Annika Krutto*, Therese Haugdahl Nøst and Magne Thoresen

A heavy-tailed model for analyzing miRNA-seq raw read counts

https://doi.org/10.1515/sagmb-2023-0016
Received April 25, 2023; accepted May 2, 2024; published online May 29, 2024

Abstract: This article addresses the limitations of existing statistical models in analyzing and interpreting highly skewed miRNA-seq raw read count data that can range from zero to millions. A heavy-tailed model using discrete stable distributions is proposed as a novel approach to better capture the heterogeneity and extreme values commonly observed in miRNA-seq data. Additionally, the parameters of the discrete stable distribution are proposed as an alternative target for differential expression analysis. An R package for computing and estimating the discrete stable distribution is provided. The proposed model is applied to miRNA-seq raw counts from the Norwegian Women and Cancer Study (NOWAC) and the Cancer Genome Atlas (TCGA) databases. The goodness-of-fit is compared with the popular Poisson and negative binomial distributions, and the discrete stable distributions are found to give a better fit for both datasets. In conclusion, the use of discrete stable distributions is shown to potentially lead to more accurate modeling of the underlying biological processes.

Keywords: breast cancer; discrete stable distributions; extremes; lung cancer; miRNA-seq raw read counts; TCGA

1 Introduction

Micro ribonucleic acids (miRNAs) form a class of small single-stranded non-coding RNA molecules that serve as regulators of gene expression at the post-transcriptional level. The human genome may encode over 1900[1] miRNAs. With rapid development in high-throughput RNA sequencing (RNA-seq), the quantitative assessment of miRNAs has become more accessible. Recent research has identified miRNAs as possible biomarkers for various diseases,[2] such as cancer. A variety of tests under development promise easier and more complete (miRNA-seq based) cancer-screening capabilities.[3] These screening methods build on mathematical models combined with advanced computation. This has made the analysis of miRNA-seq data a highly demanded topic in the field of biomedical research. The aim of this paper is to provide a novel heavy-tailed approach to the quantitative analysis of miRNAs.

High-throughput miRNA-seq data involve sequencing small RNA molecules within a single experiment, measuring the number of sequencing reads for each miRNA in a sample, termed miRNA-seq raw read counts or

1 E.g., the miRNA database at https://www.mirbase.org/cgi-bin/mirna_summary.pl?org&tnqx3d;hsa.
2 E.g., the Human miRNA Disease Database (HMDD) at https://www.cuilab.cn/hmdd.
3 E.g., https://www.science.org/content/article/catching-cancer-extremely-early.

***Corresponding author: Annika Krutto**, Oslo Centre for Biostatistics and Epidemiology, Department of Biostatistics, University of Oslo, Oslo, Norway, E-mail: annika.krutto@medisin.uio.no
Therese Haugdahl Nøst, Department of Community Medicine, Department of Community Medicine, UiT The Arctic University of Norway, Tromsø, Norway; and Department of Public Health and Nursing, K.G. Jebsen Center for Genetic Epidemiology, UiT The Arctic University of Norway, Trondheim, Norway, E-mail: therese.h.nost@uit.no
Magne Thoresen, Oslo Centre for Biostatistics and Epidemiology, Department of Biostatistics, University of Oslo, Oslo, Norway, E-mail: magne.thoresen@medisin.uio.no

raw expressions. These raw expressions represent discrete non-negative integer-valued random variables with countably infinite support, suitable for modeling using discrete probability distributions. Raw expression data often exhibit high skewness, spanning a wide range from zero to millions, indicative of potential heavy-tailed distributions. This heavy-tailed behavior in miRNA expression counts could be attributed to the inherent biological nature of these molecules, reflecting the substantial variability and diverse expression levels of miRNAs in biological systems. Factors like post-transcriptional regulation, varying cellular conditions, and the intricate networks in which miRNAs operate might contribute to this observed variability, possibly leading to the heavy-tailed distribution observed in miRNA expression counts.

However, the state-of-the-art models view these heavy tails as more of an erratic phenomenon caused by technological variances, and as a result, data pre-processing techniques such as transformation or normalization are utilized as a standard. One of the rationales for this is to reshape the miRNA-seq raw read count distribution in order to streamline the analysis and use classical and well-established statistical methodologies and distributions. For example, in `limma` (Ritchie et al. 2015), a popular `R` package for gene discovery through differential expression analyzes, the raw read counts of the RNA sequence are converted to the logarithmic scale, which compresses the (extreme) scale of the data, and then linear modeling and classical *t*-tests are used. Although logarithmic transformations can be a valuable tool for stabilizing variance, there is a potential for loss of information. Compression of the data scale during logarithmic transformation may result in a loss of precision in variability, especially for large values. This, in turn, might lead to inappropriate conclusions in differential expression analysis (e.g., if you compare the means of log-transformed variables, you are comparing geometric rather than arithmetic means). Other commonly used `R` packages for differential expression analysis are based on fitting the well-known Poisson (e.g., the `DEGseq` package by Wang et al. 2009) or negative binomial (e.g., `EdgeR` package by Robinson et al. 2010, `DESeq` package by Love et al. 2014, `DEseq2` package by Kalecky et al. 2020) distributions, neither of which are heavy-tailed. Another example is miRNA network learning, where the Poisson distribution is widely assumed: a Poisson graphical model was proposed in Yang et al. (2015), a Poisson Markov network model was developed in Žitnik and Zupan (2015), a penalized Poisson graphical model in Choi et al. (2017), and a hierarchical Poisson model is studied in Sinclair and Hooker (2019). Fitting these distributions requires certain data pre-processing procedures, such as tailoring the data to be approximately Poisson (e.g., Allen and Liu 2013) or applying some data normalization technique. For example, in the miRNA quantile normalization (see e.g., Zhao et al. 2020) raw counts of each miRNA are transformed so that the distribution of counts for each miRNA has the same quantile distribution across all samples, which is achieved by rearranging the values for each miRNA so that they have the same rank-order distribution. A special case is the median normalization, where the median expression level of each miRNA is calculated for each sample and then replaced by a reference median value across all samples, typically the median of the median expression levels of all miRNAs in all samples. Another frequent pre-processing approach is the total read normalization, which involves dividing the number of raw read counts for each miRNA by the total number of raw read counts in the sample, and multiplying by a scaling factor, such as 10^6, the reads per million (RpM) value. However, recent studies (e.g., Li et al. 2023; Zhao et al. 2020) emphasize the fact that all known normalization techniques have flaws, can introduce biases and inaccuracies, and can result in a significant loss of information. Relevant information about variability, skewness and tails may be lost. Moreover, it is not clear whether the obtained statistical conclusions would also be valid in the original data. The choice of normalization method can greatly influence the results and an inappropriate choice can lead to incorrect conclusions. Additionally, the choice of probabilistic model used to analyze the miRNA-seq expressions can also have a significant impact on the conclusions. Different distributions may lead to different outcomes, and choosing the improper distribution could impact the reliability of the inferences made from the data. In conclusion, approaches directly modeling miRNA-seq raw read counts are expected to be more suitable as they align closely with the inherent nature of the data. Furthermore, the use of raw counts enhances the interpretability of our findings and avoids unnecessary transformations that could complicate the communication of results.

Taking into account all the foregoing, a vital mathematical perspective that has not yet been fully explored is the properties of an underlying probability distribution that could be applied to the miRNA-seq raw read counts (expressions), without data manipulation. Deciding on a suitable probability distribution is crucial, as it lays

the foundation for highly relevant topics in the discovery of miRNA biomarkers, such as differential expression analysis and miRNA networks. Given the nature of the miRNA-seq raw read counts, it would be natural (and a novel alternative to existing approaches) to model the miRNA-seq raw read counts via discrete heavy-tailed distributions.

This paper introduces a novel model for miRNAs based on the application of the heavy-tailed family of discrete stable distributions, with the Poisson distribution as a special case. The Poisson distribution corresponds to the lightest tail of the discrete stable distributions, and all other family members have heavy tails. The heavier the tails, the more prone are the distributions to extreme values. Due to their properties, discrete stable distributions can be fitted directly to the miRNA-seq raw read counts. In addition to proposing them as the appropriate underlying probability distribution, a novel heavy-tailed approach is introduced for differential expression analysis via their parameters.

While exploring heavy-tailed distributions in the discrete domain, several alternatives such as the popular Poisson-Tweedie distributions (e.g., Baccini et al. 2016) and the generalized Poisson inverse Gaussian distribution (as seen in Qian et al. 2020) are noted. In particular, these distributions, among others, are tempered forms derived from the discrete stable distribution (Grabchak 2018; Grabchak 2022). Tempering functions are used to adjust the heaviness of the tails, for example, to allow for the presence of mean and variance. However, tempered discrete stable distributions, despite their capability to model heavy tails, do not attain the extreme levels allowed by nontempered discrete stable distributions. Given the expansive range from zero to millions observed in miRNA expressions and that, to the best of our knowledge, this is the first application of heavy-tailed distributions in analysis of miRNA count data, we suggest that nontempered discrete stable distributions should be considered initially. We note that modifications of the Poisson distribution, such as the generalized Conway-Maxwell-Poisson distribution (e.g., Qian and Zhu 2023), have also been suggested to address issues such as overdispersion in count data. However, while flexible, these modifications are often limited in their ability to effectively model the extreme count levels observed in miRNA data, unlike discrete stable distributions.

In the applications, two well-studied sources of miRNA-seq read counts are used. First, the mature isoform expression quantification measured in blood samples, provided by the Norwegian Women and Cancer (NOWAC) study. Second, the miRNA expression quantification measured in primary breast tumor tissues, provided by The Cancer Genome Atlas (TCGA) Research Network. All computational algorithms and functions are implemented in R.

Note that for the previous generation miRNA microarray intensity data, heavy-tailed models have been successfully applied (e.g., Misra and Kuruoglu 2016; Purdom and Holmes 2005). However, to the best of our knowledge, this study is going beyond the common standard as the first to model next-generation high-throughput RNA sequencing raw read counts via discrete heavy tailed distributions. In addition, to the best of our knowledge, this study provides the first software application for the calculation and estimation of discrete stable distributions. The first promising results introduced in this paper create a fertile ground for further developments of innovative discrete heavy tailed models in the miRNA based diagnosis and treatment of cancer and other diseases.

2 Methods: modeling via discrete stable distributions

2.1 Discrete stable distributions

Discrete stable distributions form a class of probability distributions defined on nonnegative integers and allowing heavy tails. These features make discrete stable distributions attractive for fitting miRNA-seq raw read counts. Discrete stable distributions are characterized by two parameters, a tail index parameter $\alpha \in (0, 1]$ and a parameter $\lambda > 0$, in this paper, denoted by $DStable(\alpha, \lambda)$. The Poisson distribution forms a special case of discrete stable distributions with $\alpha = 1$ and corresponds to the lightest tail of the discrete stable distributions. Discrete stable distributions with $\alpha < 1$ exhibit heavy tails (e.g., Soltani et al. 2009), also referred to as Paretian tails

(e.g., Rémillard and Theodorescu 2000, Remark 2.7). The lower the tail index $\alpha \in (0,1)$, the slower the decay, the heavier the tail and the more prone to extreme values is the distribution.

