples are generated from the count expression matrix before any normalization or data transformation. However, due to the large dataset size, we split the dataset into smaller sample sizes of about 200. The smaller samples are then fed through the model, and results are re-combined afterwards.

Controls are found by using the pre-trained DGD model to predict new data points for each sample. To do so the model parameters are fixed (decoder and GMM), and a control is found by optimizing the representation in latent space for the sample.

The scheme used to optimize the representation consists of two steps. In the first round, one representation per component for each sample is initialized, i.e. 45 representations are initialized from the component means for each sample [20]. These 45 representations are each optimized for 10 epochs, and afterwards, the best representation with the lowest reconstructions loss for the sample is selected [18]. Overall, this means the representation for a sample is initialized by assigning the optimal component to it. After finding an optimal initialization, the second round of the optimization scheme entails further optimizing the representation for 50 epochs[18].

After running the optimization, the representation found for a sample is passed though the decoder to generate the distribution parameters for the sample ($f_\theta(z)$), this being the means for the negative binomial distributions over expression counts for each gene. The means generated from a sample are then scaled by the average number of counts across the sample, resulting in the *generated controls*.

3.3 Data Pre-processing

3.3.1 Filtering

To ensure a robust comparison between miReact utilizing the global median and miReact incorporating DGD, mRNAs that are not shared between the two models are excluded. This results in the removal of 3,658 genes, from our dataset, initially comprised of 19,784 genes, leaving us with 16,126.

3.3.2 Normalization

To ensure consistency and comparability between the two methods, data is normalized in the same way as done by Nielsen and Pedersen. First, we normalize the mRNA expression count data by the total expression count of each respective sample, and then scale by the median of the library size, resulting in the relative expression levels:

$$S_i = \sum_{i=1}^{m} x_{ij}$$

$$x'_{ij} = \frac{x_{ij}}{S_i} \cdot \text{median}\,(S_1, S_2, ..., S_n)$$

Where x_{ij} denotes the mRNA expression count of gene j for sample i, m is the number of genes and n the number of samples. To complete the normalization step we transform the data by $\log_2(x'_{ij} + 1)$ to compress the range and stabilize variance. The pseudo-count of 1 is to avoid taking the log of zero.

It should be noted that when we normalize the rescaled DGD outputs (generated controls), we scale by the library size of the raw counts and rather than the library size of the generated controls themselves. This ensures comparability between the two methods, as the generated controls may have a different library size than the original counts.

3.4 Computing Fold Changes

The count expression matrix and the re-scaled DGD output (generated controls) are normalized and log transformed as described in 3.3.

Fold changes are computed from the generated controls by subtracting the generated values for each gene i for each sample j from the corresponding gene and sample of the expression matrix, i.e.

$$\Delta e_{i,j} = e_{i,j} - e_{i,j}^{DGD} \tag{2}$$

where e denotes the normalized expression counts, and e^{DGD} is the expression values of the generated controls.

For our baseline method, the fold changes are computed by the use of the global median as described in section 2.3.1.

3.5 Inferring miRNA Activity

To infer miRNA activity, the same approach as Nielsen and Pedersen has been used and as described in 2.3.1. See section 8 for miReact resources.

3'UTR sequences
3'UTR sequences are obtained from GENCODE release 33. For each gene the longest 3'UTR is chosen, and 3'UTRs with lengths less than 20 bases or more than 10.000 are removed, as done per Nielsen and Pedersen.

Probability of observing motifs in sequences
Regmex is used for computing sequence specific p-values for observing miRNA target sites within the 3'UTR sequences, as described in 2.3.1.

Calculating miRNA activity
miReact is used to infer miRNA activity scores as per 2.3.1. For the new approach, the sequence specific p-values will be ranked by the fold changes computed using the generated controls and not by the global median. Otherwise, the modified rank sum test is and activity score computation is performed as previously described in 2.3.1.

For our baseline method, activity scores are computed using the global median across all genes, with no alterations to the test setup.

3.6 Performance Evaluation

As a benchmark of our performance, we compute the correlation between the observed miRNA expression data and the inferred activity. The correlation is computed for each miRNA across the approximately 10,000 samples of which we have miRNA data available. This results in 2,450 correlation coefficients. As an overall measure, we use the average correlation across the 2,450 miRNA correlation coefficients. The average correlation between inferred activity and observed expression is computed for both methods of controls.

As a correlation measure, we use both Pearson and Spearman respectively. It should be noted that we do not necessarily expect a linear relationship between expression and activity, as activity is not a direct measure of expression level but captures motif enrichment in strongly up or downregulated mRNA, whereby we might expect a monotonic relationship between expression and activity.

3.7 Extracting Gaussian Mixture Components

From the DGD model, we extract the density scores per GMM component per sample. This score is a measure of how each representation fits under each GMM component, scaled by its corresponding mixing coefficient. By selecting the maximum score across all components for each sample, we determine which cluster, or component, the sample is most likely mapped to in the bulkDGD mode.

To get the tissues associated with these components we obtain the original found during training from Prada-Luengo et al.. They had been assigned to the most predominant sample site in the clusters of the original GTEX training data. This predominant tissue type was then assigned to the respective component as its representative tissue. They were then used to analyses regarding the mapping of the samples: see *DGD performance*.

4. Results

4.1 Comparative performance evaluation

We observe a strong correlation between inferred activity and expression levels for many miRNAs for our miReact model integrated with DGD. However, when evaluating the method by the average correlation across all miRNAs, we do not see a significant improvement compared to the global median based approach.

Using the quantitative benchmark as described in 3.6, we obtain an average Pearson correlation of 0.0083 across all 2450 miRNAs using generated controls and an average of 0.0081 using global median as a control substitute. Using Spearman, we obtain an average correlation of 0.0046 and 0.0047 using generated controls and global median respectively. The distributions of correlation coefficients can be seen on Figure 6.

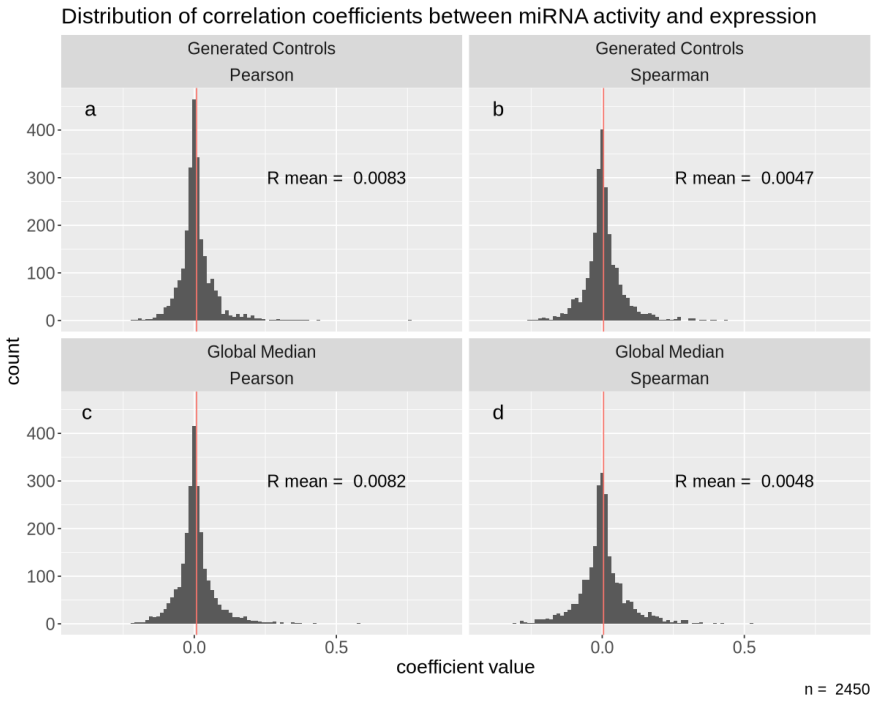


Figure 6: Distributions of correlation values between inferred activity and observed miRNA expression. **a.** distribution of Pearson correlation values between expression and activity inferred activity using generated controls. **b.** Distribution of Spearman correlation for generated controls. **c.** Distribution of Pearson correlations for global median control **d.** Distribution of Spearman correlations for global median control.

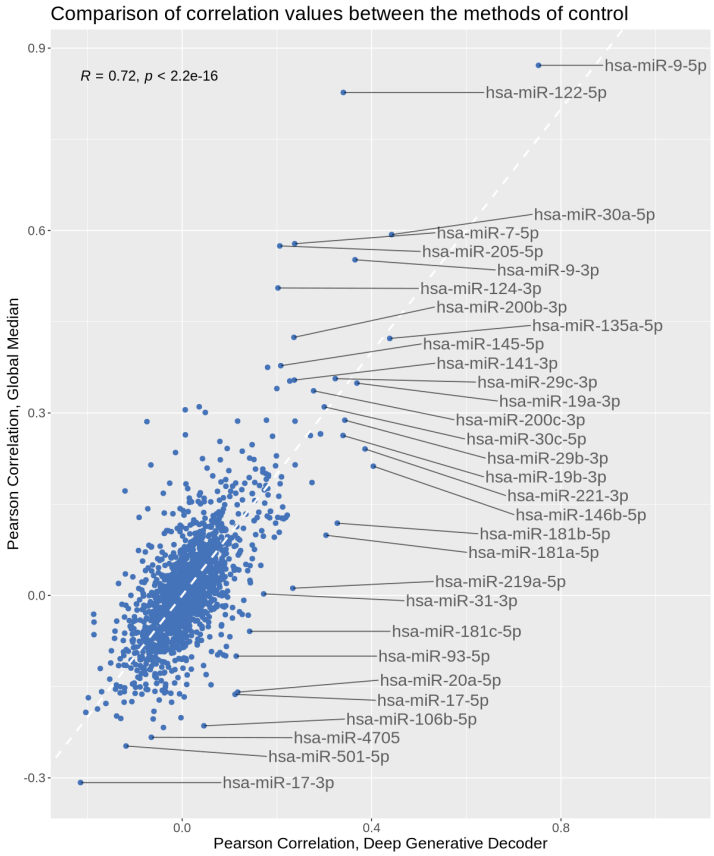


Figure 7: Comparison of Pearson correlations between both methods of inferring miRNA activity. x-axis is correlations of miRNA activity for generated controls, and y-axis is activity obtained using global median as control. miRNA with high correlations and outliers are marked.

