PLOS ONE

RESEARCH ARTICLE

Bayesian multilevel model of micro RNA levels in ovarian-cancer and healthy subjects

Paweł Wiczling[1☯]*, Emilia Daghir-Wojtkowiak[1☯], Roman Kaliszan[1], Michał Jan Markuszewski[1], Janusz Limon[2], Magdalena Koczkowska[2], Maciej Stukan[3], Alina Kuźniacka[2], Magdalena Ratajska[2]

1 Department of Biopharmaceutics and Pharmacodynamics, Medical University of Gdańsk, Gen. J. Hallera, Gdańsk, Poland, **2** Department of Biology and Genetics, Medical University of Gdańsk, Dębinki, Gdańsk, Poland, **3** Department of Gynecological Oncology, Gdynia Oncology Centre, Powstania Styczniowego, Gdynia, Poland

☯ These authors contributed equally to this work.
* wiczling@gumed.edu.pl

Citation: Wiczling P, Daghir-Wojtkowiak E, Kaliszan R, Markuszewski MJ, Limon J, Koczkowska M, et al. (2019) Bayesian multilevel model of micro RNA levels in ovarian-cancer and healthy subjects. PLoS ONE 14(8): e0221764. https://doi.org/10.1371/journal.pone.0221764

Editor: Bernard Mari, Institut de Pharmacologie Moleculaire et Cellulaire, FRANCE

Received: December 13, 2018

Accepted: August 14, 2019

Published: August 29, 2019

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: This work was supported by the Polish National Science Centre project: 2011/02/A/NZ2/00017 and 2012/07/E/NZ7/04411. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

In transcriptomics, micro RNAs (miRNAs) has gained much interest especially as potential disease indicators. However, apart from holding a great promise related to their clinical application, a lot of inconsistent results have been published. Our aim was to compare the miRNA expression levels in ovarian cancer and healthy subjects using the Bayesian multilevel model and to assess their potential usefulness in diagnosis. We have analyzed a case-control observational data on expression profiling of 49 preselected miRNA-based ovarian cancer indicators in 119 controls and 59 patients. A Bayesian multilevel model was used to characterize the effect of disease on miRNA levels controlling for differences in age and body weight. The difference between the miRNA level and health status of the patient on the scale of the data variability were discussed in the context of their potential usefulness in diagnosis. Additionally, the cross-validated area under the ROC curve (AUC) was used to assess the expected out-of-sample discrimination index of a different sets of miRNAs. The proposed model allowed us to describe the set of miRNA levels in patients and controls. Three highly correlated miRNAs: miR-101-3p, miR-142-5p, miR-148a-3p rank the highest with almost identical effect sizes that ranges from 0.45 to 1.0. For those miRNAs the credible interval for AUC ranged from 0.63 to 0.67 indicating their limited discrimination potential. A little benefit in adding information from other miRNAs was observed. There were several miRNAs in the dataset (miR-604, hsa-miR-221-5p) for which inferences were uncertain. For those miRNAs more experimental effort is needed to fully assess their effect in the context of new hits discovery and usefulness as disease indicators. The proposed multilevel Bayesian model can be used to characterize the panel of miRNA profile and to assess the difference in expression levels between healthy and cancer individuals.

Introduction

MicroRNAs (miRNAs) are abundant classes of endogenous, small non-coding RNAs of 17–25 nucleotides in length generated from 70–100 nucleotides-long hairpin precursors, which

regulate gene expression post-transcriptionally by affecting the translation of target messenger RNAs (mRNAs) [1]. mRNA target recognition by a single miRNA is found in different regions of mRNA, particularly in the 3' untranslated region (3'UTR), 5' untranslated region (5'UTR) and in the coding sequences [2], depending solely on a complementarity with the 6–8 5′ nucleotides of the miRNAs. The same miRNA may have different effects on the same disease. A single miRNA can affect hundreds of mRNA targets acting as oncogenes or tumor suppressors in a cellular-dependent context and depending on the genes targeted [3, 4]. Accumulated evidences have shown that miRNA expression is altered in most types of cancer being involved in a regulation of a wide range of developmental, physiological and cellular processes e.g. proliferation, adhesion, apoptosis and angiogenesis [5].

Therefore, a lot of effort has been paid towards searching for promising miRNA hits for diagnosis and treatment of various types of cancer e.g. breast cancer [6], leukemia [7,8], liver cancer [9,10], ovarian cancer [11], pancreatic and prostate cancer [12,13], and other diseases as well (cardiovascular, metabolic diseases, neurodegenerative disorders) [14,15,16].

Traditional experiments towards searching for novel miRNA-disease associations cost a lot of manpower, material and financial resources. For this reason, much effort is undertaken towards building effective and accurate computational models to reveal the potential relationship between disease and miRNA according to the hypothesis that miRNAs with similar functions are likely to be involved in diseases with similar phenotypes and vice versa (Bandyopadhyay, et al., 2010).

According to a state-of-the-art of existing miRNA-disease association studies, computational prediction models have been divided into four categories, (i) score function-based, (ii) complex network algorithm-based, (iii) machine learning-based, and (iv) multiple biological information-based models (comprehensively described in the review by Chen et al. [17].

Briefly and generally, the score function-based models assume that there is higher probability of association between functional-related miRNAs and phenotypically similar diseases. As its foundations lie in the probabilistic theory, assumption of prior knowledge on data distribution may affect prediction especially if the data informational content is poor. However, due to the lack of experimentally supported miRNA–target interactions, score function-based models provide high rates of false-positive and false-negative results. In this family of models, the most up-to-date models is The Within and Between Score for MiRNA–Disease Association prediction (WBSMDA) [18].

The complex network algorithm-based methods involve different aspects of miRNA similarity networks and disease similarity networks. This method is based on the use of topological information of the miRNA-disease bilayer network assuming that functionally similar miRNAs are more likely to be involved in a similar disease and vice versa which is in accordance with biological experiments. However, the drawback of this methods lies in a difficulty in their application to a new disease unless more experimental data on miRNA/disease function interaction network is collected. One of the most up-to date examples of this algorithms are: random walk-based computational model of Random Walk with Restart for MiRNA–Disease Association (RWRMDA) [19], random walk on the miRNA–disease bilayer network (MIDP) [20], Path-Based computational model for MiRNA–Disease Association (PBMDA) [21], Heterogeneous Graph Inference for MiRNA–Disease Association prediction (HGIMDA) [22], Random Walk and Binary Regression-based MiRNA-Disease Association prediction (RWBRMDA) [23].

The machine learning-based prediction models use machine learning algorithms for predictions via extracting the most relevant features or solving specific optimization problems. These kinds of methods predict the potential miRNAs for a new disease, without any previous associated disease. Machine learning-based model can incorporate different covariates for the final prediction offering improvement in the prediction performance. The most up-to-date

examples of such algorithms are KRLSM for predicting miRNA–disease associations using Kronecker RLS based on heterogeneous omics data [24], Matrix Completion for MiRNA–Disease Association prediction model (MCMDA) [25], Ranking based k-nearest-neighbors for MiRNA–Disease Association prediction (RKNNMDA) [26], Adaptive Boosting for MiRNA-Disease Association prediction (ABMDA) [27], Negative Samples Extraction based MiRNA-Disease Association prediction (NSEMDA) [28]. Multiple biological information-based models assume integration of information between miRNA–gene and disease–protein associations to explore miRNA-related and disease-related associations. The most up-to-date examples of such algorithms are computational model to infer miRNA–Protein–Disease associations (miRPD) [29] and computational framework named KBMFMDI [30], Adaptive Multi-View Multi-Label learning(AMVML) [31], Matrix Decomposition and Heterogeneous Graph Inference for miRNA-disease association prediction (MDHGI) [32].