2.1.1 Definition

Discrete stable distributions are characterized by their probability generating function. The probability generating function of the random variable X is defined as follows: $G_X(z) = E(z^X)$, operating on the interval $[-1,1]$ and mapping to the range [0,1]. The probability generating function for discrete stable laws is defined as follows (e.g., Devroye 1993; Klebanov and Slámová 2013; Soltani et al. 2009; Steutel and van Harn 1979, Formula (3.7)):

$$G(z) = \mathrm{e}^{-\lambda(1-z)^{\alpha}},\ \alpha \in (0,1],\ \lambda > 0,\ |\ z\ | \leq 1. \tag{1}$$

As is typical for heavy-tailed distributions, not all moments exist. Let X be discrete stable with exponent $\alpha \in (0,1)$ then $\mathbf{E}X^r < \infty$ only for $0 < r < \alpha$ (e.g., Steutel and van Harn 2003, Chapter V, Formulas (5.15)–(5.17)).

2.1.2 Probability calculations

Except for the Poisson distribution, probability mass functions of discrete stable distributions cannot be explicitly expressed in terms of elementary functions. In particular, when calculating $\Pr_{\theta}(X = k)$ for a discrete stable distribution with parameters $\boldsymbol{\theta} = (\alpha, \lambda)$ and $X \sim DStable(\alpha, \lambda)$, no explicit analytic form is available. In (Christoph and Schreiber 1998, Formula (8), p. 245) a recursion formula was derived (see also Steutel and van Harn 2003, Formula (5.18)):

$$\Pr_{\theta}(X = k+1) = \lambda \sum_{m=0}^{k} (-1)^m \frac{m+1}{k+1} \Pr_{\theta}(X = k-m) \binom{\alpha}{m+1}, \tag{2}$$

where $k = 1, 2, 3, \ldots$ and $\Pr_{\theta}(X = 0) = \mathrm{e}^{-\lambda}$. Explicit formulas follow: $\Pr_{\theta}(X = 1) = \alpha\lambda\mathrm{e}^{-\lambda}$, $\Pr_{\theta}(X = 2) = \frac{\alpha\lambda}{2}\mathrm{e}^{-\lambda}(\alpha(\lambda-1)+1)$ and $\Pr_{\theta}(X = 3) = \frac{\alpha\lambda}{3}\mathrm{e}^{-\lambda}[\frac{\alpha\lambda}{2}(\alpha(\lambda-1)+1) - \alpha(\alpha-1)\lambda + \frac{1}{2}(\alpha-1)(\alpha-2)]$.

The impact of changing λ for a fixed tail index α is illustrated in Figure 1. Here, the higher values of λ not only shift the distribution rightward, but also disperse the primary distribution mass (Table 1).

The recursive formula (2) is straightforward, but the calculation time grows rapidly as k increases, and it is computationally time consuming, especially when handling extremes. For calculating the probabilities of large values, (Christoph and Schreiber 1998, Formulas (11) and (12)) proved a tail asymptotic formula for $k \to \infty$:

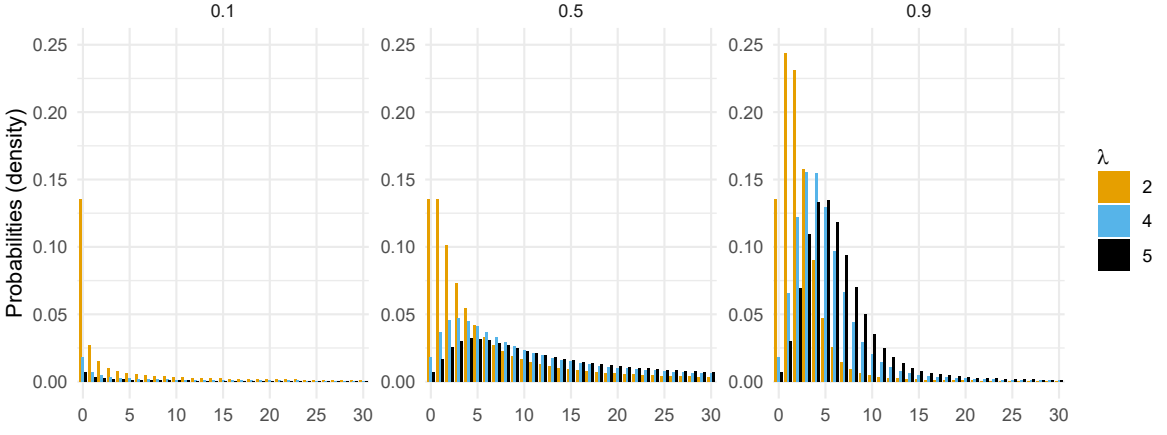

Figure 1: For tail indices $\alpha = 0.1, 0.5, 0.9$ and parameters $\lambda = 2, 4, 5$, the values of $\Pr_{\theta}(X = k)$ of $X \sim DStable(\alpha, \lambda)$ are calculated using the recursive formula (2).

Table 1: For tail indices $\alpha = 1, 0.9, 0.8, 0.4, 0.2$ and parameter $\lambda = 1$, the values of $\Pr_\theta(X = k)$ of $X \sim DStable(\alpha, \lambda)$. Calculations up to $k \leq 10^4$ are made using the recursive formula (2), and for values $k \geq 10^4$, the asymptotic tail formula (3) is used.

k	$\alpha = 1 \equiv$ Pois(1)	$\alpha = 0.9$	$\alpha = 0.8$	$\alpha = 0.4$	$\alpha = 0.2$
0	0.368	0.368	0.368	0.368	0.368
10	1.01×10^{-7}	1.71×10^{-3}	3.76×10^{-3}	9.01×10^{-3}	6.54×10^{-3}
10^2	3.94×10^{-159}	1.55×10^{-5}	4.55×10^{-5}	2.82×10^{-4}	4.98×10^{-4}
10^3	0	1.89×10^{-7}	6.97×10^{-7}	8.92×10^{-6}	3.47×10^{-5}
10^4	0	2.38×10^{-9}	1.10×10^{-8}	2.82×10^{-7}	2.39×10^{-6}
10^5	0	2.99×10^{-11}	1.74×10^{-10}	8.92×10^{-9}	1.58×10^{-7}
10^6	0	3.77×10^{-13}	2.76×10^{-12}	2.82×10^{-10}	1.03×10^{-8}

$$\Pr_\theta(X = k) = \frac{1}{\pi} \sum_{j=1}^{[(\alpha+1)/\alpha]} \frac{(-1)^{j+1}}{j!} \lambda^j \sin(\alpha j \pi) \Gamma(\alpha j + 1) k^{-\alpha j - 1} + O(k^{-\alpha-2}). \tag{3}$$

Note that (Doray et al. 2009, pp. 2008) derived an explicit formula for calculating $\Pr_\theta(X = k)$, which for $k = 1, 2, 3, \ldots$ can be expressed via Gamma functions as follows

$$\Pr_\theta(X = k) = (-1)^k \mathrm{e}^{-\lambda} \sum_{m=0}^{k} \sum_{j=0}^{m} \binom{m}{j} \binom{\alpha j}{k} (-1)^j \frac{\lambda^m}{m!} \tag{4}$$

$$= (-1)^k \frac{\alpha \mathrm{e}^{-\lambda}}{k!} \sum_{m=1}^{k} \lambda^m \sum_{j=1}^{m} (-1)^j \frac{\Gamma(\alpha j)}{(j-1)!(m-j)!} \frac{1}{\Gamma(\alpha j + 1 - k)}, \tag{5}$$

where $\alpha \in (0, 1]$, $\lambda > 0$. For computations in R, the reciprocal Gamma function $1/\Gamma(\alpha j - k + 1)$ can be calculated as follows (Prodanov 2019) $1/\Gamma(-z) = -\frac{\sin \pi z}{\pi} \Gamma(z + 1)$. However, for further calculations, in R, there is a limitation when using '`factorial`' and '`gamma`' functions for larger values (see R Core Team 2020).

Hence, for further analysis, we adopt a hybrid formula: combining the recursive formula (2) with the tail asymptotic formula (3), with a change point for values bigger than pre-defined $K > 0$ between the formulas. In this paper we set the change point $K = 1000$.

2.1.3 Simulations

For simulations of discrete stable random variables, one can use the fact that discrete stable random variables can be represented as certain Poisson mixtures (e.g., Devroye 1993; Steutel and van Harn 1979, Corollary 6.8). A discrete stable random variable with parameters α, λ is distributed as a Poisson random variable with parameter $\lambda^{1/\alpha} S_\alpha$, where S_α is a positive stable random variable with support $[0, \infty)$ and tail index $\alpha \in (0, 1)$ Simulations of positive continuous stable random variables S_α can be performed using the R package `STABLE`® (Robust Analysis Inc 2017).

2.1.4 R package

The calculation of the probability generating function, probabilities and simulations are implemented in the R package `dstabledist` (Krutto 2023).

2.2 Estimating the parameters of discrete stable distributions

Modeling via discrete stable distribution means estimating its parameters α and λ. Because the probability mass function does not have a closed analytical form and not all moments exist, the popular maximum likelihood or

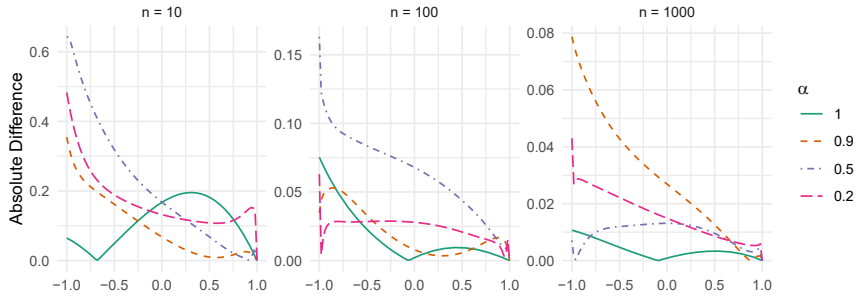

Figure 2: The absolute error between the empirical and theoretical probability generating function, $|\widehat{G}_n(z) - G(z)|, z \in [-1, 1]$, where $\widehat{G}_n$ is based on a sample from *DStable*$(\alpha, 1)$ with sizes $n = 10, 100, 1000$.

moment-based techniques are not directly applicable for discrete stable distributions. A popular approach is to use the empirical probability generating function (EPGF): let $\boldsymbol{X} = \{X_1, \ldots, X_n\}$ be a random sample from a probability distribution, then the empirical probability generating function G_n: $[-1, 1] \to [0, 1]$, is defined as

$$G_n(z) = \frac{1}{n}\sum_{i=1}^{n} z^{X_i}, \mid z \mid \leq 1. \tag{6}$$

To illustrate the behavior of the empirical probability generating function $G_n(z)$ compared to the probability generating function of discrete stable laws $G(z)$, as given by Equation (1), individual samples of discrete stable distributions are generated by simulation. For each sample, the absolute differences in the values of the theoretical and empirical functions, $|\widehat{G}_n(z) - G(z)|$ for $z \in [-1, 1]$, are presented in Figure 2.