As illustrated in Figure 6, the average correlation between activity and expression is slightly positive for both control methods.

The low average correlation for both methods is due to the reasons as described in 3.6. We do not expect activity to show a strong relationship between activity scores and expression, as activity is not a direct prediction of expression count. Additionally, for some miRNAs, we do not expect any significant correlation at all. As mentioned in 2.1, miRNA are separated into two strands creating a 5p and a 3p variant, where one strand (the "guide" strand) is functional and leads to mRNA degradation, whilst the other strand (the "passenger" strand) is typically degraded and does not play a significant role in gene regulation. In some cases, however, both strands are functional and partake in gene regulation. Which strand that will acts as a guide strand can vary [10]. An underlying assumption within the miRNA activity score is that miRNAs will result in degradation of their target mRNAs. Consequently, non-functional miRNA strands that do not interact with their assumed targets will result in low correlation between activity and expression for a large number of miRNAs. Therefore, a low average correlation is expected.

4.1.1 Evaluation of miRNA activity inference at the single miRNA level

To asses whether the methods perform equally on the same sets of miRNA or are better at inferring activity on different groups of miRNAs, we compare correlation coefficients of indi-

vidual miRNAs against each other. As described in 3.6, we compute the correlation between inferred activity and observed expression across all samples for each miRNA. This is done for both methods and the results are then compared against each other.

Overall, a correlation is observed between the correlation scores of both methods, where the highest performing miRNAs are often shared between both methods, as illustrated on figure 7. Notably, the brain and nervous system specific miR-9 is the highest ranking for both methods, and the liver-specific miR-122 achieve a strong correlation for both methods when assessed using Pearson correlation. When comparing Spearman correlations between the two methods, a similar trend is observed (a larger version of Figure **??** also containing a comparison of Spearman correlations can be found in the appendix A). The highest ranking miRNA based on Spearman correlation are members of the miR-200 family; a family that controls the expression of many genes that play important roles in cancer cells [3]. Similarly, members of the let-7 family, functioning as tumor suppressors in cancer by targeting specific genes, exhibit high correlations.

It is generally observed that miRNA of the same family often have similar Spearman correlation values, which is not unexpected as miRNA families can often bind to similar or overlapping sets of target mRNAs, due to miRNA within a family often sharing similar sequences and binding regions.

4.2 Case Studies

To delve deeper into the performance and behavior of the inferred activity based on using generated controls, we examine specific cases of miRNAs. Particularly, the tissue-specific miR-122-5p, which is also a case study used in the paper on *miReact* by Nielsen and Pedersen. Additionally, we investigate let-7a, a tumor biomarker.

4.2.1 Inferred activity for tissue specific miRNAs

As previously mentioned, miR-122 is an established liver specific miRNA. The inferred activity scores when using miReact with the global median based approach accurately reflects miR-122 heightened levels of expression in the activity score as illustrated on figure 8. The activity is highly up-regulated in samples drawn from liver hepatocellular carcinoma (LIHC) tissue as well as samples from cholangiocarcinoma (CHOL)(type of bile duct cancer), while other tissue types do not show much activity of significance. Using generated controls, we also observe higher activity scores in LIHC but not of the same level, and no significant activity scores in CHOL samples. Particularly, normal samples (samples from healthy patients originating from the same tissue-type as the tumor samples) score very differently between the two methods. As can be seen in Figure 8, normal samples obtain activity scores close to 0 for both LIHC and CHOL samples, while both the observed expression and global median activity indicate higher levels for normal samples.

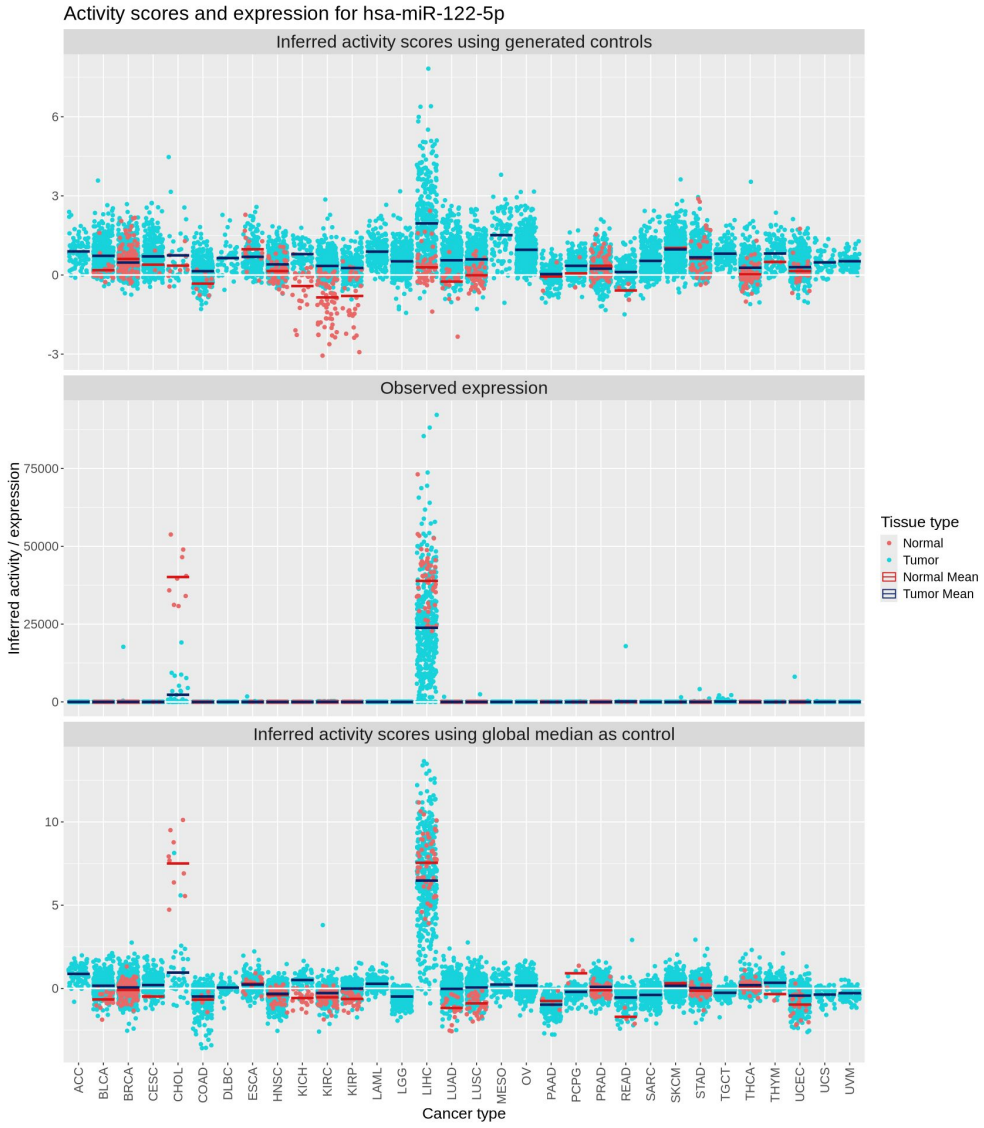


Figure 8: Barplot of activity or expression across all cancer types within the TCGA dataset. Each point represents the activity score of miR-122-5p for a sample. Red points indicate normal samples, while blue represents cancer samples. Line segments is the average activity of cancer samples (blue) or normal samples (red) within the cancer type. The first panel shows the activity from generated controls, the second panel shows the observed expression, and the third panel shows the global median used as a control.

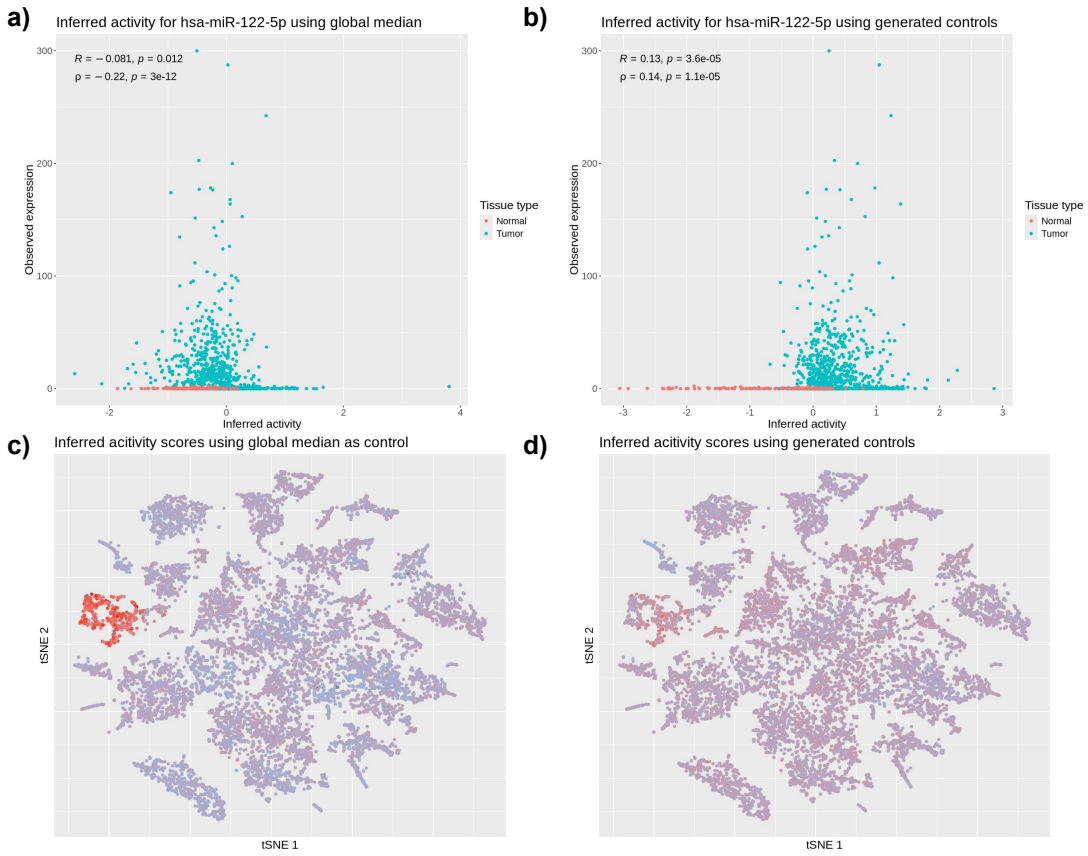


Figure 9: **a)** and **b)** are scatter plots of observed miR-122-5p expression for each sample, plotted against the two methods of control: global median and generated controls respectively, in kidney tissue. **c)** and **d)** show the activity scores of miR-122-5p overlaid the t-sne plot for the global median and generated controls methods, respectively. Each point represents a sample; red indicates high activity, while blue indicates low activity. Note the red area in figure **c)** generally corresponds to liver samples.