The above-mentioned computational-based methods have their own strengths and weaknesses. The latter may result from: (1) rare existence of identified miRNA–disease associations; (2) unavailable data on negative miRNA–disease associations; (3) limited biological data sets about miRNAs; (4) difficulties in applying computational models to miRNAs without any prior knowledge on associated diseases [17].

At present, as a complement to existing computer-based methods, more interest is paid to those based on Bayesian statistics i.e. neoteric Bayesian model (KBMFMDA) which combines kernel-based nonlinear dimensionality reduction, matrix factorization and binary classification [33] and Bayesian probabilistic matrix factorization (MDBPMF), in order to discover novel miRNA-disease associations [34], or variational Bayesian Gaussian mixture model (VB-GMM) to predict miRNA target genes [35].

The general idea of all computational models is to find the candidate miRNAs potentially associated with the disease of interest and further confirm these top miRNAs in experiments. As mentioned earlier, the pros of this approach lie in saving a lot of experimental effort with respect to miRNA-disease association.

Apart from computational models which assess miRNA-disease associations, a different aspect of analysing miRNA data from meta analyses or a single experiment involves estimation with quantified uncertainty based on effect size and credible intervals. This approach was used by Eftekharian et al. [36] and Sayad et al. [37] and is attributed to multilevel Bayesian models. This approach represents the idea that data generated in experiments can be described via mathematical models. The concept behind fitting a model to the data lies in generating thousands of random samples of the actually-observed data in order to estimate the values of model parameters with appropriate uncertainty. The more data with greater informational content, the better precision of estimation (and otherwise, less data with lesser information content, worse precision of estimation). Bayesian multilevel models encourage shifting to estimation with uncertainty and magnitude of uncertainty aiming at estimating precision rather than testing hypotheses promoting black-and-white thinking. [38]. Such distinction between null hypothesis significance testing (NHST), on the one hand, and estimation with quantified uncertainty on the other, has indirectly been promoted by the presence of "noise" in the data. For this reason, the effect size statistics seems to be a useful measure providing the information on (i) the magnitude or strength of outcomes, (ii) power for future studies and may be useful in meta-analyses while summarizing the effect sizes across independent studies [39]. To estimate the effect size for ANOVA models, the following statistics are usually used: (i) eta-squared (η^2), (ii) partial eta-sqaured ($\eta_p{}^2$) and (iii) omega-squared (ω^2). Moreover, to simply judge on the mean differences between two measurements one can calculate: R^2 if one performs regression and evaluates the correlation between 2 variables, or Cohen's *d* if one performs a t-test and want to know mean differences in a t-test [40].

Therefore, to assess the relationship between the presence of ovarian cancer and miRNA expression and to judge on the importance of the effect of disease on miRNAs concluding how certain their magnitude can be estimated, we develop a data-driven multilevel Bayesian model. The model included correlations between miRNAs and accounted for inter-individual and assay variability. We also discuss the obtained results in the context of traditional (Frequentists) approach based on controlling the false discovery rate (FDR).

Material and methods

This study was approved by the Research Ethics Committee of the Medical University of Gdansk (NKBBN/399/2011-2012). Written informed consent was obtained from all individual participants included in the study.

Structure of the dataset

Dataset used in this study consisted of vector Y of size N = 8722 (number of observations, n = 1 . . . N) representing centered and standardized miRNA levels measured in plasma and transformed to a natural logarithmic (log) scale for $K = 49$ ($k = 1 \ldots K$) different miRNAs determined in $I = 178$ individuals ($i = 1 \ldots I$, 119 controls and 59 patients) under two replicates. Y is related to the measured quantification cycle, *CT*, through the standard equation (40 –CT) log(2). A control miRNA (UniSp6) constituted cDNA synthesis control and was not included in the data analysis as it was used as an internal control of miRNA profiling. A set of vectors was used to denote indexes representing study design with *k[n]* denoting an indicator for miRNA and *i[n]* denoting indicator for a subject. Health status (0 corresponding to control and 1 to patients) constituted the available discrete covariate and was denoted as *I* x 1 vector DIS. Two continuous covariates were available: age denoted as *I* x 1 vector AGE and body weight denoted as *I* x 1 vector BW. The mean and standard deviation (SD) of data prior to centering and standardization equaled 4.03 and 1.93. Raw data are in S1 Data. The details on the experimental procedure regarding miRNA expression profiling can be found in the S1 Appendix.

Model development

The following multilevel model was used to describe the miRNA data:

$$y_n \sim N(\mu_{k[n]} + \beta_{DIS,k[n]} DIS_{i[n]} + \beta_{BW,k[n]} BW_{i[n]} + \beta_{AGE,k[n]} AGE_{i[n]} + \eta_{i[n],k[n]}, \sigma_{k[n]}) \quad (1)$$

$$\eta_{i,1\ldots K} \sim MVN(0, \boldsymbol{\Omega}) \quad (2)$$

where *N* and *MVN* denote the normal and multivariate normal distribution, a tilde (~) denotes "has the probability distribution of", i.e. the values of y_n and $\eta_{i,1\ldots K}$ are randomly drawn from the given (normal and multivariate normal) distribution, y_n represents the dependent variable; μ_k is the typical miRNA level in a healthy subject of age 52.7 years and body weight of 67.3 kg, $\beta_{DIS,k}$ describes the effect of disease for a particular miRNA; $\beta_{AGE,k}$ and $\beta_{BW,k}$ correspond to the effect of age and body weight covariates on miRNA levels, σ_k denotes standard deviation associated with measurement error for k^{th} miRNA. $\eta_{i.1\ldots K}$ is the between-subject variability of 49^{th} miRNA that was modeled using a MVN distribution with covariance matrix Ω.

The single missing value of a disease status was modeled assuming Bernoulli distribution parametrized using the proportion of cancer/healthy subjects in the data set.

$$DIS_i \sim Bern(0.33) \quad (3)$$

Similarly, the missing values for AGE and BW were assumed to be normally distributed with mean zero and standard deviation equal to 1 (thus to be approximately in a range of ages and body weight of subjects included in the study).

$$AGE_i, BW_i \sim N(0,1) \tag{4}$$

The following prior distribution was assumed during model building process:

$$\sigma_k \sim N(0,1)T(0,) \tag{5}$$

$$\mu_k \sim N(0,5) \tag{6}$$

$$\beta_{DIS,k}, \beta_{BW,k}, \beta_{AGE,k} \sim N(0,1) \tag{7}$$

For the standard deviation, half-normal distribution (expressed as T(0,) in Eq 4) ensuring positive values was used (Eq 5). We also assumed the normal distribution with mean zero and standard deviation of 5 for the mean level of miRNAs (Eq 6). A scaled inverse-Wishart prior was used for the variance-covariance matrix. This was necessary as it allows to estimate the scale parameters and the correlations from the hierarchical data. To implement it we expanded Ω to:

$$\Omega = diag(\zeta) Q\ diag(\zeta) \tag{8}$$

where Q is the unscaled covariance matrix being given the inverse-Wishart model and ζ_k is a scaling factor being given a half-normal model for each miRNA:

$$\mathrm{Q} \sim Inv - Whishart_{K+1}(I_K) \tag{9}$$

$$\zeta_k \sim N(0,1)T(0,) \tag{10}$$

where I_K is a scale, here K x K identity matrix and K+1 denotes degrees of freedom. The ζ and Q parameters cannot be interpreted separately, but allow to calculate the covariance matrix $\Omega = diag(\omega)\rho\ diag(\omega)$, and the most interesting quantities, like standard deviations and correlation matrix [41]:

$$\omega_k = \sqrt{\Omega_{kk}} = |\zeta_k|\sqrt{Q_{kk}} \tag{11}$$

$$\rho_{kk'} = \zeta_k \zeta_{k'} Q_{kk'} / (\omega_k \omega_{k'}) \tag{12}$$

To illustrate the magnitude of the difference between patients and the control group we calculated the effect size for each miRNA on the scale of data variability (d_k).