By the strong law of large numbers, $G_n(z) \to G(z)$ almost surely as $n \to \infty$ for each $z \in [-1, 1]$. Estimation based on empirical probability generating functions has been suggested in Doray et al. (2009) and Slámová and Klebanov (2014). However, the former has shortcomings in terms of search procedures and the latter requires the estimation of additional parameters. In this paper, computationally straightforward closed-form estimators are used, proposed in Marcheselli et al. (2008). Fix two points $z_1, z_2 \in (-1, 1), z_1 \neq z_2$ and solve the corresponding system of equations of (1) for α and λ. Substituting $G(z)$ with $G_n(z)$ leads to point estimators for α and λ:

$$\alpha_n(z_1, z_2, \boldsymbol{X}) = \frac{\log(\log G_n(z_1)/\log G_n(z_2))}{\log((1 - z_1)/(1 - z_2))} \text{ and } \lambda_n(z_1, z_2, \boldsymbol{X}) = -\frac{\log G_n(z_1)}{(1 - z_1)^{\alpha_n}}. \tag{7}$$

The preference for estimators (7) is rooted in their advantageous characteristics: they offer closed-form simplicity and computational efficiency. Generally, these estimators (7) are versatile and applicable for evaluation in any $z_1, z_2 \in (-1, 1)$ where $z_1 \neq z_2$. Although the approach is attractive from a computational perspective, determining the choice of z_1, z_2 to achieve the best estimates remains an open question, analogous to related studies in continuous stable distributions (Krutto 2018; Lember and Krutto 2022). However, note that $G_n(-1) = \frac{1}{n}\sum_{i=1}^{n}(-1)^{X_i}$, $G_n(0) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i=0\}}$, and $G_n(1) = 1$ while referring to Equation (1), for discrete stable laws the probability generating function gives: $G(-1) = \mathrm{e}^{-\lambda 2^{\alpha}}$, $G(0) = \mathrm{e}^{-\lambda}$, and $G(1) = 1$. Consequently, at $z = 1$, we have $G_n(1) = G(1) = 1$. This can be interpreted as the values of $G_n(z)$ and $G(z)$ being closer as z approaches 1 (compared to other values of z in [−1,1]), a trend also observed in Figure 2. Hence, it is suggested (e.g., Marcheselli et al. 2008) to consider arguments $z_1 \neq z_2$ close to 1.

To identify the optimal choices of z_1, z_2, maximum likelihood calculations can be used. Consider a realization $\boldsymbol{k} = (k_1, \ldots, k_n)$ of a discrete random vector (e.g., miRNA-seq read counts) $\boldsymbol{X} = (X_1, \ldots, X_n)$, where X_i represents the random variable associated with observation *i*. In this context, the vector $\boldsymbol{k}$ represents the observed values, while $\boldsymbol{X}$ represents the corresponding random variables.[4] The likelihood function $L(\boldsymbol{\theta})$ is then given by

4 The use of *k* instead of *x* in the context of observed values is a common convention. In particular, *x* is often used to represent continuous random variables, while *k* is frequently used for discrete values, counts or categories.

$L(\boldsymbol{\theta} \mid \boldsymbol{k}) = \prod_{j=1}^{n} p_X(k_j \mid \boldsymbol{\theta})$, where $p_X(\boldsymbol{k} \mid \boldsymbol{\theta}) = \Pr_{\boldsymbol{\theta}}(X = k)$ and $\boldsymbol{\theta} = (\alpha, \lambda)$ represent the discrete stable distribution parameters. The maximum likelihood estimate $\widehat{\boldsymbol{\theta}}_n^*(z_1^*, z_2^*, \boldsymbol{k})$ is the maximizer of $L(\widehat{\boldsymbol{\theta}}_n(z_1, z_2) \mid \boldsymbol{k})$, where $\widehat{\boldsymbol{\theta}}_n(z_1, z_2, \boldsymbol{k}) = (\widehat{\alpha}_n(z_1, z_2, \boldsymbol{k}), \widehat{\lambda}_n(z_1, z_2, \boldsymbol{k}))$ is computed via formulas (7) at z_1, z_2 from a (sub)grid of $[0, 1) \times [-1, 1)$. Calculating $\Pr_{\boldsymbol{\theta}}(X = k)$, where $X \sim DStable(\alpha, \lambda)$, $\alpha \in (0, 1]$, and $\lambda > 0$, involves combining the recursive formula (2) with the tail asymptotic formula (3). For details, see Algorithm 1.

Algorithm 1. Parameter Estimation Algorithm

1: Choose a $z \in [0, 1)$ and construct a sufficiently dense grid $[z, 1) \times [z, 1)$.
2: For each pair z_1, z_2 in the grid $[z, 1) \times [z, 1)$ and the discrete random vector $\boldsymbol{k} = (k_1, \ldots, k_n)$:
3: Compute the parameter estimates $\widehat{\boldsymbol{\theta}}_n(z_1, z_2, \boldsymbol{k}) = (\widehat{\alpha}_n(z_1, z_2, \boldsymbol{k}), \widehat{\lambda}_n(z_1, z_2, \boldsymbol{k}))$ from Equations (7).
4: If $\widehat{\alpha}_n > 1$, set $\widehat{\alpha}_n = 1$.
5: Set the change point $K > 0$ between the recursive formula (2) and the asymptotic formula in the tail (3).
6: Calculate the likelihood function $L(\widehat{\boldsymbol{\theta}}_n(z_1, z_2) \mid \boldsymbol{k})$, where
7: **if** $k_i < K$ **then**
8: for $\Pr_{\boldsymbol{\theta}}(X = k_i)$ formula (2) is used.
9: **else**
10: for $\Pr_{\boldsymbol{\theta}}(X = k_i)$ formula (3) is used.
11: **end if**
12:
13: Choose the maximum likelihood estimate $\widehat{\boldsymbol{\theta}}_n^*(z_1^*, z_2^*, \boldsymbol{k})$ by maximizing $L(\widehat{\boldsymbol{\theta}}_n(z_1, z_2) \mid \boldsymbol{k})$, i.e.,
$$\widehat{\boldsymbol{\theta}}_n^*(z_1^*, z_2^*, \boldsymbol{k}) = \underset{z_1, z_2}{\operatorname{argmax}}\, L(\widehat{\boldsymbol{\theta}}_n(z_1, z_2) \mid \boldsymbol{k}).$$

In both of our applications in Sections 3 and 4, we use the grid $[0.5, 1) \times [0.5, 1)$ and set the change point $K = 1000$.

2.3 The goodness-of-fit of the Poisson, the negative binomial and the discrete stable distributions

The goodness-of-fit of the Poisson, the negative binomial and the discrete stable distributions is compared using Akaike Information Criteria (AIC). Recall that $\text{AIC} = -2l(\boldsymbol{\theta} \mid \mathbf{k}) + 2s$, where s is the number of parameters and $l(\boldsymbol{\theta} \mid \boldsymbol{k})$ is the log-likelihood function. Given $\boldsymbol{k} = (k_1, \ldots, k_n)$, the realization of a discrete random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$, the log-likelihood function is given by $l(\boldsymbol{\theta} \mid \boldsymbol{k}) = \sum_{j=1}^{n} \log p_X(k_j \mid \boldsymbol{\theta})$, where $p(k_j \mid \boldsymbol{\theta}) = \Pr_{\boldsymbol{\theta}}(X_j = k_j)$ and $\boldsymbol{\theta}$ is the vector of distribution parameters. For the discrete stable distribution, the values of $\Pr_{\boldsymbol{\theta}}(X_j = k_j)$ are calculated using the formulas (2) and (3). For the Poisson and negative binomial distributions, the log-likelihood functions are computed using the R package `stats`. When fitting the data, both the Poisson and the negative binomial distributions are fitted using the Moment Matching Estimation (MME), available in the R package `fitdistrplus` (Delignette-Muller and Dutang 2015). The moment method is chosen because of computational challenges encountered when running the Maximum Likelihood Estimation (MLE) method for the negative binomial distribution. However, note that for the Poisson distribution, the MLE and moment estimators yield the same results.

To further assess goodness-of-fit, we present QQ plots and estimate Cramer von Mises statistic ω^2 and Anderson–Darling statistics A (e.g., Stephens et al. 1986, pp. 100) for a selection of miRNAs. These are based on calculations performed on the cumulative distribution functions, for which the discrete stable distribution lacks a closed analytical form. Nevertheless, we can estimate the cumulative distribution function for the discrete stable distribution based on the probability mass function, calculated using the hybrid formula described in Section 1.1. We mention that for the Poisson and the negative binomial distributions, advanced forms of Cramér-von Mises and Anderson–Darling tests are implemented in the R package `goftest` (Faraway et al. 2021). However, to have comparable results for all, we calculate same basic statistics, which can be used as a measure of distance to assess fit. For the Cramér-von Mises w^2, we use:

$$n\omega^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left[\frac{2i-1}{2n} - F(X_{(i)})\right]^2, \tag{8}$$

and for the Anderson–Darling A, we use:

$$A^2 = -n - \sum_{i=1}^{n}\frac{2i-1}{n}\left[\ln(F(X_{(i)})) + \ln\left(1 - F(X_{(n+1-i)})\right)\right], \tag{9}$$

where $X_{(1)}, \ldots, X_{(n)}$ represent ordered data, and F is the estimated distribution function of the Poisson, negative binomial and discrete stable, respectively.

Caution is advised when interpreting the QQ plots and the Cramer-von Mises and Anderson–Darling statistics for the discrete stable distribution due to the lack of a closed-form solution.