For multiple miRNAs a consistent pattern is observed: normal samples are often assigned activity scores around zero, while this is not he case for the observed expression nor global median based inferred activity. This behavior appears to stem from the fact that activity scores computed using generated controls do not reflect a sample's miRNA activity relative to all other samples across various tissue types. Instead, they primarily indicate a sample's activity relative to its healthy state. In this context, activity scores of zero indicate that no dysregulation of miRNA activity take place, while non-zero activity scores indicate aberrant activity.

When using the global median as a substitute control to compute gene relative expression, it reflects whether a gene is up or downregulated compared to all other samples across many different tissue types. Hence the relative expression captures not only cancer-induced up or downregulation, but will also reflect if a gene is relatively highly expressed due to being tissue-specific. In miReact, this will shift the ranking of the gene (3'UTR) sequences, and clustering of motifs in sets of tissue-specific genes leads to high activity scores for those motifs. In summary, sample activity scores strongly reflect inter-sample differences.

In contrast, when using generated controls, relative expression serves as a metric for how gene expression has deviated compared to a healthy sample counterpart, focusing solely on expression alterations due to cancer. In the miReact test setup, genes are subsequently ranked based on the changes induced by cancer in their expression levels. Consequently, normal samples are expected to exhibit no cancer-induced gene up or downregulation, resulting in a seemingly

arbitrary ranking of genes and hence no clustering of motifs. Therefore, miRNA in normal samples are expected to obtain activity scores close to zero. The activity score obtained by using generated controls minimally reflects tissue-specificity of miRNA. As illustrated by Figure 9.c, activity inferred by use of the global median is strongly upregulated in the cluster of liver samples. The same tendency is not seen using generated controls in Figure 9.d, where we only observe slight activity regulation throughout the plot across multiple samples (can be seen by many points being slightly more red overall).

While activity scores obtained from the DGD based approach is not a direct improvement in estimation of miRNA expression levels, they often lead to an improved distinction of activity between tumor and normal samples. As illustrated in Figure 8, generated controls yield normal samples with activity scores around 0, while tumor samples are estimated to have upregulated activity. The difference in mean activity between normal samples and tumor samples is more pronounced than when using the global median. The distinction between normal and tumor is particularly evident when examining miR-122-5p expression in kidney tissue. As illustrated in Figure 9.a, activity based on the global median fails to clearly distinguish between tumor and normal samples, with many tumor samples assigned low activity scores regardless of their observed expression levels. However, as illustrated in Figure 9.b, generated controls allow for a clear distinction of normal and tumor sample activity. These findings align with existing literature, which has demonstrated the potential of miR-122-5p as a biomarker for certain types of kidney cancer, where elevated levels of miR-122-5p is indicative of advanced cancer and reduced survival time[4].

4.2.2 Inferred activity for tumor biomarkers

As mentioned in section 2.2, the let-7 family regulates oncogenes prominently expressed in tumor samples. Consequently, abnormal levels of miRNAs from the let-7 family can serve as biomarkers for cancer. Specifically, let-7a has been found to be downregulated in a number of different tumor samples.[16] Another known tumor suppressor is miR-200.[2]

let-7's role as a tumor suppressor is evident in our inferred activity levels using DGD for controls, as seen in Figure 10.a, where we examine the ability to distinguish between tumor and normal activity. For the typical guide strand, let-7a-5p, the activity is noticeably downregulated, for tumor samples in comparison to normal samples. The inferred activity scores using generated controls are predominantly downregulated, whilst the median of normal activity centered around 0 (indicating no abnormal activity). This represents a significant improvement compared to the inferred activity scores using global median as controls, where tumor samples still exhibit some significant values but not with any clear direction. This behaviour is also reflected in other tumor suppressors like the typical guide strand miR-200c-3p, that also has a slight negative drift in comparison to its normal samples.

Given the fundamental role tumor suppressors play in cellular regulation, it's expected that when analyzing miRNA tumor biomarkers, their expression levels will typically be high across all or multiple tissue types. However, tumor suppressor expression levels are often downregulated in cancer.

Examining let-7a-5p further, we observe high expression counts across various cancer types, as shown on Figure 10.b (mid panel). Notably, the expression levels in normal samples are often slightly higher than in tumor samples, as expected.

Using the global median approach, let-7a activity varies significantly depending on the cancer type (Figure 10.b, bottom panel), and see both cases of strong up or downregulation. However, we do observe higher mean activity in normal samples compared to tumors. Consequently, in this setting we can only detect aberrant miRNA activity if we have activity scores for normal samples available for comparison. As previously mentioned, activity inferred by the global median approach also captures tissue- and sample-specific activity, and not just cancer-related activity, making it difficult to distinguish between these factors and their contributions to activity.

In contrast, with generated controls (Figure 10.b, top panel) let-7a activity is consistently negative across all tumor samples regardless of cancer type, while normal samples center around activity levels of zero. This allows us to determine the direction of activity even without normal samples for comparison, as the activity score reflects only cancer-related changes.

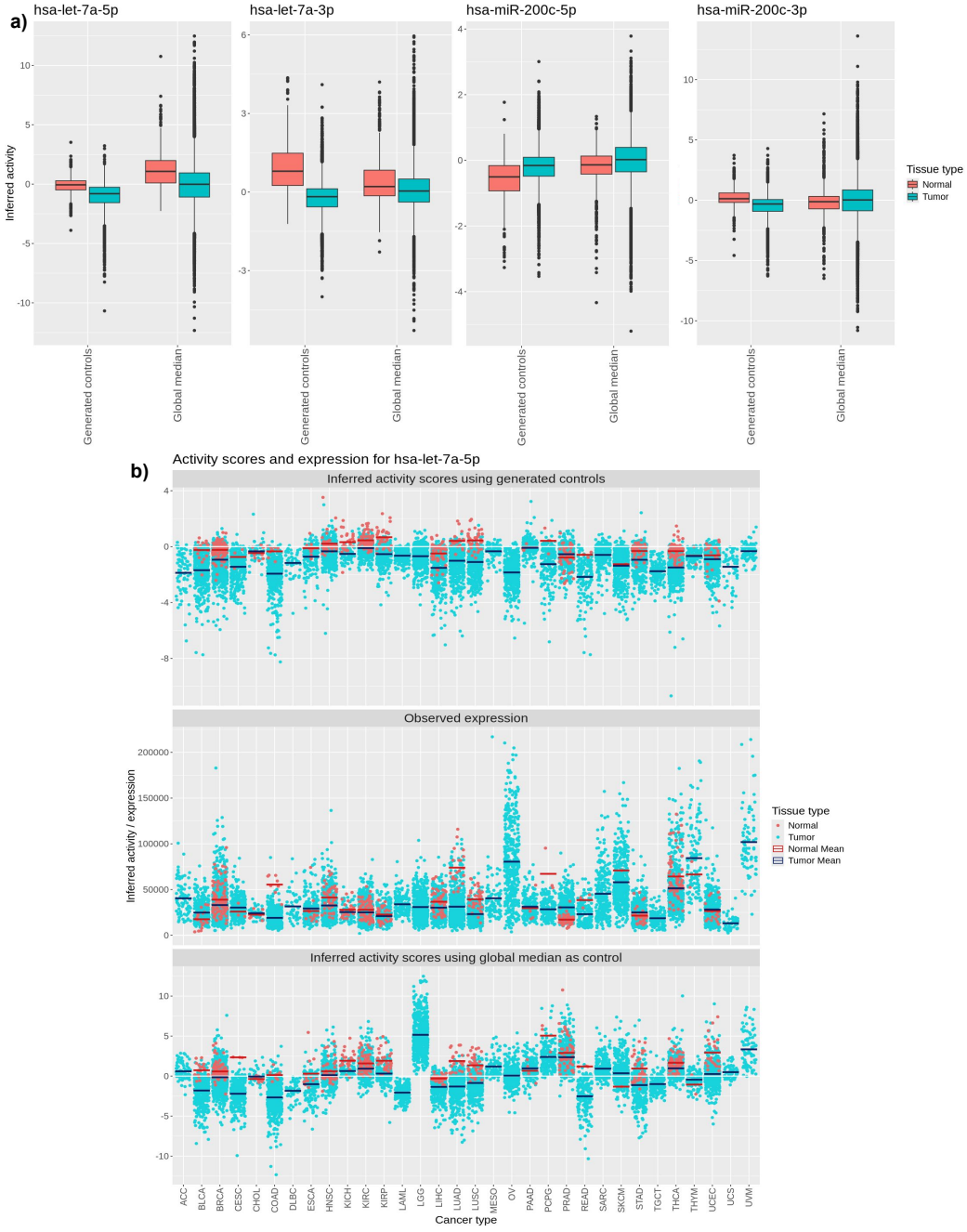


Figure 10: Overview of the activity of tumoral supressors let-7a and miR-200 inferred by both methods of controls. **a)** Box plots illustrating the spread of normal vs. tumor samples for act method for both passenger and guide strand. **b)** Barplot of activity and expression across all cancer types within the TCGA dataset. Each point represents the activity score of let-7a-5p for a sample. Red points indicate normal samples, while blue represent cancer samples. Line segments is the average activity of cancer samples (blue) or normal samples (red) within the cancer type. The first panel shows activity from generated controls, the second shows observed expression, and the third shows activity using the global median as a control.

In summary, activity scores inferred using miReact with DGD strongly reflect cancer-specific activity, minimizing the tissue-specific influences typically seen in scores computed by miReact. Using generated controls, we generally achieve a stronger distinction between activity score values for normal and tumor samples, potentially enabling easier analysis of miRNA behaviour in cancer. miRNA activity in normal samples centers around zero, while non-zero scores seem to indicate aberrant miRNA activity. This distinction is particularly evident for miRNAs that serve as cancer biomarkers.