$$d_k = \beta_{DIS,k} / \sqrt{\omega_k^2 + \sigma_k^2} \tag{13}$$

The effect size along with the associated uncertainty is a useful measure to assess the potential diagnostic value of a single miRNA. The values of d_k larger than 1.5 indicate that the miRNA levels in cancer and patient subjects differ considerably (the underlying normal distributions are almost baseline separated).

AUC under the ROC

To assess the expected out of sample discrimination potential of a subset of miRNA, AUC under the ROC curve was calculated using 10-fold cross-validation. For that purpose the patients from the original data were randomly partitioned into 10 subsamples. Out of the 10

subsamples, a single subsample was excluded from the analysis. The remaining 9 subsamples were used to calculate the probability of cancer for each excluded subject (p_i). The cross-validation process was then repeated 10 times, with each of the 10 subsamples used exactly once as the validation data. The results from the folds were combined and summarized as AUC under the ROC curve. The probability of having cancer for a particular individual was calculated transforming the proposed linear model to its logistic representation [42]:

$$logit(p_i) = \log(0.333) - 0.5(2\mu_{k\epsilon j} + \beta_{DIS,k\epsilon j})\Sigma^{-1}_{k\epsilon j}\beta^{T}_{\beta_{DIS,k\epsilon j}} + y_{k\epsilon j}\Sigma^{-1}_{k\epsilon j}\beta^{T}_{\beta_{DIS,k\epsilon j}} \quad (14)$$

where j denote a subset of miRNA used for predictions, y_j denote miRNA observations from one repetition only, Σ is a sum of the inter-individual and residual variability $\Sigma = \Omega + diag(\sigma^2)$; log(0.33) denotes the ratio of prior probabilities here assumed equal to the proportion of cancer/healthy subjects in the data.

Model assessment

For model diagnostic purposes we plotted (i) weighted residuals versus miRNA and (ii) weighted residuals versus fitted values. This graph evaluated the variability of the observations across each miRNA and assessed the presence of a pattern or trend in the residuals. The weighted residuals should be distributed across zero line with standard deviation near one. If miRNA observations falls outside this range, it indicates model misspecification.

False discovery rate approach

The FDR method was adopted by ranking the raw p values from the lowest to the highest, multiplying each p value by the number of variables, and dividing by its rank order. If the FDR-corrected p-value is less than the significance level 0.05 a variable is conventionally labelled statistically significant.

Technical

The model was developed using JAGS 4.0.0. with *rjags*, *runjags* and *coda* packages in R environment. Three MCMC chains of 100000 iterations were simulated. The first 1000 iterations of each chain were discarded and every 3rd sample was retained. Thus 1000 MCMC samples were used for subsequent analyses. Model convergence was assessed by Gelman-Rubin diagnostics available in JAGS. The MCMC chains were assumed to have reached the stationary distribution if Gelman-Rubin values were less than 1.2 for all parameters. Furthermore, the trace history of MCMC samples for all chains were examined visually for all parameters, for which 'fuzzy caterpillar' suggests that MCMC chains had reached a stationary distribution. The code for the model is available in the S1 Appendix. The FDR was calculated given a set of p-values adjusted using Benjamini & Hochberg method with *stats* package in R environment.

Results

Biological concept of the study

The whole study design initialized with the determination of 752 miRNA levels in 59 samples (first stage of the study): (i) control (n = 16), (ii) ovarian cancer with no BRCA1/2 mutation (-/-) (n = 33) and (iii) ovarian cancer with BRCA1 or BRCA2 mutation (+/+) (n = 10). Further (second stage of the study), based on concentration differences reported between patients and controls in the first stage (based on p-value with FDR correction) and the available literature reports, 49 miRNAs were selected out of 752 hits and further measured in 178 individuals: (i) control (n = 118), (ii) ovarian cancer with no BRCA1/2 mutation (-/-) (n = 49) and (iii)

Fig 1. Raw data (centered and standardized miRNA levels) summarized as boxplots for 49 miRNAs in patients and controls. The box and whiskers plots depict mean, 25th and 75th percentiles. Blue dots overlaid are individual data points.

https://doi.org/10.1371/journal.pone.0221764.g001

ovarian cancer with BRCA1 or BRCA2 mutation (+/+) (n = 10). The second stage of the study included a separate group of individuals not included in the first stage of the study.

Dataset characteristics

The raw data used in this study covered 49 miRNAs measured in 178 individuals (59 ovarian cancer patients and 119 controls) (Fig 1). Detailed characteristics of the available covariates is presented in Table 1. The mean for age and weight of individuals was 52.6 (±13.7) and 67.2 (±11.6).

Table 1. Demographic characteristic of subjects included in the study.

	All subjects	Percent of missing data
Health status	-	0.56%
Age, mean (±SD)	52.67 (±13.7)	2.2%
Weight, mean (±SD)	67.29 (±11.66)	36.5%