2.4 Using discrete stable distributions for differential expression analysis

Differential expression analysis for miRNAs aims to identify miRNAs whose expression levels change notably under different experimental conditions, such as various cancer subtypes, healthy versus diseased states, or pre- and post-treatment stages. The objective is to understand the roles of these molecular regulators in biological processes, disease mechanisms, and as potential therapeutic targets. This analysis involves comparing miRNA expression profiles to discover specific miRNAs that might be crucial for driving biological changes. These changes can involve overexpression (increased expression levels) or underexpression (reduced expression levels), providing critical information on disease progression and treatment responses. The primary challenge in differential expression analysis is to distinguish genuine biological expression changes from the inherent variability of the underlying nature of miRNA expression. Extremely large values often observed in miRNA expression have traditionally been considered experimental noise, leading to their elimination through normalization and preprocessing procedures. However, this heavy-tailed nature might represent the underlying distribution rather than mere experimental noise, as discussed in the Introduction section. Therefore, the development of alternative methods for differential expression analysis directly handling raw data becomes crucial. At the outset, we highlight the availability of popular robust non-parametric tests, recommended for data containing extreme values (e.g., Staudte and Sheather 2011; Wilcox 2022). For example, the rank-based Brunner-Munzel test (e.g., Kume et al. 2017) for the stochastic equality of two populations. This test accommodates ties and does not assume equal variances between the groups. However, if the null hypothesis is rejected, it implies differences between distributions, but it provides limited insight into the nature of these differences. Additionally, the Moody's median test can be applied to test for differences in medians. Although the median is robust against heavy-tailed distributions, a limitation arises as our interest extends beyond merely determining median based central tendency. That is, the median serves as a singular metric, and no significant differences in medians does not necessarily imply that heavy-tailed distributions are identical or vice versa. However, we recommend using both tests as additional tools in the context of heavy-tailed differential expression analysis. In `R`, the Brunner-Munzel test can be applied by using the package `brunnermunzel` (Ara 2020) and the median test via the package `agricolae` (Felipe de Mendiburu 2023). In this paper, we propose that examining differences in the parameters α and λ of the discrete stable distribution could be more sensitive and suitable for discovering group differences, considering that miRNAs are derived from heavy-tailed distributions. For each miRNA, let $\widehat{\lambda}_n > 0$ and $\widehat{\alpha}_n \in (0, 1]$ be the estimates of the parameters of the discrete stable distribution, fitted to the raw expressions of the miRNA sequence belonging to a group of interest. We denote the (true) parameters of some group A by λ_A, α_A and corresponding estimates by $(\widehat{\lambda}_n)_A$, $(\widehat{\alpha}_n)_A$. To measure the difference between the parameter estimates, for each miRNA, calculate (1) the ratios of the parameter estimates of λ of the reference A to some comparison group B: $(\widehat{\lambda}_n)_A/(\widehat{\lambda}_n)_B$ and (2) the differences of the corresponding tail index estimates: $(\widehat{\alpha}_n)_A - (\widehat{\alpha}_n)_B$. (we do not use absolute value, as the sign indicates the direction of the tail heaviness. However, for illustrative purposes, the absolute difference may be used in some situations). This approach allows for a visual representation of the differences in estimates in high dimensions: assuming that the x-axis represents $(\widehat{\alpha}_n)_A - (\widehat{\alpha}_n)_B$ and the y-axis represents $(\widehat{\lambda}_n)_A/(\widehat{\lambda}_n)_B$. All miRNAs with minimal differences are located around zero. MiRNAs with similar tail

index estimates but differences in parameter lambda estimates align around a vertical line drawn from zero, parallel to the y-axis. In general, differences in both α and λ are of interest. Differences in the tail index α estimates indicate a possible difference in population tails, which affects the proneness to extreme values. However, in a differential expression analysis context, the parameter λ might be more interesting. Since λ impacts both the center and concentration of the primary mass of the distribution in discrete stable distributions (see the discussion of Figure 1 in Section 1.1), it might provide a more comprehensive understanding of differences between distributions across groups. However, as we see from formulas (7), differences in the tail index estimates α complicate the direct interpretation of the related difference (or lack thereof) in the estimates of λ. On the other hand, minor variations in the tail index estimates α indicate comparable tail characteristics, potentially facilitating a more straightforward comparison of the λ estimates. In practical terms, a difference of up to 0.1 could be considered small, indicating similar tail index. Importantly, as differences are identified from parameter estimates, the incorporation of a formal test proves to be beneficial. For similar tail indexes, differences in λ estimates might signal differences in typical patient expressions: higher λ estimates indicate a tendency toward elevated miRNA expressions by shifting and/or spreading the distribution's mass over a wider range.

Regarding the significance of differences in the parameter λ, a permutation test (e.g., Wilcox 2022) based on a statistic such as:

$$D = \left(\lambda_n^A - \lambda_n^B\right) \tag{10}$$

can be used to assess the significance of differences in λ between groups A and B by comparing the statistic computed from the original sample with those obtained from permuted samples. Permutation tests involve multiple random reassignments of observations, simulating the null hypothesis distribution to evaluate the statistical significance of observed differences. However, testing all possible permutations can be computationally intensive. Hence, in practice, random sampling from permutations is typically adopted. The corresponding p-value based on $i = 1, \ldots, n_{\text{perm}}$ values $D(i)$ is calculated as:

$$p_{\text{perm}} = \frac{\#|D(i)| > |D_o|}{n_{\text{perm}}}, \tag{11}$$

where $D(i)$ represents the difference based on the ith sample permutation, D_o is the observed value of the original data, and n_{perm} is the number of permutations.

3 Data and exploratory data analysis

This paper aims to explore the discrete probability distribution underlying miRNA-seq counts in a broader sense. Therefore, we analyze and model different types of miRNA data. First, we used the **sum expression for all reads aligned per miRNA** (quantification of miRNA expression) measured from malignant breast neoplasm (breast cancer primary tissue). Data were obtained from the database generated by The Cancer Genome Atlas (TCGA) Research Network for BReast CAncer (**TCGA-BRCA**), available at https://www.cancer.gov/tcga. TCGA-BRCA data consist of expressions of 1881 miRNAs, measured in primary breast tumor tissue samples $n = 1076$. For further analysis, the 400 miRNAs with the highest values of median absolute deviation were selected. Modeling was carried out on the expressions of all samples together ($n = 1076$) and on expressions stratified according to the PAM50 molecular subtypes of breast cancer: LumA ($n = 568$), LumB ($n = 202$), HeR2 ($n = 81$), Basal ($n = 186$) subtypes, and the normal-like subgroup ($n = 40$). The PAM50 classification list is available in `R/Bioconductor` package `TCGAbiolinks` (see Mounir et al. 2019). Second, we use the expressions of the **isoforms** (quantification of mature isoform expression) from blood samples with/without a later diagnosis of lung cancer. Data were obtained from the database of the Norwegian Women and Cancer Study for LUng CAncer (**NOWAC-LUCA data**). More information about the NOWAC study is available at https://site.uit.no/nowac/ and an exhaustive cohort overview is given in (Lund et al. 2007). NOWAC-LUCA data consist of 2013 miRNA expressions, measured in $n = 240$ blood samples. The 198 miRNAs for which there were a maximum of five samples with no detected reads were kept in the analyzes. Modeling was performed on expressions from all samples together ($n = 240$), and

expressions from samples with a later diagnosis of lung cancer (case, $n = 124$) compared to data from matched samples without cancer diagnosis (control, $n = 116$).

3.1 Exploratory data analysis for heavy tails

The summary of some classic and robust statistics for the TCGA-BRCA data is given in Table 2 and for the NOWAC-LUCA data in Table 3.

Table 2: TCGA-BRCA: for $p = 400$ miRNAs, the summary of summary statistics for data from all samples ($n = 1076$).

Measure	Min.	1st qu.	Median	Mean	3rd qu.	Max.
			Classic measures			
Minimum	0	0	9	557	93	36,292
Maximum	245	1880	7934	175,963	39,582	5,736,079
Range	245	1879	7892	175,405	39,432	5,699,787
Mean	21	104	449	14,131	2739	914,977
St.Dev.	25	150	594	16,135	3479	822,512
Skewness	1.9	3.2	4.3	6.1	7.0	22.7
Ex.Kurt.	4.8	15.5	30.1	82.7	82.1	622.4
			Robust measures			
Median	1	5	40	5703	590	648,067
IQR	20	84	370	12,161	2081	736,563
MAD	1	4	33	4727	524	490,266
Medcouple	0.00	0.33	0.37	0.39	0.44	0.86
			Comparison with Poisson distribution			
Var/mean	2.5	30.9	220.7	13,088.7	1832.3	739,392.0
Mean·Skw^2	11	437	3660	162,332	24,203	12,989,100
Mean·Ex.Kurt	19	693	5606	242,156	38,466	15,638,396

Table 3: NOWAC-LUCA: for $p = 198$ miRNAs, the summary of summary statistics for data from all samples ($n = 240$).

Measure	Min.	1st qu.	Median	Mean	3rd qu.	Max.
			Classic measures			
Minimum	0	0	0	2	1	84
Maximum	145	1396	3772	38,069	19,646	1,281,836
Mean	28	120	289	3359	1558	109,551
St.Dev.	24	151	382	3963	1963	157,626
Ex.Kurt.	4.4	18.4	29.8	43.8	60.0	158.3
Skewness	1.9	3.6	4.7	5.3	6.4	11.7
			Robust measures			
Median	19	81	201	2292	1175	61,156
IQR	20	80	192	2244	1087	77,908
MAD	13	55	134	1420	682	44,374
Medcouple	0.05	0.27	0.32	0.32	0.37	0.51
			Comparison with Poisson distribution			
Var/mean	22	156	467	4947	2566	226,799
Mean·Skw^2	114	2470	8652	94,522	45,277	2,492,023
Mean·Ex.Kurt	136	3513	11,333	126,248	64,386	3,036,658

Both Tables 2 and 3 show misalignment between the classic and robust central tendency and dispersion measures, and both tables suggest that the underlying probability distribution might be skewed to the right and heavy tailed. We point out that even if the theoretical mean does not exist due to the heavy-tailed nature of the distribution, the sample average (sample mean), as presented in Tables 2 and 3, can still be calculated from the observed data. However, in this case, the sample mean will not serve as an estimate of the non-existent theoretical mean. With regard to the popular assumption that miRNAs follow a Poisson distribution, we also draw attention to the inconsistency with this distribution. Theoretically, for the Poisson distribution, the variance-to-mean ratio, the product of mean and skewness squared, and the product of mean and excess kurtosis are all equal to 1. A summary of these measures for the TCGA-BRCA data is given in Table 2 and, correspondingly, for the NOWAC-LUCA data in Table 3. From both Tables 2 and 3, it follows that the miRNA-seq count data are overdispersed, overskewed, and have heavier tails, relative to the Poisson distribution. However, note that sample skewness and excess kurtosis can be severely biased in finite samples (e.g., Joanes and Gill 1998), so these results should be treated with care.