4.3 Distribution of miRNA activity scores

Previous findings reveal that activity scores from the two methods reflect distinct biological signals and behaviors. Specifically, miRNA activity scores in normal samples exhibit notable differences in value ranges between the two methods. Additionally, many cancer biomarker miRNAs demonstrate a consistent direction of activity regulation across all samples with the DGD-based approach, whereas this pattern is not observed using the global median-based approach.

To better understand the differences in miRNA activity scores between the two methods, we analyze their distributions. We are particularly interested in whether cancer biomarkers generally display predominantly positive or negative activity scores across all samples, or conversely, if miRNAs with highly distinct distributions could potentially be biomarkers.

4.3.1 Average activity scores of miRNA

The distribution of activity scores differ strongly between the two methods, as illustrated in Figure 11.a. When miRNA activity scores are computed using the global median, the average activity scores for miRNA center around zero, with minimal instances of values that are marginally above or below zero. However, the scores obtained using generated controls show large variability around zero, extending across a broad range of considerably high to low activity values. Overall, we observe that miRNAs with a high activity score tend to be strongly upregulated globally across all samples, and vice versa for miRNAs with low activity. Computing activity by the traditional method does not exhibit the same distinct tendency for a miRNA to be globally up or downregulated.

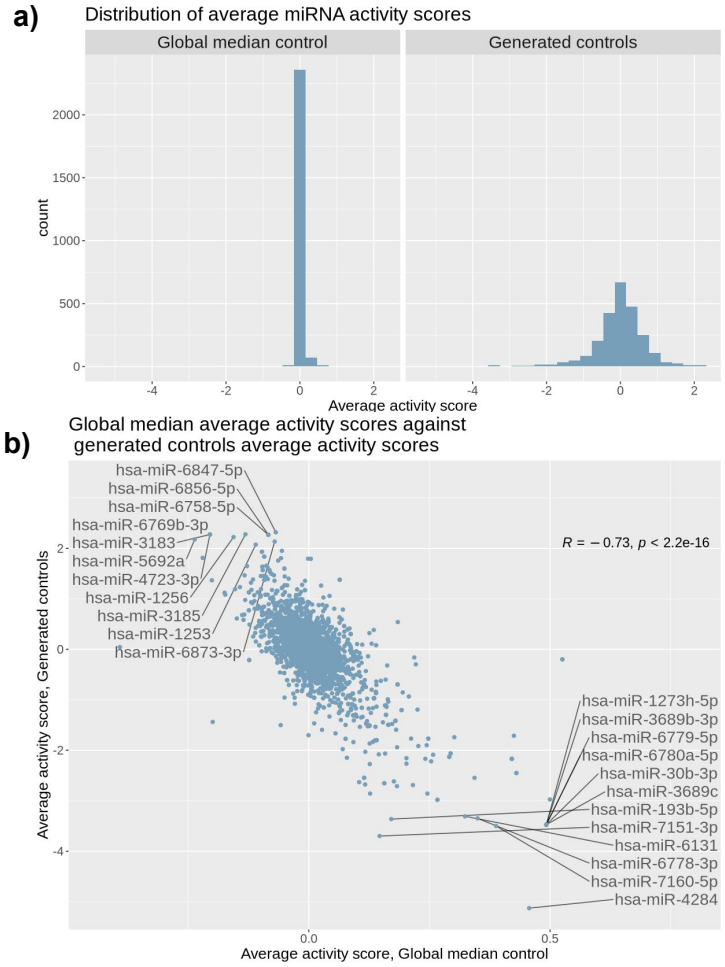


Figure 11: a) The distribution of average miRNA activity score for both methods. **b)** Comparison of miRNA activity scores by the global median method against DGD method. Each point represents a miRNA.

miRNAs show less differentiation in average activity across samples when computed using the global median. This may be due to the signal of cancer-related activity being minor compared to tissue-specific or sample-specific activity, or other conditions that dominate activity scores of a sample. As activity can depend on various biological signals unique to each sample, there is large variability in activity scores, averaging to approximately zero across all samples for a given miRNA. Hence we obtain a very narrow distribution of activity scores.

In the case of the DGD based approach, where other sources of variability except cancer are removed, we see a broader distribution of average activity scores. The average activity of a given miRNA could be indicative of its general activity levels in cancer tissue (note the majority of samples are tumor samples and so the average activity across all samples mainly reflects tumorous activity). As previously described, activity scores computed using generated controls will generally assign normal samples an activity of zero, while non-zero activity scores should indicate abnormal miRNA activity related to cancer. Hence miRNA that are globally up or downregulated may signify a given miRNAs overall role in cancer either as a tumor suppressor or as an oncogenic miRNA.

This behaviour is illustrated in Figure 12.a for the previously described miRNAs miR-122-5p, miR-200c-5p, and let-7a-5p. The densities show the distribution of the miRNAs activity

scores across all normal and tumor samples respectively. As seen on the figure, the distributions of activity scores (generated controls) in tumor samples are slightly skewed for the aforementioned miRNAs. This is consistent with the literature; as described in section 2.2, let-7a (red distribution in the figure) is an oncogenic miRNA, and is its levels are known to be downregulated in numerous cancer types including lung, breast, colorectal, pancreatic, prostate and kidney, among others[19]. Many of these are present in the TCGA dataset, and we do observe that let-7a activity is negatively skewed. Oppositely, we observe miR-122-5p (yellow) to be positively skewed, though miR-122-5p can act as either a tumor-suppressor or tumor promoter. However, in the case of kidney cancer, we observed upregulated expression levels in section 4.2.1. miR-200c-5p is very slightly negatively skewed and known to be commonly downregulated, but its expression can be context-dependent.

The distributions of the normal samples for miR-122-5p, miR-200c-5p, and let-7a-5p do not show the same skewness, but are generally symmetrical around zero activity, corresponding to what was expected.

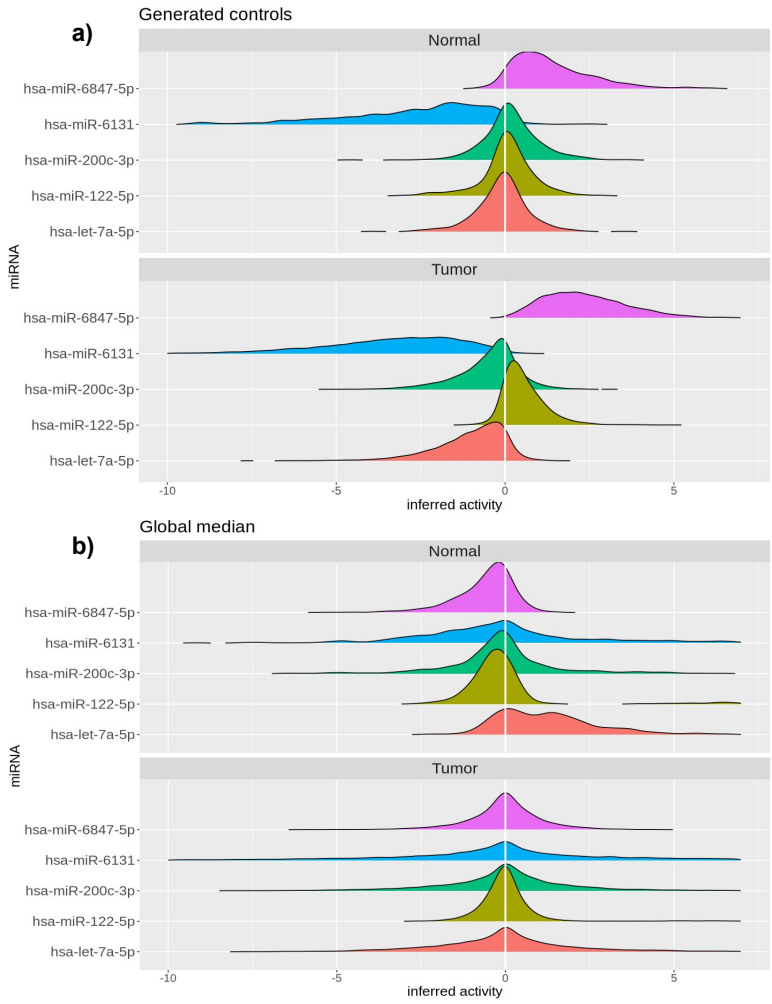


Figure 12: a) Distribution of activity scores from DGD method for a selected set of miRNA. Upper panel is distribution of scores for normal samples, lower is tumor samples. **b)** same as d, with activity scores from global median method.

The activity score distributions obtained by using the global median as control are depicted in Figure 12.b. As can be seen on the distributions for tumor samples, the distributions are centered strongly around zero with varying variance. Hence, the distributions do not signal

the activity scores having a cancer-specific tendency, but reflects varying factors across all samples. Moreover, due to the majority of samples being from tumor tissue, the global median used as control mostly represents an average tumor sample. The effect of this is particularly evident in the distribution of normal samples in Figure 12.b. The relative expression will reflect how much the gene expression level deviates from the average tumor samples, which can cause normal samples to obtain a ranking of genes as if some expression levels are abnormally up or downregulated. If motifs then cluster in sets of genes that diverge from the average tumor expression, it can result in high or low activity scores that don't reflect actual abnormal activity.

When comparing the average activity scores for miRNA produced by using the global median as control against using generated controls, we observe a negative correlation as illustrated in Figure 11.b. Specifically, we observe that miRNAs that score a positive high average activity by the global median method, generally will obtain a negative average activity when using generated controls. This is particularly evident for miRNA with outlying activity scores, which are marked by name in Figure 11.b. The negative correlation is likely a product of the mentioned difference between the two methods; when using the global median as control we are essentially comparing expression levels against an overall tumor sample, whilst the generated controls enable comparison to healthy samples. This may cause a shift in the ranking of genes, such that a set of genes that may be upregulated compared to the average tumor sample may become downregulated when comparing to a healthy sample. If a given motif or miRNA binding site clusters within this set of genes, the direction of the activity score will change from negative to positive. This also applies in the opposite scenario, where a set of genes goes from being downregulated to upregulated between the two methods of control, and hence activity scores become negatively correlated.