https://doi.org/10.1371/journal.pone.0221764.t001

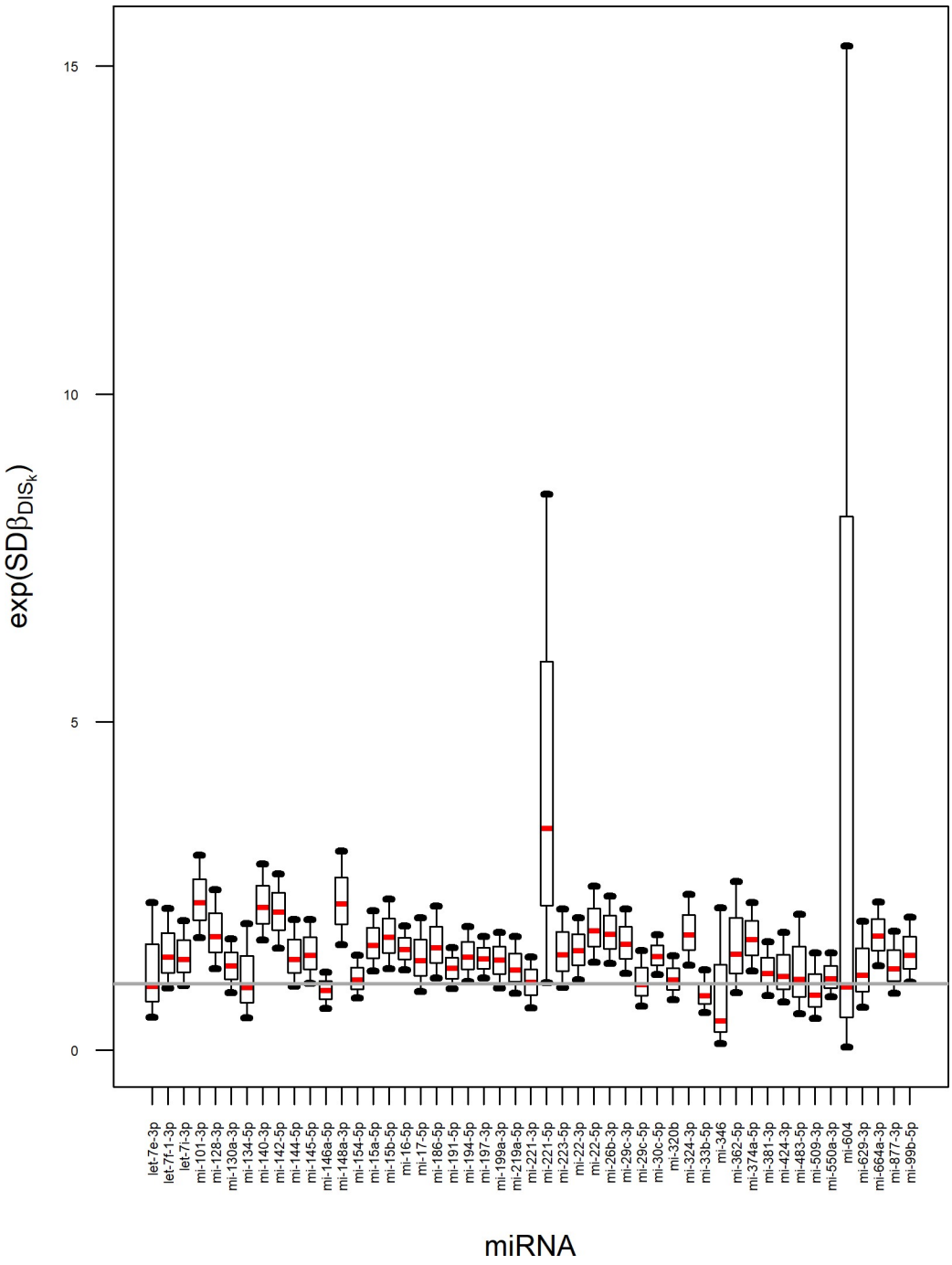

Fig 2. The summary of a marginal posterior distribution representing fold change between disease and control subjects for 49 miRNA. The distribution was summarized as a boxplot with 5th, 25th, 50th, 75th and 95th percentile. Grey line denotes no effect for miRNAs.

https://doi.org/10.1371/journal.pone.0221764.g002

Effect of health status on miRNA levels

Developing the multilevel model we evaluated the effect of health status on miRNA levels (via fold change) by plotting exp($\beta_{DIS,k}$) and associated uncertainty for each miRNA (Fig 2). The miRNA levels are generally higher in patients than in healthy individuals with different level of

uncertainty. The uncertainty is higher for miRNAs with larger number of missing measurements. The miRNAs for which 90% credible interval is above or below the grey horizontal line could be claimed to be associated with the disease (their levels differ between patients and controls) assuming the model and the available data. As an example the fold changes (median (5^{th}-95^{th} percentiles) of $\exp(SD{\cdot}\beta_{DIS,4})$ = 2.24 (1.7–2.96)), $\exp(SD{\cdot}\beta_{DIS,9})$ = 2.11 (1.54–2.68)) and $\exp(SD{\cdot}\beta_{DIS,13})$ = 2.22 (1.59–3.03)) were determined for miR-101-3p, miR-142-5p and miR-148a-3p. The above-mentioned miRNAs were characterized by quite low percentage of missing data i.e. 4.49%, 0% and 26.4%. On the other hand, those miRNA with high proportion of missingness, i.e. miR-221-5p and miR-604 (94.38% and 98.88%) provide very uncertain predictions (their credible interval is consistent with a large range of possible fold changes). To decrease this uncertainty, more data for these miRNAs should be gathered.

Usefulness of miRNAs in cancer detection

By simulations (rather than simply point estimates of parameters), the inferential uncertainty can be propagated into other interesting quantities, like effect size. In this work we estimated the effect size to discuss the difference in miRNA levels between healthy and control subjected on the scale of data variability (Fig 3). The larger the difference the more promising the miRNA for the purpose of diagnosis (to calculate probability of disease). The effect sizes for miRNAs that are characterized by large negative (miR-346) or positive (miR-221-5p) values indicate that for those miRNA it is worth to do more experiments to fully confirm their usefulness in diagnosis. On the other hand if one is willing to select one miRNA for diagnosis based on this data only, three miRNAs (miR-101-3p, miR-142-5p and miR-148a-3p) with effect sizes that are far away from zero would be a good choice as they have the greatest probability of having small effect size. Since they are highly correlated (> 0.95), they carry essentially the same information about health status of the patients.

To calculate the AUC under the ROC curve and further evaluate which combination of miRNAs has the greatest discrimination ability we used 10-fold cross-validation. The one-miRNA-at-a-time AUC under the ROC curve are presented in Table A in S1 Appendix. For the mentioned three miRNAs, i.e miR-101-3p, miR-142-5p and miR-148a-3p, AUC was estimated at 0.65 (0.64–0.66), 0.65 (0.64–0.67), 0.65 (0.62–0.67) suggesting their limited discrimination potency. There is a limited benefit in using more than one miRNA for discrimination, i.e. the use of three miRNAs together led to a very similar AUC of 0.65 (0.63–0.67). There is also a little benefit in adding other miRNAs (i.e. miRNA with missing values being less than 10%). For this subset, the AUC increased to 0.72 (0.69–0.75). This small increase is a consequence of a high correlation between miRNA levels measured in the study.

Model evaluation and estimation of model parameters

The plot of weighted residuals and weighted residuals versus fitted values (Fig A in S1 Appendix) indicated that the variability of the observations were rather constant across miRNAs, with fairly similar spreads at the fitted values. We therefore conclude no bias or trend in model prediction and therefore conclude good specification of the model. The summary of posterior distributions for standard deviations of assay and inter-individual variability σ_k, ω_k for each miRNA were demonstrated in Fig 4.

False Discovery Rate *p* values

Four miRNA levels were found to be significantly different between patients and controls i.e. miR-101-3p, miR-140-3p, miR-142-5p, miR-148a-3p based on FDR p-values. The corresponding effect size (d_{FDR}) for these miRNA was 0.63, 0.64, 0.6, 0.56. We also observed that two