The importance of exploratory data analysis to understand heavy-tailed distributions and the properties of underlying discrete probability distributions is highly emphasized; However, the identification of heavy tails often lacks formal criteria and relies instead on visual inspection (e.g., Embrechts et al. 2013, Chapter 6). Common methods in exploratory data analysis for heavy tails, such as probability and QQ plots, the mean excess function plots, and Gumbel's method of exceedance, tend to be better suited for a single-variable or few-variables analysis (e.g., Embrechts et al. 2013, Section 6.2), and might become impractical in high-dimensional settings. A simple tool for exploring heavy tails and the finiteness of moments is the ratio of maximum to sum (e.g., Embrechts et al. 2013, Section 6.2.6). Suppose $X_1, \dots, X_n$ are iid random variables and for any positive r define the quantity $R_n(r) = M_n(r)/S_n(r)$, where $S_n(r) = \mid X_1 \mid^r + \cdots + \mid X_n \mid^r, M_n(r) = \max\left(\mid X_1 \mid^r, \dots, \mid X_n \mid^r\right), r > 0, n \geq 1$. The following equivalence holds: $R_n(r) \to^{a.s.} 0$ if and only if $E \mid X \mid^r < \infty$. Therefore, $R_n(r)$ should be small for large n provided that $E \mid X \mid^r < \infty$, otherwise it indicates heavy tails. The advantage of using $R_n(r)$ is its ability to visually combine many variables into a plot, which is especially useful in high-dimensional data analysis. For $r = 1, 2, 3, 4$, the ratios of maximum to sum $R_n(r)$ for the TCGA-BRCA data are presented in Figure A.1-F.1 and for the NOWAC-LUCA data in Figure A.3-F.1. From both Figures A.1-F.1 and A.3-F.1, the results suggest that most of the analyzed miRNAs may not have moments of order $r = 1, 2, 3, 4$. In summary, exploratory data analysis for heavy tails of miRNA expressions indicates that the underlying distributions are likely to be heavy tailed.

4 Results from the TCGA-BRCA data

In this section, the results of the analysis of the data obtained from the database generated by the TCGA-BRCA Research Network are presented.

4.1 TCGA-BRCA: estimates of discrete stable distribution parameters

The discrete stable distributions were fitted to $p = 400$ miRNAs, both for data from all samples (all, $n = 1076$) and for data from the molecular subtypes, normal-like (Normal, $n = 40$), luminal A (LumA, $n = 568$), luminal B (LumB, $n = 202$), basal-like (Basal, $n = 185$), HER2-amplified (positive) (Her2, $n = 81$). A summary of the corresponding parameter estimates $\widehat{\alpha}_n$ and $\widehat{\lambda}_n$ is given in Table 4. Table 4 displays estimates derived from various samples. 'All samples' encompass all subtypes together, while samples from specific subtypes are selected as subsets. The parameters are then estimated separately for each of these subsets.

From Table 4, the tail index α of each of the $p = 400$ miRNAs was estimated to be less than 1, which means that the underlying distribution was estimated to be heavy tailed. In particular, for data from all samples, half of the miRNAs are estimated to have a tail index less than 0.65. Focusing on the subgroups, also here all tail index estimates was less than 1. However, the estimates are slightly larger (i.e., slightly lighter tails) than the ones for all data. Estimates of the parameter λ range from almost zero to hundreds of thousands, both for samples from all data and from molecular subtypes. When comparing the subtypes, the λ estimates are similar to those for all

Table 4: TCGA-BRCA: summary of parameter estimates $\widehat{\alpha}_n$ and $\widehat{\lambda}_n$ for the discrete stable distribution for raw read counts of $p = 400$ miRNAs; data from all samples and samples of molecular subtypes.

	Min.	1st qu.	Median	Mean	3rd qu.	Max.
			Summary of $\hat{\alpha}_n$ estimated from all samples			
All samples	0.1513	0.5988	0.6577	0.6402	0.7104	0.9957
			Summary of $(\hat{\alpha}_n)_{\text{Subtype}}$ estimated from molecular subtypes			
Normal-like	0.3105	0.6394	0.7182	0.7312	0.8351	0.9939
LumA	0.1860	0.6118	0.7063	0.7071	0.8533	0.9820
LumB	0.1419	0.6468	0.7319	0.7326	0.8625	0.9810
Basal	0.1979	0.6403	0.7136	0.7177	0.8268	0.9883
Her2	0.2066	0.6697	0.7411	0.7446	0.8691	0.9862
			Summary of $\hat{\lambda}_n$ estimated from all samples			
All samples	0.73	3.06	9.71	2853.87	35.52	224,116
			Summary of $(\hat{\lambda}_n)_{\text{Subtype}}$ estimated from molecular subtypes			
Normal-like	0.75	4.59	11.90	843.43	50.52	187,481
LumA	0.60	4.44	11.56	130.27	35.05	61,281
LumB	0.50	4.06	10.35	621.34	43.14	1,689,641
Basal	0.63	4.28	10.95	223.48	38.88	11,680
Her2	0.63	4.61	13.03	905.24	48.81	229,910

samples, except for the mean and maximum. Indeed, the maximum of the λ estimates is affected by assigning data entries to subgroups, which accordingly affects the mean. However, in all groups, the median of $\widehat{\lambda}_n$ is around 10, while at least a quarter of λ estimates are more than 50. More analysis of the differences between the groups, at the miRNA level, is given in Sections 3.3 and 3.4.

4.2 TCGA-BRCA: comparison of the goodness-of-fit

The goodness-of-fit of the Poisson, negative binomial, and discrete stable distributions are compared using AIC. For comparison, for every miRNA, the AIC ratio of the estimated discrete stable to the estimated Poisson distribution was evaluated, and the corresponding ratio of the estimated discrete stable to the estimated negative binomial was evaluated. To summarize these results, box plots of the ratios are presented in Figure 3.

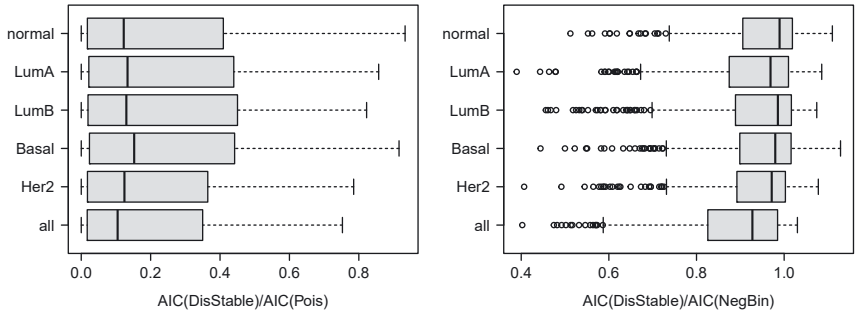

Figure 3: TCGA-BRCA: for $p = 400$ miRNAs, the ratios of AICs of discrete stable models to Poisson and to negative binomial for data from all samples ($n = 1076$) and from samples of molecular subtypes, normal-like ($n = 40$), LumA ($n = 568$), LumB ($n = 202$), Basal ($n = 185$), Her2 ($n = 81$).

The ratios are given both for the models fitted to the data from all samples together (labelled 'all') and for data from the molecular subtypes (labelled correspondingly). If the ratio is less than 1, then the discrete stable distribution gives a better fit. The smaller the ratio, the better the fit of the discrete stable distribution (in comparison to the Poisson and negative binomial distribution). From the left side of Figure 3, it can be seen that for all miRNAs, the discrete stable distribution gives a much better fit than the Poisson, both for the data from all samples together and for the data from the molecular subgroups. In all samples, for around 200 miRNAs the AIC of the discrete stable distribution is approximately 90 percent reduced, and for around 300 miRNAs the values are at least 50 percent reduced. For no miRNA, the Poisson distribution was a better fit than the discrete stable distribution. From the right side of Figure 3, it follows that for all samples together, the discrete stable distribution gives a better fit than the negative binomial for around 344 (86 %) of the miRNAs. For the remaining 56 (14 %) miRNAs, the ratios are close to 1, meaning that the goodness-of-fit is similar. For data from the LumA, LumB, Basal, Her2 and normal-like subtypes, the discrete stable distribution gives a better fit than negative binomial for 58 %, 69 %, 63 %, 73 % and 58 % of the miRNAs, respectively. For a more thorough investigation of goodness-of-fit for a selected set of miRNAs, see Section 3.4.

4.3 TCGA-BRCA: differential expression via discrete stable distributions

In this section, differential expression analysis is performed through estimates of the parameters of discrete stable distributions within the molecular subtypes of breast cancer. In particular, the estimates for the subtypes LumA ($n = 568$), LumB ($n = 202$), HeR2 ($n = 81$) and Basal ($n = 186$) are compared with the ones for the normal-like subgroup ($n = 40$). The general summary of the summed expression for all reads aligned per 400 miRNA (measured in primary breast cancer tissue) is given in Table 5.

Certain tendencies can be observed from Table 5. For example, the minimum expression in each group is 0, and the quartiles are also broadly similar. However, for all quartiles, the values of the other subtypes are slightly lower (i.e., underexpressed) compared to the normal-like subgroup, which is in agreement with previous findings (e.g., Fontana et al. 2021, and references therein.) Extremely large maximum values are present in all groups, but no obvious differences appear.

The general summary of the parameter estimates of the fitted discrete stable distributions of samples of molecular subtypes is given in Table 4. Next, for each miRNA the differential expression analysis was performed between normal-like and the other molecular subtypes, as explained in Section 1.4.

For each miRNA, we calculated (1) the ratios of estimates for the parameter λ of the normal-like subgroup and the other subgroups and (2) the differences in estimates for the tail index λ of the normal-like subgroup and the other subgroups. For each miRNA, both calculations are plotted in Figure 4.

Note that in Figure 4, the ratios of the estimates of λ larger than 1 represent underexpressions, while the ratios less than 1 represent overexpressions, compared to the normal-like subgroup.