4.3.2 miRNA activity scores reflecting motifs of interest

Many of the outlying miRNAs with very high or low average activity scores exhibit a clear positive or negative skewness of their activity score distributions. This is illustrated for the miRNAs miR-6847-5p (pink) and hsa-miR-6131 (blue) in Figure 12.a. Both miRNAs are marked on the figure 11.b, where miR-6847-5p scores a high average activity whilst hsa-miR-6131 scores among the miRNAs of lowest average activity. However, many of the outlying miRNAs highlighted in Figure 11.b obtain activity scores that do not reflect their observed expression. Regardless of whether their average activity is predicted to be high or low, the miRNAs often are *very* lowly expressed for both normal and tumor samples across all tissue types, indicating these miRNAs may have low functionality. An example is illustrated in Figure 13, which depicts miR-6847-5p activity across all cancer types. The observed expression is minimal for all cancer types, with some having practically no expression. However, the activity computed using generated controls systematically assigns a positive activity score to all samples, many of which obtain a significant score. This behaviour is especially prevalent amongst miRNAs with high average scores.

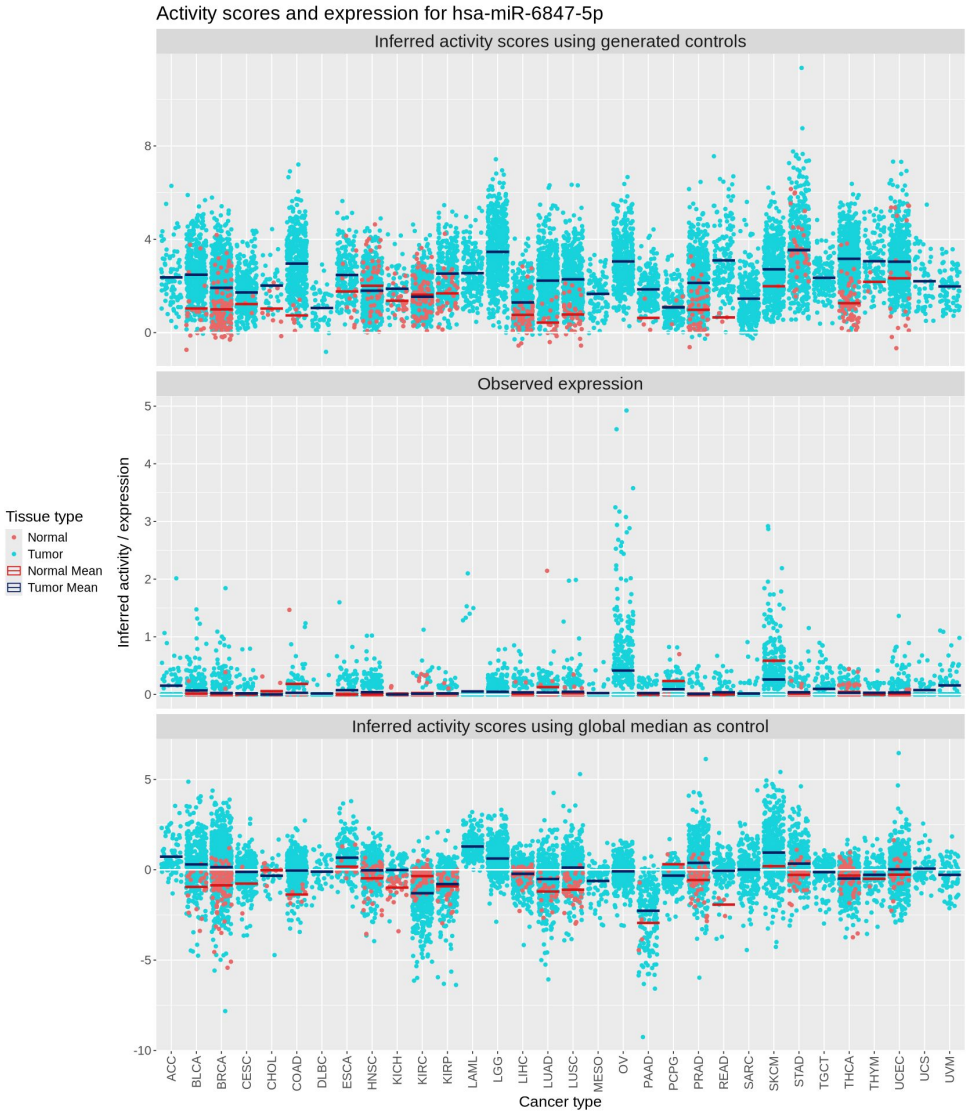


Figure 13: Barplot of activity and expression values for miR-6847-5p across cancer types. Upper panel is scores from the DGD method, mid is observed expression levels and last is scores of the global median method.

The outlying miRNAs are often those with a high numerical designation, which reflects the chronology of their discovery. These miRNAs are often discovered later due to not being very abundantly expressed or only expressed in specific conditions or tissues, making their discovery and characterization challenging. Hence, the high activity scores of these miRNAs may not be a reflection of their expression and actual activity, but may relate to the presence of certain motifs. As previously described, non-functional miRNAs will show low correlation between activity scores and expression, and the presence of their target sites on up or downregulated genes may not relate to miRNA-mRNA interaction. As illustrated in Figure 12.a, computing activity scores by generated control results in a strong distinction and aberrant activity scores for these miRNAs compared to using the global median, as depicted in Figure 12.b. This distinction made by generated controls may indicate that the underlying motifs of the miRNAs relate to certain motifs or sequences being enriched in cancer, where the indicated enrichment is not related to miRNA gene regulation. Further study of the enriched genes would be necessary to determine what regulatory mechanism or factors are involved, and their relation to cancer.

Ultimately, by analyzing the distributions of activity scores, we observe that many cancer-related miRNAs demonstrate a clear skewness in tumor samples, indicating a distinct behavioral pattern in cancer. In contrast, normal samples tend to center around zero, supporting the hypothesis that significant non-zero activity scores indicate aberrant miRNA activity. This potentially allows for predicting cancer-related miRNAs without needing normal controls for comparison. Comparatively, significant activity levels cannot be expected to reflect cancer-related behaviour in the traditional miReact setting. Lastly, integrating miReact with DGD may potentially enable not only the capturing of motifs related to miRNA-mRNA interaction, but also other motifs of interest relating to cancer.

4.4 DGD performance

To ensure a comprehensive evaluation of miReact's performance using generated controls, it is essential to also examine the performance of the DGD model itself. This examination will enhance our understanding of the robustness and reliability of our results.

4.4.1 Exploring representations

To evaluate how bulkDGD performs on the TCGA data, we begin by assessing the first goal of DGD, which is to compute low-dimensional representations of the samples.
The representations play a vital role as inputs in the decoder part of the DGD model. Therefore, exploring whether they retain sample information is an essential step in analysing the performance of the bulkDGD model. Since we have an increased number of components in comparison to primary sites, it might also be able to capture smaller subgroups of tissues, that have different characteristics.
The representations ability to capture, or retain, tissue information is visualized by creating a tSNE projection of the representations, as illustrated in Figure 14. The representations seem to cluster well in regards to both their primary site, as shown in Figure 14.a, and the models components, as shown in Figure 14.b. The clear grouping indicates that bulkDGD successfully retains the underlying biological structure of the data in the representations.
In Figure 14.b we also see how some tissues are divided between multiple components. This is clearly shown in cluster containing the samples originating from the brain (circled with red in Figure 14.b). This makes sense as 7 of our 45 components have 'Brain' as their associated GMM component.

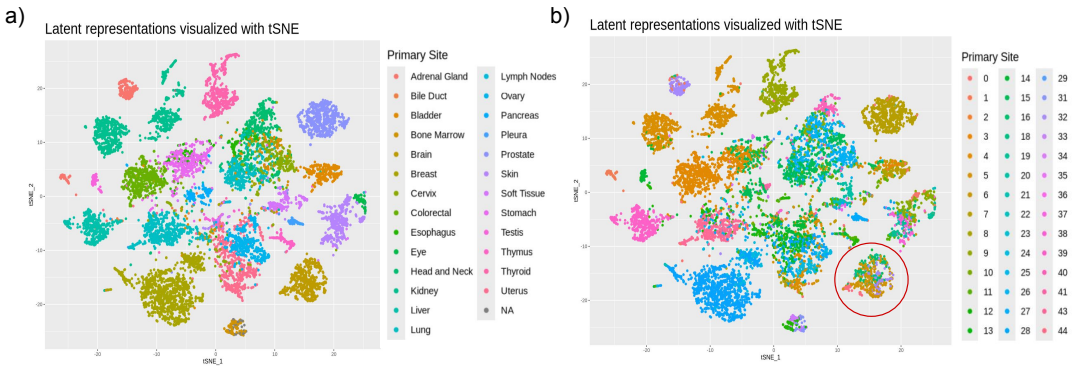


Figure 14: Latent representations of TCGA. **a** Visualized using tSNE colored by the samples primary site. **b** Colored by their GMM components representative tissue.

4.4.2 Exploring closest-normal comparison sets for TCGA samples

To delve deeper into the performance of bulkDGD, we assessed its ability to generate useful representations for cancer samples, by looking at whether bulkDGD matches our TCGA tumor samples to the same GMM component as their normal GTEx counterpart, see section 3.7. This is illustrated in the association matrix in Figure 15.a showing the percentage of TCGA samples assigned to the GMM components.

Since we used a pretrained model, we had a mismatch of primary sites. The bulkDGD model had been trained on a subset of the primary sites from GTEx. This subset contained 15 of the TCGA primary sites along with additional types. Consequently, only 15 out of 27 tissues in our analysis had components trained to capture them. This is illustrated in the association matrix in Figure 15.a, where a substantial part of the columns are almost empty, indicating that they do not capture any of the samples.

Most of our comparable tissues were represented by a single component; however, some were split across multiple components. To enhance the clarity and interpretability of the association matrix, we combined these components. The resulting, more refined association matrix is displayed in Figure 15.b.

From the cleaned association matrix we see that 9 out of 26 TCGA primary sites had more than 70% of the samples assigned to a single GMM representative tissue (the dark blue). 7 of these 9 were assigned to the matching component, while 89.7% of 'Pleuara' were assigned to 'Adipose' and 73.75% of 'Eyes' samples were assigned to 'Skin'.

The lack of directly comparable GMM components resulted in components capturing a diverse array of tissues. For example, the 'Stomach' component, marked in red in Figure 15.b, was the assigned component for significant portions of multiple tissue types. This phenomenon is further highlighted in a tSNE plot of mRNA expression values (Figure 15.c), where the samples assigned to the component for stomach is highlighted.