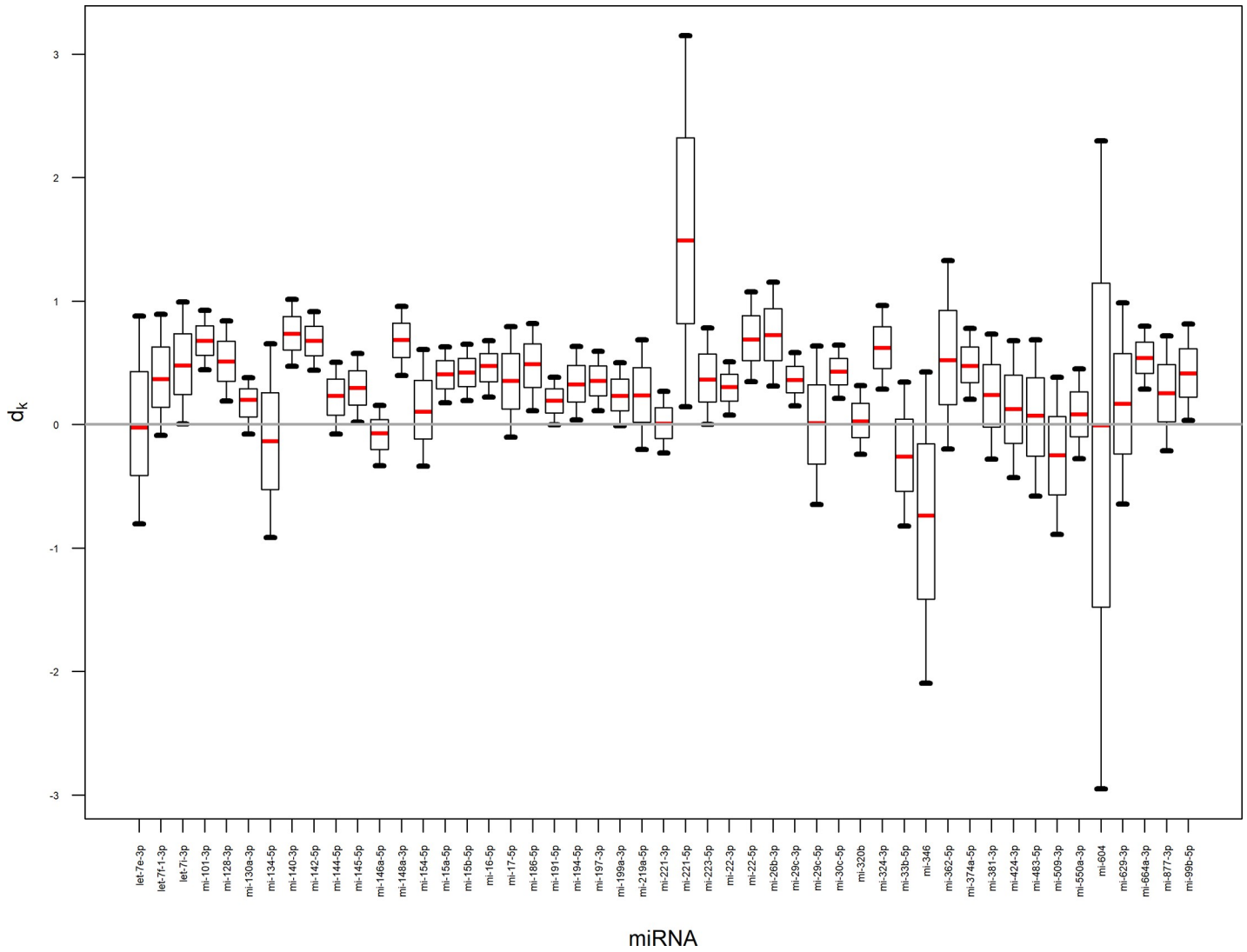

Fig 3. The summary of a marginal posterior distribution of an effect size for 49 miRNA. The distribution was summarized as a boxplot with 5th, 25th, 50th, 75th and 95th percentile. Grey line describes no effect.

https://doi.org/10.1371/journal.pone.0221764.g003

miRNAs which were not significant (according to null hypothesis significance testing) after FDR correction were characterized by the $d_{FDR} > 1$ i.e. miR-221-5p ($d_{FDR} = 1.77$) and for miR-346 ($d_{FDR} = 1.43$). The $d_{FDR} > 1$ without significance after FDR correction could be a consequence of high proportion of missingness for these miRNAs (94.38% and 95.51%). The four miRNAs which were significantly different between patients and controls were also characterized by a relatively low percentage of missingness (Table A in S1 Appendix).

Discussion

Unlike computational–based models aimed at finding a candidate miRNA potentially associated with the disease accompanied by further confirmation its relevance in experiments, the general idea of this work was to judge on the practical relevance of miRNAs in the presence of measurement noise and data variability by proposing the data generating process.

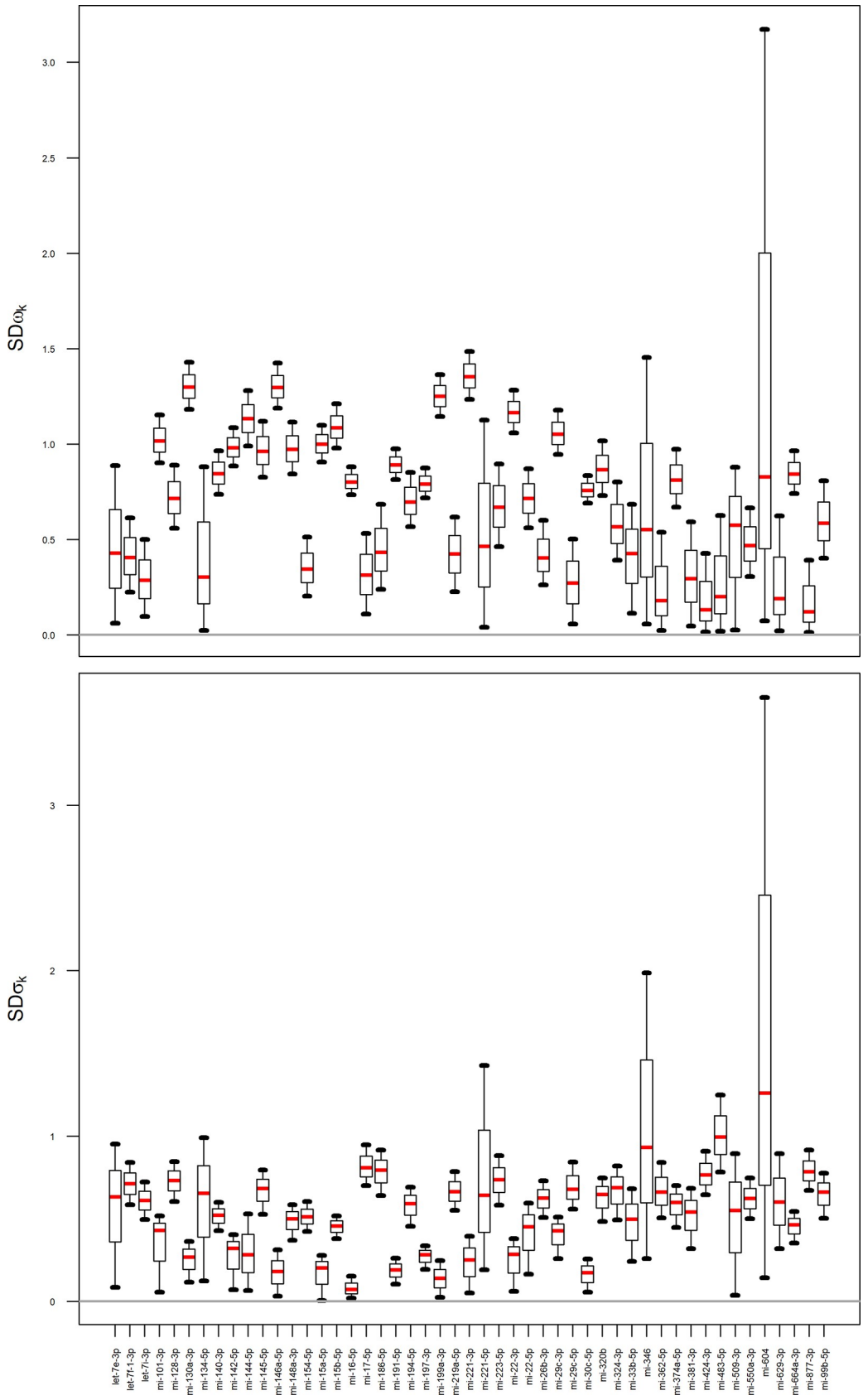
SDωk
3.0
2.5
2.0
1.5
1.0
0.5
0.0
SDσk
3
2
1
0
let-7e-3p
let-7f-1-3p
let-7i-3p
mi-101-3p
mi-128-3p
mi-130a-3p
mi-134-5p
mi-140-3p
mi-142-5p
mi-144-5p
mi-145-5p
mi-146a-5p
mi-148a-3p
mi-154-5p
mi-15a-5p
mi-15b-5p
mi-16-5p
mi-17-5p
mi-186-5p
mi-191-5p
mi-194-5p
mi-197-3p
mi-199a-3p
mi-219a-5p
mi-221-3p
mi-221-5p
mi-223-5p
mi-22-3p
mi-22-5p
mi-26b-3p
mi-29c-3p
mi-29c-5p
mi-30c-5p
mi-320b
mi-324-3p
mi-33b-5p
mi-346
mi-362-5p
mi-374a-5p
mi-381-3p
mi-424-3p
mi-483-5p
mi-509-3p
mi-550a-3p
mi-604
mi-629-3p
mi-664a-3p
mi-877-3p
mi-99b-5p