In Figure 4, compared to the normal-like subgroup, the five miRNAs that are most underexpressed (irrespective of differences in the tail index α estimates) in each subtype are as follows:

- LumA: hsa-mir-21, hsa-mir-10a, hsa-mir-199b, hsa-mir-199a-2, hsa-mir-182;
- LumB: hsa-mir-143, hsa-mir-21, hsa-mir-199b, hsa-mir-93, hsa-mir-199a-2;

Table 5: TCGA-BRCA: for the raw read counts of $p = 400$ miRNAs, for data from the molecular subtypes, the summary of descriptive statistics.

	n	Min.	1st qu.	Median	Mean	3rd. qu.	Max.
Normal-like	40	0	7	53	10,617	706	3,961,021
LumA	568	0	6	43	10,280	609	5,736,079
LumB	202	0	5	38	6897	516	3,941,286
Basal	185	0	6	44	6679	620	4,218,093
Her2	81	0	6	45	7608	594	4,536,709

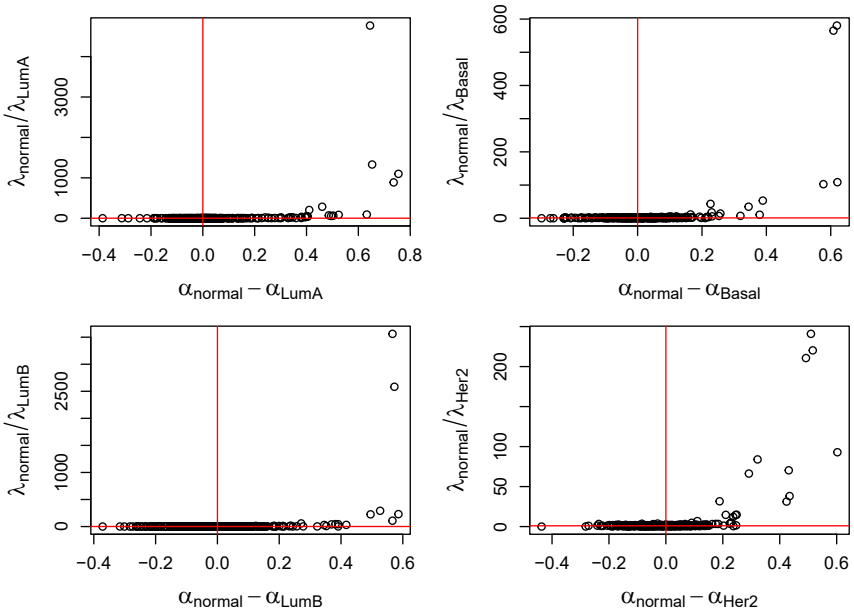

Figure 4: TCGA-BRCA: for $p = 400$ miRNAs, the ratios of estimates for parameter λ and the differences in estimates for the tail index α, between the normal-like subgroup and the other breast cancer subgroups, labeled correspondingly.

- Basal: hsa-mir-199a-2, hsa-mir-199b, hsa-mir-199a-1, hsa-mir-126, hsa-mir-100;
- Her2: hsa-mir-199b, hsa-mir-183, hsa-mir-199a-2, hsa-mir-151a, hsa-mir-182.

Although the aforementioned miRNAs have the highest ratios of λ estimates, they also have large differences in the tail index estimates, making comparison of the estimates of λ less straightforward. As such, we can conclude that for these miRNAs, the discrete stable distribution parameters were estimated to be different. However, as explained in Section 1.4, in Figure 4 we are even more interested in the cases where the difference in estimates for the tail index α is relatively small, while the corresponding ratio of estimates for the parameter λ is different from one. In that case, the interpretation of λ differences is more straightforward. For that purpose, a zoom into Figure 4, where only the part with difference in tail index α estimates between the groups of interest is less than 0.1, is presented in Figure 5. In Figure 5, the ratios within the LumA subtype appear to fluctuate less around the value of 1 than within other subtypes. This agrees with previous studies in which miRNA expressions of the LumA subtype have been found to be similar to the normal-like subgroup (e.g., Fontana et al. 2021).

The ratios of λ estimates that stand out are all larger than 1, representing underexpressions compared to the normal-like subgroup. The top 10 ratios (larger than 1) from Figure 5, with the corresponding estimated differences in tail index, are given in Table 6.

In particular, in Table 6 the miRNAs with ratios larger than 3.5 (i.e., most underexpressed), are as follows: (1) for LumA: none; (2) LumB: hsa-mir-1262, hsa-mir-202, hsa-mir-138-1; (3) Basal: hsa-mir-26a-1, hsa-mir-26a-2, hsa-mir-29c, hsa-mir-202, hsa-mir-101-1; (4) Her2: hsa-mir-135a-2, hsa-mir-202, hsa-mir-93.

For all the miRNAs that stand out with especially large differences in the discrete stable distribution parameter estimates, a further analysis needs to be done for exploring them as potential biomarkers for the (molecular subtypes of) breast cancer. As an example, for the miR-200 family, a more detailed analysis with robust testing is provided in Section 3.4.

4.4 TCGA-BRCA miR-200: an illustrative example for the miR-200 family

The miR-200 family is one of the most frequent groups of miRNAs whose expression is altered in (breast) cancer (e.g., Cavallari et al. 2021; Fontana et al. 2021; Wen et al. 2021; Ye et al. 2014). The family consists of five miRNAs:

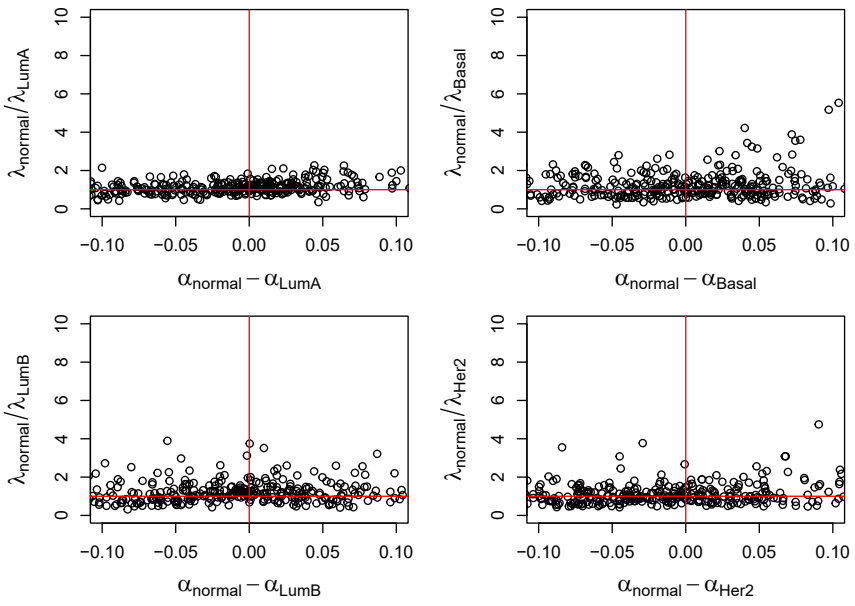

Figure 5: TCGA-BRCA: a zoom to Figure 4, where the absolute differences in estimates for tail index α is less than 0.1.

miR-141, miR-200a, miR-200b, miR-200c and miR-429. Despite comprehensive studies, the potential role of these miRNAs as biomarkers in cancer has not yet been completely understood. In this section, an illustrative example of differential expression analysis via discrete stable distributions is given for the miR-200 family for the PAM50 breast cancer molecular subtypes normal-like, LumA, LumB, Her, and Basal.

4.4.1 TCGA-BRCA miR-200: modelling via discrete stable, Poisson and negative binomial distribution

We start out by assessing goodness-of-fit. We present QQ plots in Figures A.2-A.2-F.1–F.5, which empirically confirm that the discrete stable distribution fits the miR200 family well, with an exception for miR429, where none of the distributions seem to fit very well. Additionally, we estimate Cramer von Mises and Anderson–Darling statistics (distances), see formulas (8) and (9). Results are given in Table 7.

The Poisson distribution does not provide a good fit based on either statistic. We note that compared with the Cramer-von Mises distance, the Anderson–Darling distance assigns more weight to observations in the tails of the distribution, making it more suitable for contexts with heavy tails. For both statistics, the discrete stable distribution shows the closest fit for miR200a, miR200b, and miR429, while for the two others, the distances are slightly in favor of the negative binomial distribution. However, caution is advised when interpreting these measures for the discrete stable distribution due to the numerical estimation of the cumulative distribution function. We have also obtained statistics for subtypes, which yielded similar results, alternately favoring both negative binomial and discrete stable distributions (results not presented).

The estimates of the discrete stable distribution parameters, the corresponding value of AIC, and the values of AIC for the Poisson and negative binomial distributions, are given in Table 8. These results highlight that every member of the miR-200 family is estimated to be heavy-tailed (i.e., $(\widehat{\alpha}_n)_{\text{subtype}} < 1$). From Table 8, for each miR-200 family member, the discrete stable distribution provides a better fit than the Poisson distribution; for miR-200a, miR-200c, miR-141 the discrete stable distribution provides a better fit than the negative binomial distribution; for miR-200b, miR-429, the discrete stable distribution provides a similar fit compared to the negative binomial.

Table 6: TCGA-BRCA: from Figure 5, the list of miRNAs with the top 10 ratios of parameter λ estimates.

LumA	$(\widehat{\alpha}_n)_{\text{normal}} - (\widehat{\alpha}_n)_{\text{LumA}}$	$(\widehat{\lambda}_n)_{\text{normal}}/(\widehat{\lambda}_n)_{\text{LumA}}$
hsa-mir-129-1	0.04	2.27
hsa-mir-26a-1	0.06	2.26
hsa-mir-129-2	0.02	2.11
hsa-mir-143	0.02	2.09
hsa-mir-378a	0.05	2.03
hsa-mir-542	0.04	2.03
hsa-mir-17	0.06	1.95
hsa-mir-218-2	0.05	1.94
hsa-mir-19b-2	0.07	1.89
hsa-mir-26b	0.1	1.89
LumB	$(\widehat{\alpha}_n)_{\text{normal}} - (\widehat{\alpha}_n)_{\text{LumB}}$	$(\widehat{\lambda}_n)_{\text{normal}}/(\widehat{\lambda}_n)_{\text{LumB}}$
hsa-mir-138-1	−0.06	3.89
hsa-mir-202	0	3.75
hsa-mir-1262	0.01	3.52
hsa-mir-150	0.09	3.21
hsa-mir-1258	0	3.12
hsa-mir-1295a	−0.05	2.97
hsa-mir-1304	−0.1	2.72
hsa-mir-3926-2	−0.02	2.62
hsa-mir-125b-1	0.03	2.6
hsa-mir-129-2	0.06	2.59
Basal	$(\widehat{\alpha}_n)_{\text{normal}} - (\widehat{\alpha}_n)_{\text{Basal}}$	$(\widehat{\lambda}_n)_{\text{normal}}/(\widehat{\lambda}_n)_{\text{Basal}}$
hsa-mir-101-1	0.1	5.17
hsa-mir-202	0.04	4.22
hsa-mir-29c	0.07	3.88
hsa-mir-26a-2	0.08	3.6
hsa-mir-26a-1	0.07	3.55
hsa-mir-10a	0.04	3.44
hsa-mir-143	0.04	3.24
hsa-mir-944	0.05	3.15
hsa-mir-135a-2	0.02	3.01
hsa-mir-30d	0.07	2.83
Her2	$(\widehat{\alpha}_n)_{\text{normal}} - (\widehat{\alpha}_n)_{\text{Her2}}$	$(\widehat{\lambda}_n)_{\text{normal}}/(\widehat{\lambda}_n)_{\text{Her2}}$
hsa-mir-93	0.09	4.75
hsa-mir-202	−0.03	3.78
hsa-mir-135a-2	−0.08	3.55
hsa-mir-26a-2	0.07	3.08
hsa-mir-26a-1	0.07	3.08
hsa-mir-135a-1	−0.04	3.08
hsa-mir-143	0	2.67
hsa-let-7b	0.05	2.56
hsa-mir-3926-2	−0.04	2.45
hsa-mir-92a-2	0.06	2.32