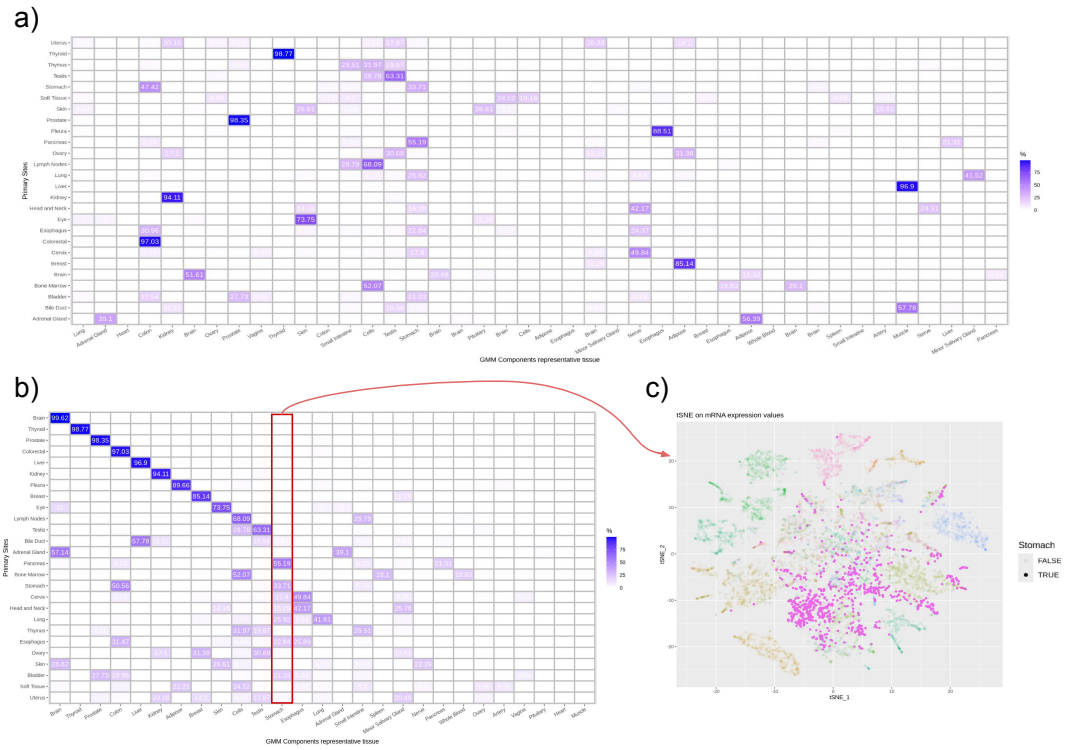


Figure 15: Closest normal representation for cancer samples. **a)** Matrix showing the percentage of TCGA tumor samples(x-axis) assigned to GMM components(y-axis) named after their majority normal primary site. This illustrates that only a few of our TCGA tissues are predominantly described by a single component (dark blue cells).**b)** Replacing the y-axis with the combined components that share the same majority type. **c)** tSNE plot of mRNA expressions colored by primary site with the samples associated with the components for liver - marked with red in B - that multiple tissues types are associated with. An enlarged version of the figure can be found in Appendix.C

Components that contribute to multiple tissues often exhibit a biological connection with those tissues. An example of this is the 'Liver' component that catches 57.8 % of all 'Bile Duct' samples. Since the tissues are connected in the body, it makes sense that the samples would have a the closest resemblance to 'Liver' when lacking their own specific component.
However, there are exceptions, such as the 'Stomach' component, which catches a notable percentage of 'Pancreas' and 'Esophagus' samples that have some biological connection to it. But, it also catches 'Bladder', 'Lung', 'Cervix', and 'Head and Neck', which are not as biologically similar.

Since the Decoder is trained alongside the GMM components this frequent misclassification of tissue types could lead to the emergence of atypical pseudo-normal samples, which do not accurately reflect the true biological nature of the tissue.

It is worth noting that some of the samples are of metastatic cancer. This means it is cancer that has spread from another part of the body, hence it might be more similar to the tissue from where the cancer originated. This would certainly make up for some of the missassigned samples, but not all, since only 3.6% of our samples are classified as metastatic. None of the samples assigned to the 'Stomach' component are metastatic despite it capturing a large fraction of multiple different tissues (see Figure 15.b and c).

5. Discussion

5.1 Potential of miRNA activity scores as indicators of cancer-related activity changes

Our aim was to integrate a deep generative decoder (DGD) to enhance miReact's performance in inferring miRNA expression levels. Although the integrated model did not significantly improve the prediction of expression levels, it shows promise for predicting aberrant miRNA activity. By using pseudo-normals generated by the DGD model for relative expression comparison, we effectively reframe the miReact setting. As described in section 4.2.2, this approach predominantly reflects cancer-related activity, eliminating other tissue, cell, or sample-specific activity.

While we observed correlations with the observed miRNA expression levels, the new setting may affect comparability, as scores now reflect a *relative* activity level rather than general miRNA activity.

Activity scores measure changes in mRNA activity, with high scores indicating upregulation and low scores indicating downregulation in a given sample. Zero activity scores suggest normal miRNA activity in the sample.

To evaluate the integrated model's performance as a measure of relative activity level, activity scores could be compared against the observed relative expression of miRNA.

5.2 Model improvements

5.2.1 Enhanced performance of miReact

The incorporation of bulkDGD has significantly improved aspects of the miReact model by addressing some of its previous limitations regarding sample variability and data skewness. miReact is dependent upon a varied sample input to ensure robust calculations, as it uses the median gene expression level across all samples, to calculate the fold change. Consequently, the used fold changes are dependent on the other samples in the input dataframe. Previously, the addition or removal of samples could dramatically alter these fold changes, potentially impacting the accuracy of the analysis, especially if the sample set was skewed towards a specific tissue type.
For instance, a predominance of tumorous liver samples would move the median, used for fold changes, towards the average expression of a cancerous liver cell, rendering their inferred activity less pronounced. Furthermore, it would affect all other cell types in the analysis, which would be compared to something that mostly resembles a cancerous liver cell. Hence, a large degree of skewness of a dataset used for miReact would make all the results obsolete.

With the integration of bulkDGD, these issues have been effectively resolved. bulkDGD calculates controls independently of the other samples in the dataset, ensuring that the addition or removal of samples no longer impacts miReact's analysis. This independence means that the results remain accurate and reliable, regardless of the composition of the dataset.

5.2.2 Adapting the DGD model to the setting

The pretrained bulkDGD model already delivers impressive results by capturing key features and patterns from the training data. However, we hypothesize that its performance could be further enhanced by addressing certain limitations.
Currently, all representations are initialized at the means of the GMM components representing the specific tissues the model was trained on. This can be challenging when 13 out of the 27 sample sites from the TCGA data are not directly comparable to the training data. If new tissue types differ significantly from those the model was originally trained on, performance

may degrade as the control will be based on the most similar type in the training data.

To mitigate this issue, it might be advantageous to train the model on an even wider variety of samples. This might also give the model a wider landscape of the cell types in the body, helping in analysing cancer samples that deviate from typical patterns.

Alternatively, a model trained on tissues directly reflecting the specific types needed might be preferable for specific cases. This would make the model more robust in its decision. In this case, excluding known metastatic samples might be needed since those samples could resemble tissues outside the scope of a limited model.
By addressing these limitations, we hypothetically could achieve even greater precision and reliability in our results.

5.3 Incorrectly matched samples

The misassignment of samples, discussed in section 4.4.2, can significantly contribute to information loss in our analysis.

The decoder in our bulkDGD model is trained to construct the most accurate normal from a given latent representation. Initially, these sample representations are set at the means of the GMM and are trained to fit well within this structure while simultaneously minimizing reconstruction loss.

Throughout our training process of the representations both the GMM parameters and the decoder parameters remain frozen. We hypothesize that the domain of the latent space, which is defined by the data and tissues that it has been trained on, has limitations when it comes to handling samples that are far from its known domain. This both applies to samples of unseen tissue types and potentially also tumorous samples that have strongly deviated from its normal counterpart that it can be hard to place in the latent space. This would mean that cancer samples that are misassigned are likely those that have diverged the most in the latent space. An example of this would be adrenal gland, where 56.39% of the samples are being assigned to a GMM component meant for brain tissue. Once processed by the decoder we hypothesise that its produced normals more closely resemble brain tissue rather than adrenal gland tissue. This is also the concern for samples that have no comparable tissue in the pretrained model. The decoder has not been trained on those tissues and will consequently not know how to generate their normals. It will try to find the closest comparison but will not find the actual structure of its normal state.

Such misclassification can greatly impact miReacts inferred activity when using DGD generated controls. For example, miR-200c-3p has been measured as prominently upregulated in studies of ovarian cancer[2]. Although miReact detects some upregulation in these samples they are also associated with many different components. Notably, only two samples from ovaries are assigned to the component associated with ovaries. A closer examination using Figure 16.b, with inferred activity for miR-200c-3p for each of the GMM components and their associated tissue, reveals the misassignment of the ovary samples. The majority of our 'Ovary' samples with a high inferred activity are assigned the component associated with 'Testis' and 'Breast'. The samples belonging to components 'Cells', 'Kidney' and 'Minor Salivary Glad' are closer to 0.

This adds a lot of noise to the output as samples that have diverged far enough from the latent space of their normals now have an inferred activity close to zero because they are being compared to controls that resemble another tissue.

If the hypothesis is true, the misassigned samples would be of great scientific interest as they are the ones that have been changed the most by cancer. Hence, they could currently be a significant loss of information.

This hypothesis suggests that miReact could achieve more distinct results using generated controls if the assignments in bulkDGD are corrected.
A proposal for a structural modification to prevent these issues will be discussed in Section 7.1.