Fig 4. The summary of a marginal posterior distribution of σ_j, ω_j, parameters for 49 miRNA summarized as boxplots with 5^{th}, 25^{th}, 50^{th}, 75^{th}, 95^{th} percentiles. A large uncertainty was observed for those miRNA for which missing values were reported in high proportion.

https://doi.org/10.1371/journal.pone.0221764.g004

Specifically, we firstly investigated the relationship between the presence of ovarian cancer and miRNA expression via multilevel Bayesian model. Secondly, we assessed the effect of disease on miRNA levels controlling for differences in covariates and modelling the covariance matrix. The effect size and uncertainties around its estimates allowed to judge how certain the magnitude of miRNAs levels can be estimated which constituted an aid in terms of interpretation of the practical relevance of 49 miRNAs measurements for diagnostic purposes.

Under the partial-pooling scenario served by the multilevel Bayesian framework, the parameters' estimates were pulled toward the population mean with a standard deviation set to unity (weekly-informative prior), leading to a reduction of false–discoveries. Another aspect of partial pooling is related to information sharing allowing to estimate individual model parameters. In other words, using multilevel modeling we can make inference on each miRNA borrowing information from other miRNAs.

In the context of gene expression study, there are many factors that affect the obtained results e.g. missing values identified during data generating process evolving mainly from detection sensitivity, contamination, error induced in experimental operations or inappropriate data pre-processing. Their occurrence is unavoidable leading to uncertainty of model estimates and affecting final conclusions if not properly accounted for [43]. For studies considering differentially expressed genes, estimation of a fold–change is the simplest and still widely used measure to identify gene-specific changes, however rarely with a measure of its reliability (due to assumption that all genes exhibit the same level of noise). In the literature, a "significance analysis of microarrays" (SAM) is a common and simple approach utilizing genome–wide information to account for a signal-to-noise ratio [44].

Apart from the existence of this measure and discussing the difference in miRNA levels estimated on the scale of data variability, we used the effect size with associated uncertainty to judge on the relevance of measuring miRNA levels. As evidenced by *a posteriori* distribution of effect sizes of individual miRNA among patients in relation to the control group, several effect sizes are possibly large (above 1) due to large uncertainty. Under such scenario, one should always keep in mind that the true value may lie between those bounds and to reduce this uncertainty around estimates, more data is required. The distribution of effect sizes were also characterized by a negative values which indicate that for some miRNA a decrease in cancer patients is plausible [45].

The posterior distribution of effect size and uncertainties around its estimates allowed for a more sophisticated and intuitive judgment on health status effect exerted on miRNA levels. In our opinion, probabilistic assessments of uncertainty around model parameters and predictions is more convincing especially when making decision on the validity of miRNA measurements for cancer detection [46,47].

In this study we showed that the effect size along with the associated uncertainty can be a useful measure to assess the potential diagnostic value of each miRNA. Accompanying credible intervals for the AUC constituted a kind of "double-check" of miRNAs' diagnostic value. However to prove usefulness for diagnostic purposes, more studies are needed to quantify the added predictive value of individual miRNAs' measurements.

Unlike Bayesian concept to potential markers discovery, under Frequentists approach the effect size is treated as a fixed value addressing the question how much the data differ 'significantly' from that expected under the null. The vast majority of research uses the Frequentists

concept completely ignoring the probability of raised hypothesis [48] and reporting only those results which are statistically significant. Usually, when replicating the original case-control study under the same conditions or under conditions as much similar to the original study as possible, researchers obtain negative results (zero or small effect). This is commonly explained via the presence of a wide sample-to-sample variability in the data, small sample size etc. [49,50]. For this reason, the same experiment will probably result in a substantially different *p*-value which questions reliability of results generally obtained in those experiments, even with high statistical power of a test [51,52].

In observational studies aimed at selecting potential disease indicators, we can observe a high probability of effects' overestimation, high false negative rate resulting from study design and sample size (often inadequate in terms of complexity of the task), lack of confounders adjustment and ignorance of correlations between studied features. The above-mentioned factors influence this one-at-a-time feature selection to a great extent leading to poor predictive performance of developed models and thus generating non-reproducible results. Given that biology is complex and variability in the data is always present, for data analysis purposes we should apply methods that fit better the data i.e. those based on penalization, like the proposed multilevel model [53,54].

Liu et al. however, [55] discuss whether multilevel models may outperform single-gene-at-a-time analysis or SAM in genomics studies. The authors point out that multilevel models can have good performance in case of variance stabilization, however differential expression can be more reasonably analysed with poisson and negative binomial models. Moreover, they underlie that research objective is a key when it comes to decision on the analysis method used. If the goal is gene selection, more computing intensive shrinkage approach should be considered. If we expect large signals changes, fold changes and tail probabilities appear to be the best statistics, otherwise when estimating reliably measured differential expression, the signal-to-noise ratios and Bayes factors is a good choice.

In this study, we built a data-driven model describing the effect of disease and associated covariates on miRNA level simultaneously accounting for the presence of variability. We modelled miRNA data using the normal distribution (with between-subject variability modelled via multivariate normal distribution). Under this scenario, the use of poisson or negative binomial distribution to model miRNA data was unnecessary and could only complicate the model. Although the use of easy-to-use SAM could be a more simple approach to practitioners, we could not apply this measure in this study as we observed significant age difference between patients and healthy individuals (which may be considered a limitation of the study). The use of multilevel model with age and bodyweight adjustment allowed to take the potentially confounding effects of age and body weight into account. The lack of endogenous controls (housekeeping gene) for quantitative control and normalization may also be considered as an limitation of the study.

Conclusions

The relevance of the most promising miRNAs for cancer diagnosis identified in this work (miR-101-3p, miR-142-5p, miR-148a-3p) is rather limited. There are however several miRNAs for which the inferences are uncertain. When analyzing such small data from an observational study caution is always needed as the effects could be biased due to presence of unaccounted confounders. The proposed approach should be considered a more natural statistical formalization of the scientific process of evaluating the evidence. The Bayesian posterior quantifies the uncertainty about the model parameters allowing to make various decision, i.e. assess the usefulness of miRNA for cancer detection.