4.4.2 TCGA-BRCA miR200: differential expression analysis via discrete stable distribution

In this section, we explore the differences in discrete stable estimates between the breast cancer molecular subtypes for the miR-200 family. Referring to the summary in Table A.2-T.1, we observe differential expression patterns between the normal-like subtype and other subtypes for each member of the miR-200 family. In Table 8,

Table 7: Goodness-of-fit statistics: Cramér-von Mises statistic ($n\omega^2$) and Anderson–Darling statistic (A).

miRNA	Poisson		Negative binomial		Discrete stable	
	A	$n\omega^2$	A	$n\omega^2$	A	$n\omega^2$
miR200a	445	113	23	79	20	28
miR200b	444	118	23	81	9	9
miR200c	594	115	21	62	27	116
miR141	481	113	21	63	27	97
miR429	245	111	24	91	14	38

Table 8: TCGA-BRCA: Estimates for discrete stable distribution parameters (($\widehat{\alpha}_n$), ($\widehat{\lambda}_n$)), AIC values for discrete stable (AIC.DS), Poisson (AIC.Pois), and negative binomial (AIC.NB) distributions, differences Δ in tail index α and parameter λ, along with Brunner-Munzel test (p_{BM}) and permutation test (p_{perm}) results comparing normal-like ($n = 40$) with LumA ($n = 568$), LumB ($n = 202$), Basal ($n = 186$), and HER2-amplified ($n = 81$) subgroups.

	$(\widehat{\alpha}_n)$	$(\widehat{\lambda}_n)$	AIC.DS	AIC.Pois	AIC.NB	$\Delta(\widehat{\alpha}_n)$	$\Delta(\widehat{\lambda}_n)$	p_{BM}	p_{perm}
miR200a									
Normal-like	0.32	18	687	69,289	710				
LumA	0.30	25	9116	1,497,011	10,527	0.02	−7	0.003	0.148
LumB	0.32	20	3452	470,220	3650	0.00	−2	0.461	0.256
Basal	0.31	22	3053	405,834	3369	0.01	−4	0.089	0.265
Her2	0.33	17	1439	128,918	1440	−0.01	1	0.422	0.262
miR200b									
Normal-like	0.32	16	711	68,805	705				
LumA	0.31	22	9413	1,289,291	10,258	0.01	−6	0.192	0.871
LumB	0.32	18	3422	361,738	3530	0.00	−2	0.351	0.516
Basal	0.31	21	3102	386,395	3299	0.01	−5	0.671	0.497
Her2	0.34	16	1425	113,739	1388	−0.02	0	0.223	0.43
miR200c									
Normal-like	0.78	2598	817	908,873	918				
LumA	0.78	3096	11,857	15,026,637	13,233	0.00	−498	0.074	0.583
LumB	0.83	4764	3997	4,905,796	4661	0.05	−2166	0.612	0.435
Basal	0.77	2728	3870	5,445,207	4309	0.01	−130	0.314	0.408
Her2	0.85	5347	1562	1,382,181	1832	0.07	−2749	0.944	0.160
miR141									
Normal-like	0.32	22	718	121,047	756				
LumA	0.38	50	9342	1,960,122	11,029	0.06	−28	0.002	0.025
LumB	0.45	77	3727	735,291	3916	0.13	−55	0.021	0.562
Basal	0.56	146	3490	840,206	3635	0.24	−124	0.013	0.373
Her2	0.60	199	1482	229,638	1546	0.28	−177	0.123	0.240
miR429									
Normal-like	0.50	14	575	8191	541				
LumA	0.60	31	8293	214,501	8159	0.10	−17	0.041	0.289
LumB	0.55	24	2944	60,454	2812	0.05	−10	0.471	0.676
Basal	0.57	29	2813	78,044	2696	0.07	−15	0.012	0.300
Her2	0.47	15	1217	35,756	1154	0.03	−1	0.411	0.247

you can find differences in tail index α and parameter λ, as well as Brunner-Munzel test and permutation test results in addition to parameter estimates and AIC values.

From Table 8, the following can be observed:

- For both miR-200a and miR-200b, the tail index α estimates for the LumA, LumB, Basal and Her2 subgroup are varying ± 0.02, meaning that the tails are estimated to be quite similar compared to the tail index estimate for the normal-like subgroup. From this, the parameter λ estimates are rather straightforward to compare. For both miR-200a and miR-200b, the parameter λ estimates for the LumA, LumB, Basal and Her2 subgroups are the same or slightly bigger than for the normal-like subgroup, indicating possible overexpression in these subtypes compared to the normal-like. For miR-200a, the Brunner-Munzel test identifies significant differences in distributions between the normal-like and LumA subtype ($p = 0.003$), and close to significant between normal-like and the Basal subtype ($p = 0.089$). For miR-200b, the Brunner-Munzel test did not identify any significant differences in distributions between the normal-like and the other subtypes.
- For miR-200c, compared to the normal-like subgroup, the tail index α estimates for the LumA subgroup is the same, similar for Basal and slightly bigger for LumB and Her2 (differences $+0$, $+0.01$, $+0.05$, $+0.07$, respectively). Therefor, for LumA and Basal the estimates for λ are straightforward to compare and both indicate potential overexpression. For LumB and Her2 subtypes, the estimates for λ are not as straightforward to compare, but dismissing the differences in tail estimates, both are indicating potential overexpression. The Brunner-Munzel test identifies close to significant differences only between the normal-like and LumA subtype ($p = 0.074$).
- For miR141, the estimates of the tail index α for the LumA, LumB, Basal and Her2 are most varying when comparing to the normal-like subtype (differences $+0.06$, $+0.13$, $+0.24$, $+0.28$, respectively). A larger tail index estimate defines a lighter tail, meaning very small or very large values are less likely. The λ estimates for the LumA, LumB, Basal, and Her2 subgroups range from 2 to almost 10 times larger than the λ estimates for the normal subgroup. Dismissing the differences in the tail estimates, the λ estimates for each of the subtypes seem to suggest potential overexpression compared to the normal-like. The Brunner-Munzel test shows significant differences in distributions between the normal-like and LumA subtype ($p = 0.002$), LumB subtype ($p = 0.021$), and the Basal subtype ($p = 0.013$).
- For miR429, the tail index α estimates for the LumA, LimB, Basal and Her2 are similar to the tail index estimate in the normal-like subtype (differences $+0.1$, $+0.05$, $+0.07$, -0.03, respectively). For all the subtypes, the λ estimate is indicating potential overexpression as compared to normal-like. The Brunner-Munzel test shows significant differences in distributions between the normal-like ($p = 0.041$) and Basal subtype ($p = 0.012$).

For further insight, Moody's median test was performed on the raw expressions, which examines whether samples originate from populations with same medians, results are given in Table A.2-T.2. In addition, *t*-test on the log2-transformed expressions was performed, results given in Table A.2-T.3, which examines whether samples originate from populations with same geometric mean.

The final conclusions for differential expressions, based on differences in the estimates of α less than 0.1, the values of the estimates of λ, the Brunner-Munzel test (Table 8), and the Moody's median test (Table A.2-T.2), are as follows: Compared to the normal-like subgroup, miR200a, miR200c, miR141 were found to have elevated expressions in the LumA subtype; miR200a was found to have elevated expressions in the Basal subtype.

As the final step, we performed permutation tests based on Equation (11), where the original value $D_o = \Delta(\widehat{\lambda}) = (\widehat{\lambda}_n)_{\text{normal}} - (\widehat{\lambda}_n)_{\text{subtype}}$ is found in Table 8. The resulting p-values are given in Table 8, column p_{perm}. From Table 8, only for miR141 the difference in estimates of λ is found to be significant between normal-like and LumA subtypes. However, the permutation test does involve differences in tail index and should be treated with care for final conclusions.

Another study of miR-200 TCGA-BRCA data (Fontana et al. 2021) found that miR-141-3p was underexpressed in the normal-like subtype, miR-200a-3p was underexpressed in the Her2 and LumB subtypes, and miR-200b-3p underexpressed in the Her2 subtype. However, for analysis they applied an ANOVA test on log-transformed

data; for data they used the mature isoforms expressions (not the summed expression for all reads aligned per miRNA) and applied a slightly different selection of samples.

5 Results from the NOWAC-LUCA data

To show the relevance of the discrete stable distribution also for other types of data, in this section, the results of a corresponding analysis of the mature isoforms expressions (the mature isoform expression quantification) measured in blood samples, obtained from the NOWAC-LUCA study database, are presented.

5.1 NOWAC-LUCA: estimates of discrete stable distribution parameters

Discrete stable distributions were fitted to 198 variables of miRNA expressions for data from all samples ($n = 240$) and to data from 'case' ($n = 124$) and 'control' ($n = 116$) subgroups. The summary of parameter estimates of the discrete stable distribution is given in Table 9.

For all miRNAs, the underlying distributions were estimated to be heavy tailed (i.e., $\widehat{\alpha}_n < 1$). This is true for all samples together, and when stratified by case-control status. In particular, for 50 % of miRNAS from all samples tail index is estimated less than 0.59. The estimates of λ for the data from all samples vary from 6 to 66 while around half are between 8 and 25. The minimum and 1st quartile of λ estimates for the data from cases and controls are similar while median, mean, 3rd quartile and maximum are quite different in comparison to the ones based on all samples.

5.2 NOWAC-LUCA: comparison of the goodness-of-fit

The goodness-of-fit of the Poisson, the negative binomial and the discrete stable distributions are compared using AIC. To compare goodness-of-fit, the ratios of corresponding AIC values were evaluated as before. If the ratio is less than 1, then the discrete stable distribution gives better fit. The ratios for $p = 198$ miRNAs are presented in Figure 6.

From the left side of Figure 6 it can be seen that the discrete stable distribution was a better fit than the Poisson distribution for all miRNAs. In fact, more than half of the miRNAs have an AIC value for the Poisson distribution more than 10 times bigger than those for the discrete stable distributions. From the right side of

Table 9: NOWAC-LUCA: for $p = 198$ miRNAs, the summary of parameter estimates of discrete stable distributions for data from all samples together ($n = 1076$) and from samples of cases ($n = 124$) and controls ($n = 116$), respectively.