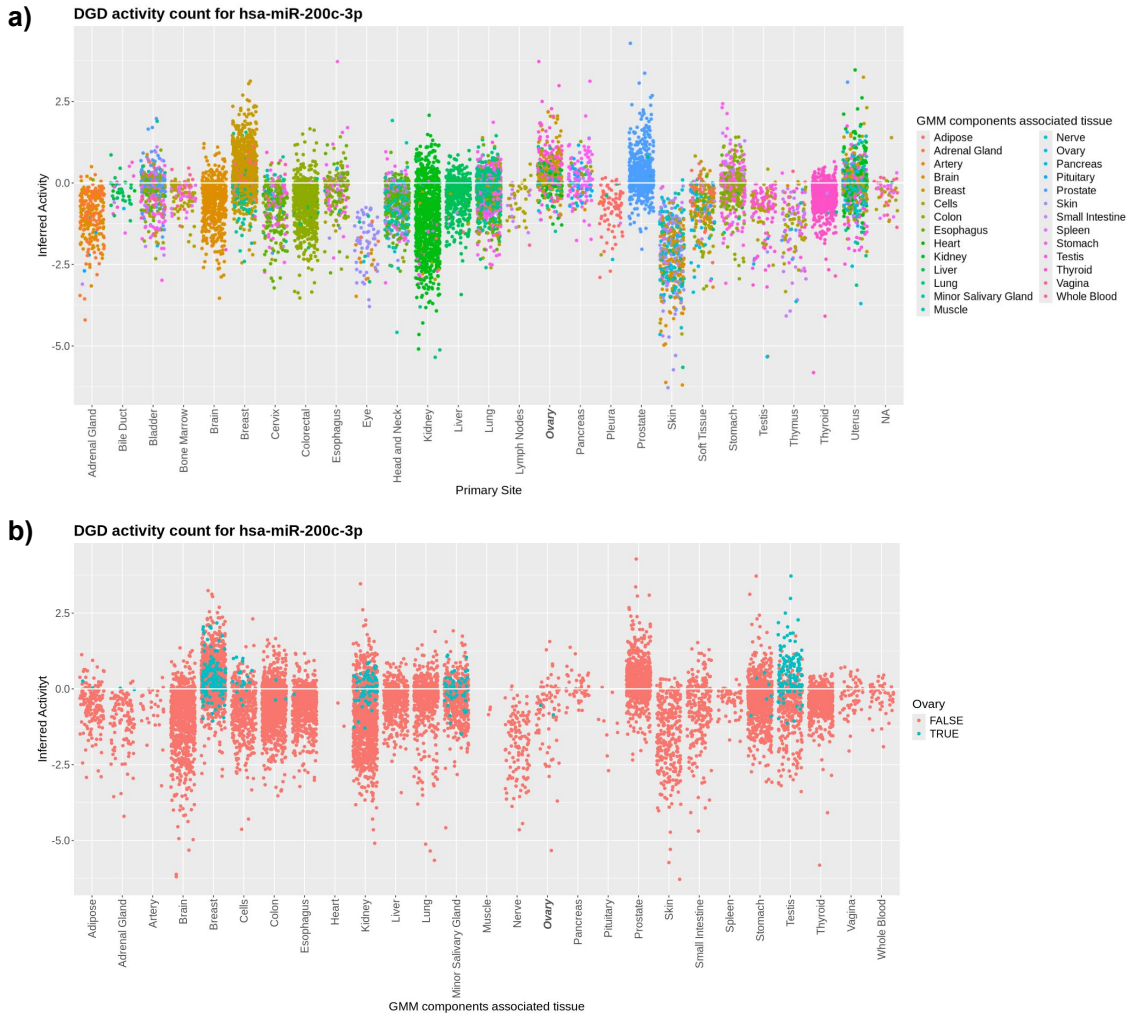


Figure 16: Barplots of inferred activity of miR-200c-3p **a)** activity across all sample sites, colored by GMM components associated tissue. **b)** activity across all the associated tissues for components, colored by whether the sample site was 'Ovary'(blue) or not (red).

6. Conclusion

In this study, we introduced and integrated the statistical tools miReact and the Deep Generative Decoder to enhance the estimation of miRNA activity from mRNA expression data, with a focus on its application in cancer research. By applying the DGD model to thousands of cancer samples, we generated normal controls, addressing the challenge in miReact of obtaining reliable control samples.

Our comparative analysis of miReact's performance, with and without the integration of

DGD-generated controls, demonstrated how the use of generated controls shifted miReact's aim. By comparing cancer samples with their own healthy counterparts, the inferred miRNA activity levels primarily reflected the impact of cancer, allowing for better analysis of cancer-specific impact on miRNA behaviour.

Additionally, the study uncovered some pitfalls of the pretrained bulkDGD model when applied to a larger dataset like ours. We suggested improvements to address these issues, which could further enhance the performance of DGD and, consequently, miReact using DGD-generated controls.

7. Directions for future research

The work with miReact and DGD has illuminated multiple avenues for further research that might improve the combined model or shed light upon its difficulties, enhancing understanding of miRNA.

7.1 Supervised Bulk DGD

In the paper "The Deep Generative Decoder: MAP estimation of representations improves modeling of single-cell RNA data" it has already been tried for other forms of DGD to develop a supervised version. That supervised version would assign a GMM component for each tissue type a priori and would then only be trained to cover the latent space of the assigned tissues' representations. This approach, however, would limit the model's ability to discover substructures and valuable information about the relationships between different structures in the latent space [20].

Our proposal to avoid this problem would be to train a model unsupervised (like it is currently). From a pretrained model, we have the possibility to extract the tissue each component is most responsible for (see Methods:Extracting Gaussian Mixture Components).

Using the pretrained model an extension could be made to make the model semi-supervised. This would be done by including the tissue each sample is taken from as an input argument. Then instead of initializing the representations at each of the GMM components means, they would only be initialized at the component associated with the corresponding tissue. This would give us a greater control over the sample identity. This would solve our problems of false assignment.
Hypothetically the more progressed a cancer is, the less it resembles the original tissue and the higher risk of false classification. Hence some of the most interesting samples are likely falsely assigned. Using a semi-supervised model would then yield more interesting results.
It is important to note that this proposal would not be usable for metastatic cancers, since they might not resemble the tissue of the sample site.

7.2 Other diseases

The model could be used to explore miRNA activity in the context of other diseases, giving insights and potential ideas for treatment. An idea could be to use the model on SARS-CoV-2 because of the extensive datasets that were created during the height and aftermath of the COVID-19 pandemic. Some of those datasets include multiple samples from the same patients at different time points following the initial diagnosis. The disruption and re-establishment of miRNA activity could be interesting to explore as well as which miRNAs that have prolonged abnormal states. This would give understanding of COVID progression and potential

therapeutic targets

8. Data and material availability

bulkDGD

The python package implementing the Deep Generative Decoder and pre-trained model can be found at the GitHub repository:
GitHub: Center-for-Health-Data-Science/bulkDGD

Documentation on the bulkDGD package and tutorials on finding best representations and computing fold changes can be found at:
bulkDGD: Finding the best representations for a set of new samples
bulkDGD: Differential Expression Analysis

miReact

The miReact software is written in R and available at the GitHub repository: [12]
GitHub: miReact

Regmex is available as a R package at [13]:
GitHub: Regmex

Git repository of the project

Code used for analysis and model integration:
GitHub: repository of the project

9. Bibliography

[1] Jorge AL, Pereira ER, Oliveira CS, Ferreira ES, Menon ETN, Diniz SN, and Pezuk JA. Micrornas: understanding their role in gene expression and cancer. *einstein(São Paulo)*, 2026. doi: 10.31744/einstein_journal/2021RB5996.

[2] Sheril June Ankasha, Mohamad Nasir Shafiee, Norhazlina Abdul Wahab, Raja Affendi Raja Ali, and Norfilza Mohd Mokhtar. Oncogenic role of mir-200c-3p in high-grade serous ovarian cancer progression via targeting the 3'-untranslated region of DLC1. *Int. J. Environ. Res. Public Health*, 18(11):5741, May 2021.

[3] Ilaria Cavallari, Francesco Ciccarese, Evgeniya Sharova, Loredana Urso, Vittoria Raimondi, Micol Silic-Benussi, Donna M. D'Agostino, and Vincenzo Ciminale. The mir-200 family of micrornas: Fine tuners of epithelial-mesenchymal transition and circulating cancer biomarkers. *Cancers*, 13(23), 2021. ISSN 2072-6694. doi: 10.3390/cancers13235874. URL `https://www.mdpi.com/2072-6694/13/23/5874`.

[4] Mahboobeh Faramin Lashkarian, Nasrin Hashemipour, Negin Niaraki, Shahrad Soghala, Ali Moradi, Sareh Sarhangi, Mahsa Hatami, Fatemehsadat Aghaei-Zarch, Mina Khosravifar, Alireza Mohammadzadeh, Sajad Najafi, Jamal Majidpoor, Poopak Farnia, and Seyed Mohsen Aghaei-Zarch. Microrna-122 in human cancers: from mechanistic to clinical perspectives. *Cancer Cell International*, 23(1), February 2023. ISSN 1475-2867. doi: 10.1186/s12935-023-02868-z. URL `http://dx.doi.org/10.1186/s12935-023-02868-z`.

[5] Andris Finkbeiner and Wendy Jiang. miRNA Processing. `https://app.biorender.com/biorender-templates/figures/all/t-63e0f719d36c205dcc1fa55f-mirna-processing`, 2024.

[6] National Cancer Institute. The cancer genome atlas program (TCGA). `https://www.cancer.gov/ccg/research/genome-sequencing/tcga`, 2023. Accessed: may 21.

[7] Catherine Jopling. Liver-specific microrna-122: Biogenesis and function. *RNA Biology*, 9 (2):137–142, 2012. doi: 10.4161/rna.18827. URL `https://doi.org/10.4161/rna.18827`. PMID: 22258222.

[8] Ariany Lima Jorge, Erik Ribeiro Pereira, Christian Sousa de Oliveira, Eduardo dos Santos Ferreira, Edmara Toledo Ninzoli Menon, Susana Nogueira Diniz, and Julia Alejandra Pezuk. Micrornas: understanding their role in gene expression and cancer. *Einstein (São Paulo)*, 19, 2021. ISSN 2317-6385. doi: 10.31744/einstein_journal/2021rb5996. URL `http://dx.doi.org/10.31744/einstein_journal/2021RB5996`.

[9] Yong Sun Lee and Anindya Dutta. MicroRNAs in cancer. *Annu. Rev. Pathol.*, 4(1): 199–227, 2009.

[10] J. C. Medley, G. Panzade, and A. Y. Zinovyeva. Microrna strand selection: unwinding the rules. *WIREs RNA*, 12, 2020. doi: 10.1002/wrna.1627.

[11] Michael O'Donnell Michael M. Cox, Jennifer A. Doudna. *Molecular Biology: Principles and Practice*. W. H. Freeman, 2015. ISBN 9781319036119.

[12] Morten Muhlig Nielsen and Jakob Skou Pedersen. mirna activity inferred from single cell mrna expression. *Scientific Reports*, 11(1), 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-88480-5.