Supporting information

S1 Data. Raw data.
(CSV)

S1 Appendix. Experimental procedures, codes for Bayesian multilevel model and supporting figure and table.
(DOCX)

Acknowledgments

This work was supported by the Polish National Science Centre project: 2011/02/A/NZ2/00017 and 2012/07/E/NZ7/04411. The authors thank all the patients and their families who participated in this study for their invaluable contribution. Dr Koczkowska is also affiliated with the Medical Genomics Laboratory, Department of Genetics, University of Alabama at Birmingham.

Author Contributions

Conceptualization: Paweł Wiczling, Emilia Daghir-Wojtkowiak.

Data curation: Magdalena Koczkowska.

Formal analysis: Paweł Wiczling, Emilia Daghir-Wojtkowiak.

Investigation: Paweł Wiczling, Emilia Daghir-Wojtkowiak, Magdalena Koczkowska, Maciej Stukan, Alina Kuźniacka, Magdalena Ratajska.

Methodology: Paweł Wiczling.

Project administration: Magdalena Ratajska.

Supervision: Paweł Wiczling, Roman Kaliszan, Michał Jan Markuszewski, Janusz Limon.

Visualization: Emilia Daghir-Wojtkowiak.

Writing – original draft: Paweł Wiczling, Emilia Daghir-Wojtkowiak.

References

1. Das J, Podder S, Ghosh TC. Insights into the miRNA regulations in human disease genes. BMC Genomics 2011; 15: 1010.
2. Lee I, Ajay SS, Yook JI, Kim HS, Hong SH, Kim NH et al. New class of microRNA targets containing simultaneous 5'UTR and 3'UTR interaction sites. Genome Res. 2009; 19: 1175–1183. https://doi.org/10.1101/gr.089367.108 PMID: 19336450
3. Esquela-Kerscher A, Slack FJ. Oncomirs—microRNAs with a role in cancer. Nat Rev Cancer 2006; 6: 259–269. https://doi.org/10.1038/nrc1840 PMID: 16557279
4. Wu D, Hu Y, Tong S, Williams BR, Smyth GK, Gantier MP. The use of miRNA microarrays for the analysis of cancer samples with global miRNA decrease. RNA 2013; 19: 876–888. https://doi.org/10.1261/rna.035055.112 PMID: 23709276
5. Wang J, Chen J, Sen S. MicroRNA as Biomarkers and Diagnostics. J Cell Physiol. 2016; 231: 25–30. https://doi.org/10.1002/jcp.25056 PMID: 26031493
6. Lu M, Ju S, Shen X, Wang X, Jing R, Yang C et al. Combined detection of plasma miR-127-3p and HE4 improves the diagnostic efficacy of breast cancer. Cancer Biomarkers 2017; 18: 143–148. https://doi.org/10.3233/CBM-160024 PMID: 27983524
7. Xu LH, Guo Y, Zhang XL, Chen JJ, Hu SY. Blood-Based Circulating MicroRNAs are Potential Diagnostic Biomarkers for Leukemia: Result from a Meta-Analysis. Cell Physiol Biochem. 2016; 38: 939–49. https://doi.org/10.1159/000443046 PMID: 26938054

8. Swellam M, El-Khazragy N, Clinical impact of circulating microRNAs as blood-based marker in childhood acute lymphoblastic leukemia. Tumour Biology 2016; 37: 10571–6. https://doi.org/10.1007/s13277-016-4948-7 PMID: 26857279
9. Zekri AN, Youssef AS, El-Desouky ED, Ahmed OS, Lotfy MM, Nassar AA et al. Serum microRNA panels as potential biomarkers for early detection of hepatocellular carcinoma on top of HCV infection. Tumour Biology 2016; 37: 12273–12286. https://doi.org/10.1007/s13277-016-5097-8 PMID: 27271989
10. Debernardi S, Massat NJ, Radon TP, Sangaralingam A, Banissi A, Ennis DP et al. Noninvasive urinary miRNA biomarkers for early detection of pancreatic adenocarcinoma. Am J Cancer Res. 2015; 15: 3455–66.
11. Shapira I, Oswald M, Lovecchio J, Khalili H, Menzin A, Whyte J et al. Circulating biomarkers for detection of ovarian cancer and predicting cancer outcomes. Br J Cancer 2014; 18: 976–83.
12. Deng T, Yuan Y, Zhang C, Zhang C, Yao W, Wang C et al. Identification of Circulating MiR-25 as a Potential Biomarker for Pancreatic Cancer Diagnosis. Cell Physiol Biochem. 2016; 39: 1716–1722. https://doi.org/10.1159/000447872 PMID: 27639768
13. Moustafa AA, Ziada M, Elshaikh A, Datta A, Kim H, Moroz K et al. Identification of microRNA signature and potential pathway targets in prostate cancer. Exp Biol Med. 2016; 242: 536–546.
14. Iorio MV, Croce CM. MicroRNA disregulation in cancer: diagnostics monitoring and therapeutics A comprehensive review. EMBO Molecular Medicine 2012; 4: 143–159. https://doi.org/10.1002/emmm.201100209 PMID: 22351564
15. Voellenkle C, van Rooij J, Cappuzzello C, Greco S, Arcelli D, Di Vito L et al. MicroRNA signatures in peripheral blood mononuclear cells of chronic heart failure patients. Physiol Genomics 2010; 2: 420–426.
16. Estep M, Armistead D, Hossain N, Elarainy H, Goodman Z, Baranova A et al. Differential expression of miRNAs in the visceral adipose tissue of patients with non-alcoholic fatty liver disease. Aliment Pharmacol Ther. 2010; 32: 487–97. https://doi.org/10.1111/j.1365-2036.2010.04366.x PMID: 20497147
17. Chen X, Xie D, Zhao Q, You ZH. MicroRNAs and complex diseases: from experimental results to computational models. Briefings in bioinformatics 2019; 22; 20(2): 515–539. https://doi.org/10.1093/bib/bbx130 PMID: 29045685
18. Chen X, Yan CC, Zhang X, You ZH, Deng L, Liu Y et al. WBSMDA: within and between score for MiRNA-disease association prediction. Sci Rep 2016; 6:21106. https://doi.org/10.1038/srep21106 PMID: 26880032
19. Chen X, Liu MX, Yan G. RWRMDA: predicting novel human microRNA-disease associations. Mol Biosyst. 2012; 8:2792–8. https://doi.org/10.1039/c2mb25180a PMID: 22875290
20. Xuan P, Han K, Guo, Li J, Li X, Zhong Y et al. Prediction of potential disease associated microRNAs based on random walk. Bioinformatics 2015; 31:1805–15. https://doi.org/10.1093/bioinformatics/btv039 PMID: 25618864
21. You ZH, Huang ZA. PBMDA: a novel and effective path based computational model for miRNA-disease association prediction. PLoS Comput Biol. 2017; 24; 13(3).
22. Chen X, Yan CC, Zhang X, You ZH, Huang YA, Yan GY, HGIMDA: heterogeneous graph inference for miRNA-disease association prediction. Oncotarget 2016; 7(40):65257–69. https://doi.org/10.18632/oncotarget.11251 PMID: 27533456
23. Niu YW, Wang GH, Yan GY, Chen X, Integrating random walk and binary regression to identify novel miRNA-disease association. BMC Bioinformatics. 2019; 28; 20(1):59. https://doi.org/10.1186/s12859-019-2640-9 PMID: 30691413
24. Luo J, Xiao Q, Liang C, Ding P, Predicting MicroRNA-disease associations using Kronecker regularized least squares based on heterogeneous omics data. IEEE Access 2017; 5:2503–13.
25. Li JQ, Rong ZH, Chen X, Yan GY, You ZH, MCMDA: matrix completion for MiRNA-disease association prediction. Oncotarget 2017; 8:21187. https://doi.org/10.18632/oncotarget.15061 PMID: 28177900
26. Chen X, Wu QF, Yan GY. RKNNMDA: ranking-based KNN for MiRNA-disease association prediction. RNA Biol. 2017; 14: 952–62. https://doi.org/10.1080/15476286.2017.1312226 PMID: 28421868
27. Zhao Y, Chen X, Yin J. Adaptive boosting-based computational model for predicting potential miRNA-disease associations.
28. Wang CC, Chen X, Yin J, Qu J. An integrated framework for the identification of potential miRNA-disease association based on novel negative samples extraction strategy. RNA Biol. 2019; 16(3):257–269. https://doi.org/10.1080/15476286.2019.1568820 PMID: 30646823
29. Mørk S, Pletscher-Frankild S, Palleja Caro A, Gorodkin J, Jensen LJ. Protein-driven inference of miRNA–disease associations. Bioinformatics 2014; 30:392–7. https://doi.org/10.1093/bioinformatics/btt677 PMID: 24273243