[13] Morten Muhlig Nielsen, Paula Tataru, Tobias Madsen, Asger Hobolth, and Jakob Skou Pedersen. Regmex: A statistical tool for exploring motifs in ranked sequence lists from genomics experiments. *Algorithms for Molecular Biology*, 13(1), 2018. ISSN 1748-7188. doi: 10.1186/s13015-018-0135-2.

[14] Broad Institute of MIT and Harvard. Genotype-tissue expression (gtex). https://gtexportal.org/home/aboutAdultGtex, 2023. Accessed: june 11.

[15] Khalid Otmani and Philippe Lewalle. Tumor suppressor miRNA in cancer cells and the tumor microenvironment: Mechanism of deregulation and clinical implications. *Front. Oncol.*, 11, October 2021.

[16] Fei Peng, Ting-Ting Li, Kai-Li Wang, Guo-Qing Xiao, Ju-Hong Wang, Hai-Dong Zhao, Zhi-Jie Kang, Wen-Jun Fan, Li-Li Zhu, Mei Li, Bai Cui, Fei-Meng Zheng, Hong-Jiang Wang, Eric W-F Lam, Bo Wang, Jie Xu, and Quentin Liu. H19/let-7/LIN28 reciprocal negative regulatory circuit promotes breast cancer stem cell maintenance. *Cell Death Dis.*, 8(1):e2569, January 2017.

[17] Iñigo Prada-Luengo and Valentina Sora. bulkDGD, 2024. URL https://github.com/Center-for-Health-Data-Science/bulkDGD.

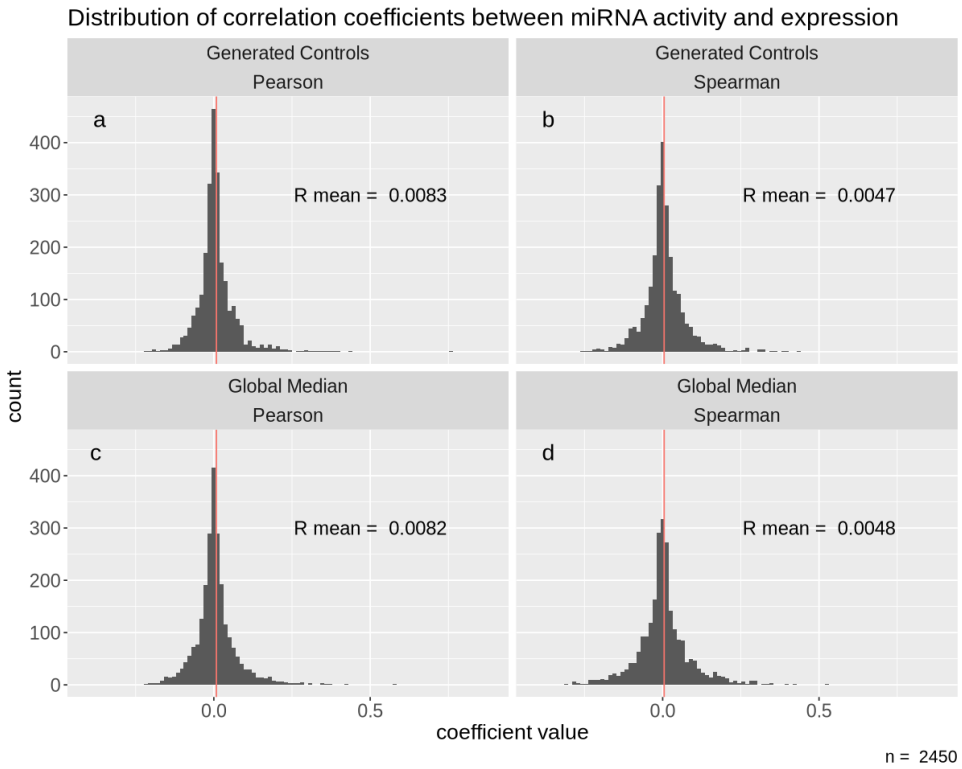
[18] Iñigo Prada-Luengo, Viktoria Schuster, Yuhu Liang, Thilde Terkelsen, Valentina Sora, and Anders Krogh. N-of-one differential gene expression without control samples using a deep generative model. *Genome Biology (Online Edition)*, 24(1), 2023. ISSN 1474-7596. doi: 10.1186/s13059-023-03104-7. Publisher Copyright: © 2023, The Author(s).

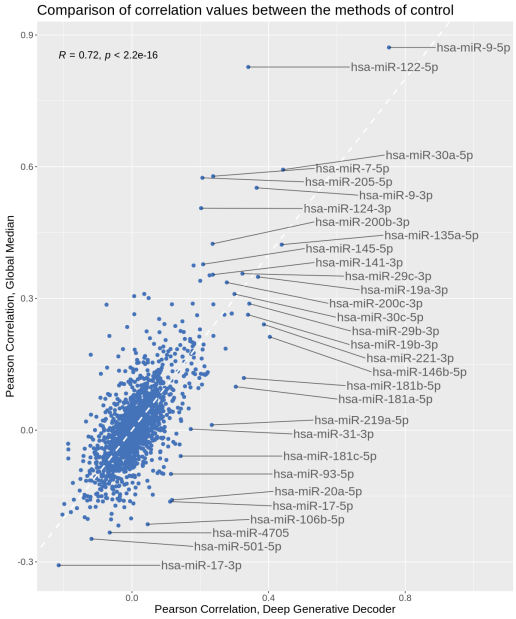
[19] Ian A Prior, Paul D Lewis, and Carla Mattos. A comprehensive survey of ras mutations in cancer. *Cancer Res.*, 72(10):2457–2467, May 2012.

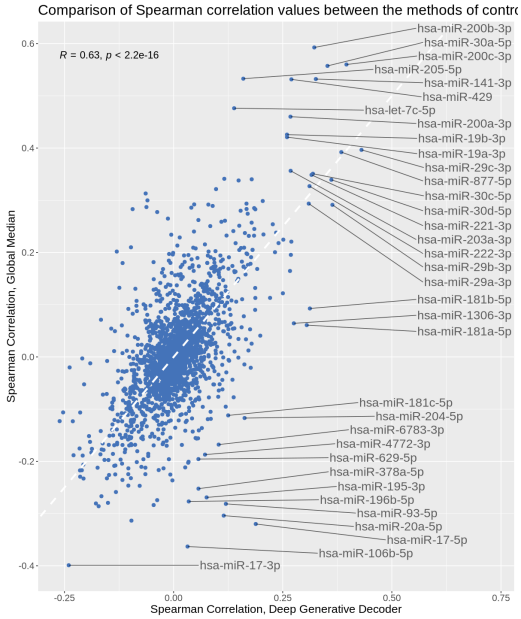
[20] Viktoria Schuster and Anders Krogh. The deep generative decoder: Map estimation of representations improves modelling of single-cell rna data. *Bioinformatics*, 39(9), August 2023. ISSN 1367-4811. doi: 10.1093/bioinformatics/btad497. URL http://dx.doi.org/10.1093/bioinformatics/btad497.

Appendices

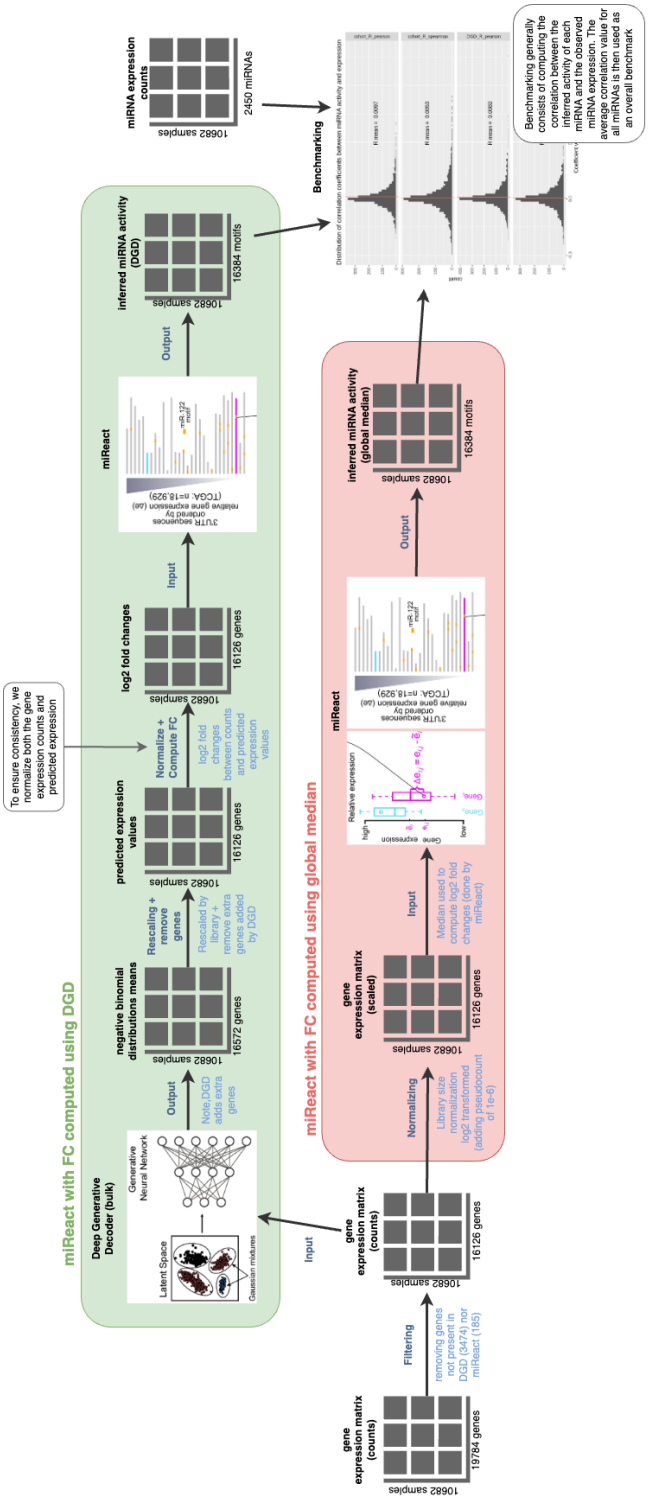
A. Comparison of correlation scores







B. Extended process overview



C. DGD Performance