30. Lan W, Wang J, Li M, Liu J, Pan Y, Predicting microRNA-disease associations by integrating multiple biological information. In: IEEE International Conference on Bioinformatics and Biomedicine. IEEE, 2015, 183–8.
31. Liang C, Yu S, Luo J. Adaptive multi-view multi-label learning for identifying disease-associated candidate miRNAs. PLoS Comput Biol. 2019; 1; 15(4):e1006931. https://doi.org/10.1371/journal.pcbi.1006931 PMID: 30933970
32. Chen X, Yin J, Qu J, Huang L. MDHGI: Matrix Decomposition and Heterogeneous Graph Inference for miRNA-disease association prediction. PLoS Comput Biol. 2018; 24; 14(8):e1006418. https://doi.org/10.1371/journal.pcbi.1006418 PMID: 30142158
33. Chen X, Li SX, Yin J, Wang CC, Potential miRNA-disease association prediction based on kernelized Bayesian matrix factorization. Genomics. 2019; 25. pii: S0888–7543(19)30191–0.
34. Mao G, Wang SL, Zhang W, Prediction of Potential Associations Between MicroRNA and Disease Based on Bayesian Probabilistic Matrix Factorization Model. J Comput Biol. 2019; 26. https://doi.org/10.1089/cmb.2019.0012 PMID: 31246500
35. Ma XL, Yang X, Fan R, Screening of miRNA target genes in coronary artery disease by variational Bayesian Gaussian mixture model. Exp Ther Med. 2019; 17(3):2129–2136. https://doi.org/10.3892/etm.2019.7195 PMID: 30867700
36. Eftekharian MM, Komaki A, Mazdeh M, Arsang-Jang S, Taheri M, Ghafouri-Fard S, Expression Profile of Selected MicroRNAs in the Peripheral Blood of Multiple Sclerosis Patients: a Multivariate Statistical Analysis with ROC Curve to Find New Biomarkers for Fingolimod. J Mol Neurosci. 2019; 68(1):153–161. https://doi.org/10.1007/s12031-019-01294-z PMID: 30895441
37. Sayad A, Taheri M, Arsang-Jang S, Glassy MC, Ghafouri-Fard S, Hepatocellular carcinoma up-regulated long non-coding RNA: a putative marker in multiple sclerosis. Metab Brain Dis. 2019; 2. https://doi.org/10.1007/s11011-019-00418-z PMID: 31049796
38. Kruschke JK, Torrin ML, The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective, Psychonomic Bulletin & Review 2018; 25(1), 178–206.
39. Fritz C, Morris PE. Effect Size Estimates: Current Use, Calculations, and Interpretation. Journal of Experimental Psychology: General 2012; 141: 1, 2–18.
40. Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front. Psychol 2013; 26.
41. Gelman A, Hill J. Data analysis using regression and multilevel/hierarchical models; Cambridge University Press; 2007.
42. Harrell FE, Lee KL, A comparison of the discrimination of discriminant analysis and logistic regression under multivariate normality. Biostatistics Statistics in Biomedical, Public Health and Environmental Sciences, edited by Sen P. K. Elsevier Science Publisher B.V.(North-Holland), 1985.
43. Yang Y, Xu Z, Song D. Missing value imputation for microRNA expression data by using a GO-based similarity measure. BMC Bioinformatics 2016; 17: 10. https://doi.org/10.1186/s12859-015-0853-0 PMID: 26818962
44. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci. 2001; 98: 5116–5121. https://doi.org/10.1073/pnas.091062498 PMID: 11309499
45. Okada K, Negative estimate of variance-accounted-for effect size: How often it is obtained, and what happens if it is treated as zero, Behavior Research Methods 2017; 49 (3): 979–987. https://doi.org/10.3758/s13428-016-0760-y PMID: 27334468
46. Koech MK, Otieno AR, Kimeli V, Koech EK. Posterior F-Value In Bayesian Analysis Of Variance Using Winbugs Mathematical Theory and Modeling www.iiste.org ISSN 2224-5804 (Paper) ISSN 2225-0522: 4, No.5, 2014.
47. Erkanli A, Taylor DD, Dean D, Eksir F, Egger D, Geyer J et al. Application of Bayesian Modeling of Autologous Antibody Responses against Ovarian Tumor-Associated Antigens to Cancer Detection. Cancer Res. 2006; 66: 1792–1798. https://doi.org/10.1158/0008-5472.CAN-05-0669 PMID: 16452240
48. Feckler A, Low M, Zubrod JP, Bundschuh M. When Significance Becomes Insignificant: Effect Sizes and their Uncertainties in Bayesian and Frequentist Frameworks as an Alternative Approach when Analyzing Ecotoxicological Data. Environ Toxicol Chem. 2018; 37(7):1949–1955. https://doi.org/10.1002/etc.4127 PMID: 29508923
49. van Aert RCM, van Assen MALM. Bayesian evaluation of effect size after replicating an original study. PLoS One. 2017; 12(4): e0175302. https://doi.org/10.1371/journal.pone.0175302 PMID: 28388646
50. Jin X, Xu C, Zhang Q, Singh VP. Parameter and modeling uncertainty simulated by GLUE and a formal Bayesian method for a conceptual hydrological model. J Hydrol. 2010; 383; 147–155.

51. Boos DD, Stefanski LA. P-Value Precision and Reproducibility. Am Stat. 2011; 65: 213–221. https://doi.org/10.1198/tas.2011.10129 PMID: 22690019

52. Halsey LG, Curran-Everett D, Vowler SL, Drummond GB. The fickle P value generates irreproducible results, Nat Methods 2015; 12.

53. Chen Q, Nian H, Zhu Y, Talbot HK, Griffin MR, Harrell FE, Too many covariates and too few cases?–a comparative study, Statist. Med. 2016; 35: 4546–4558.

54. Harrell FE, Slaughter JC, Biostatistics for Biomedical Research, Challenges of analysing high-dimensional data, Chapter 20, http://hbiostat.org/doc/bbr.pdf.

55. Liu D, Parmigiani P, Caffo B, Screening for Differentially Expressed Genes: Are Multilevel Models Helpful?. J Biomet Biostat 2014, 5:2.