	Min.	1st qu.	Median	Mean	3rd qu.	Max.
Summary of $\widehat{\alpha}_n$: estimates from all samples						
All samples	0.04164	0.32339	0.59210	0.48411	0.67225	0.79630
Summary of $(\widehat{\alpha}_n)_{\text{Group}}$: estimates from case/control groups						
Case	0.0732	0.3246	0.5458	0.4573	0.6293	0.7666
Control	0.0657	0.3352	0.6830	0.5457	0.7304	0.8291
Summary of $\widehat{\lambda}_n$: estimates from all samples						
All samples	5.577	8.698	16.79	18.09	25.21	66.13
Summary of $(\widehat{\lambda}_n)_{\text{Group}}$: estimates from case/control groups						
Case	5.54	8.367	14.29	14.71	18.70	48.22
Control	6.92	9.17	21.14	34.20	39.71	347.97

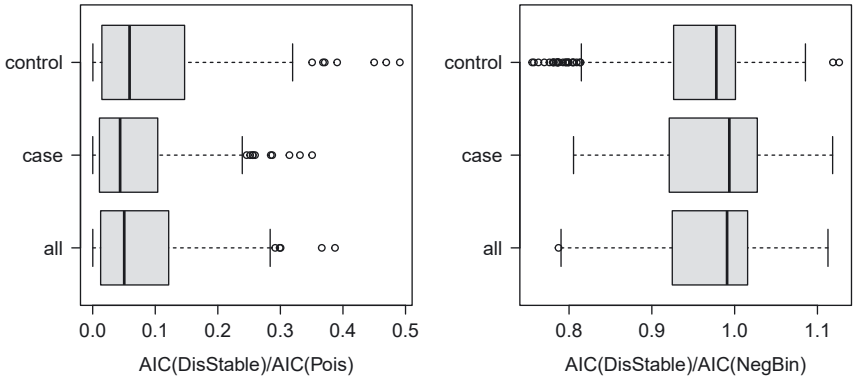

Figure 6: NOWAC-LUCA: for $p = 198$ miRNAs, the ratios of AICs of discrete stable models to Poisson and negative binomial models from all samples ($n = 240$) and from samples grouped to case ($n = 116$) and control ($n = 124$).

Figure 6, for the data from 'all', 'control' and 'case' samples, the discrete stable distribution give a better fit than the negative binomial distribution for 58 %, 73 %, 54 % of the miRNAs, respectively. For the remaining miRNAs on the right side of Figure 6, the goodness-of-fit is similar.

5.3 NOWAC-LUCA: differential expression via discrete stable distribution

In this section, the differential expression analysis via the estimates of the parameters of discrete stable distributions for the $p = 198$ miRNA read counts is performed. First, the summary of data is given in Table 10. From this table, the quantiles of reads of the data from case and control samples are quite similar, except for the maximums.

The summary of the discrete stable distribution parameter estimates is given in Table 9. Although the quartiles of the estimates of α are quite similar, they hint that the tail index for the data from the 'case' samples are slightly smaller than for the data from the 'control' samples, that is, the miRNA-seq counts in the 'case' group have heavier tails. The quartiles of estimates of λ are lower (indicating underexpression) for the data from the 'case' samples, especially the 3rd quartiles and maximums. Next, for each miRNA, the differential expression analysis between cases and controls was carried out, as explained in Section 1.4. For each miRNA, the ratio of the λ estimates and corresponding difference in α estimates are presented in Figure 7.

On the left side of Figure 7, the ratios that stand out represent underexpressions of certain miRNAs in the 'case' subgroup as compared to the controls. In particular, the top 5 underexpressed miRNAs (i.e., the largest ratios of λ estimates) in decreasing order are hsa-miR-10b-5p, hsa-miR-186-5p, hsa-miR-128-3p, hsa-miR-222-3p, hsa-miR-101-3p, hsa-miR-146b-5p, hsa-miR-98-5p, hsa-miR-342-3p. However, while these miRNAs have clear differences in the λ estimates, they also have the biggest differences in the tail index α estimates, which means the parameter λ differences should be treated with care. For more insights, a zoom into the left side of Figure 7, where only the part with differences in tail index estimates is less than 0.1, is presented on the right side of Figure 7. In Table 11, the miRNAs with ratios of λ estimates most different from 1, from the right side of Figure 7, and the corresponding differences in the tail index estimates, are listed.

Table 10: NOWAC-LUCA: for $p = 198$ miRNAs, the summary of data for samples from the 'case' and 'control' subgroups.

	n	Min.	1st qu.	Median	Mean	3rd. qu.	Max.
Case	116	0	74	239	3064	1,207	877,569
Control	124	0	73	240	3636	1206	1,281,836

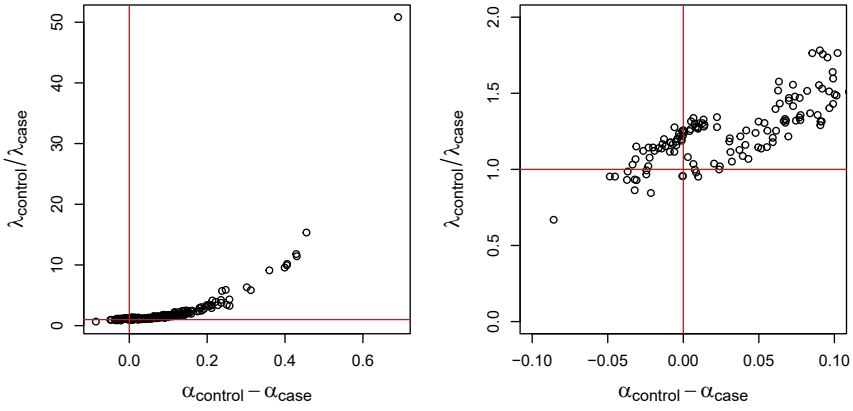

Figure 7: NOWAC-LUCA: for $p = 198$ miRNAs, the ratios of estimates for parameter λ and the differences in estimates for tail index α, between the 'control' and 'case' groups, labeled accordingly.

Table 11: NOWAC-LUCA: from the right side of Figure 7, miRNAs with ratios of the λ estimates clearly different from 1, where the differences in α estimates do not exceed 0.1 in absolute value.

miRNA	$(\hat{\lambda}_n)_{\text{control}}/(\hat{\lambda}_n)_{\text{case}}$	$(\hat{\alpha}_n)_{\text{control}} - (\hat{\alpha}_n)_{\text{case}}$
hsa-miR-142-3p	0.67	−0.09
hsa-miR-215-5p	1.51	0.1
hsa-miR-19b-3p	1.52	0.08
hsa-miR-182-5p	1.52	0.06
hsa-miR-339-5p	1.53	0.09
hsa-miR-328-3p	1.55	0.09
hsa-miR-99a-5p	1.56	0.07
hsa-miR-140-3p	1.58	0.06
hsa-miR-29c-3p	1.6	0.1
hsa-miR-486-3p	1.64	0.1
hsa-miR-652-3p	1.74	0.1
hsa-miR-99b-5p	1.76	0.09
hsa-miR-30a-5p	1.76	0.09
hsa-miR-23b-3p	1.78	0.09

For all these miRNAs we observe an underexpression of miRNAs among cases compared to controls, except for hsa-miR-142-3p, for which we observe an overexpression. In addition, results from an even more conservative approach regarding the tail differences, is presented in Table 12. In particular, from the right side of Figure 7, the top 20 ratios of λ estimates with constrained differences in tail index ≤ 0.01, is listed in Table 12.

Here, all ratios of the λ estimates represent underexpression of miRNAs in the 'case' subgroup. Although the ratios of the λ's in both Tables 11 and 12 are not too remarkable, the listed miRNAs could be considered as potential biomarkers. Similarly, from the left side of Figure 7, miRNAS with the largest relative differences in λ estimates as well as differences in α estimates should be studied further as potential biomarkers for lung cancer.

In (Nøst et al. 2023) nine candidate miRNAs were proposed as potential early markers of lung cancer; miR-320d, miR-320c, miR-320b, miR-92b-3p,miR-130b-3p, miR-200c-3p, miR-375-3p, miR-335-5p,and miR-323a-3p. This study found miR-320c and miR-320 elevated in pre-diagnostic specimen, i.e., they can be suggested as early markers of lung cancer. Based on Table 12, we would in addition highlight hsa-miR-335-5p (and hsa-miR-320a-3p from the miR-320 family).

Table 12: NOWAC-LUCA: from the right side of Figure 7, miRNAs with the largest relative differences in λ estimates, where the differences in α estimates do not exceed 0.01 in absolute value.

miRNA	$(\widehat{\lambda}_n)_{\text{control}}/(\widehat{\lambda}_n)_{\text{case}}$	$(\widehat{\alpha}_n)_{\text{control}} - (\widehat{\alpha}_n)_{\text{case}}$
hsa-miR-16-2-3p	1.2	0
hsa-let-7g-5p	1.2	0
hsa-miR-181a-5p	1.21	0
hsa-let-7i-5p	1.22	0
hsa-miR-93-5p	1.23	0
hsa-miR-146a-5p	1.24	0
hsa-miR-150-5p	1.25	0
hsa-let-7f-5p	1.25	0
hsa-miR-30e-5p	1.25	0
hsa-miR-185-5p	1.26	0
hsa-miR-486-5p	1.26	0
hsa-miR-199a-3p	1.27	0.01
hsa-miR-199b-3p	1.27	0.01
hsa-miR-425-5p	1.28	0.01
hsa-miR-335-5p	1.28	−0.01
hsa-miR-484	1.28	0.01
hsa-miR-126-5p	1.3	0.01
hsa-miR-320a-3p	1.3	0.01
hsa-miR-423-5p	1.32	0.01
hsa-miR-27b-3p	1.34	0.01

6 Conclusions

A new approach to analysis of miRNA-seq data is introduced. In particular, a novel application of modelling via the heavy tailed family of discrete stable distributions is proposed. In addition, the parameters of the discrete stable distribution are suggested as a possible alternative heavy tailed target of inference in differential expression analysis.

Based on two sources of miRNA-seq raw read counts data, the results of this paper show that (1) the miRNA-seq raw read counts arise from a heavy tailed distribution: exploratory analysis for heavy tails shows that both TCGA-BRCA summed expressions per miRNA as well as NOWAC-LUCA mature isoforms expressions per miRNA are most likely following a heavy tailed distribution. Moreover, in both datasets, the estimates for the tail index α imply a heavy tailed distribution; (2) the proposed heavy-tailed discrete stable distributions are suitable for modelling the miRNA-seq raw read counts: the goodness-of-fit of the discrete stable distributions are better than the popular Poisson and negative binomial distributions. Following (1) and (2), we conclude that in the differential expression analysis, a heavy-tailed approach (and possibly supported by some robust testing) would be most suitable.

Additionally, the proposed heavy-tailed approach for miRNA analysis via discrete stable distributions offers improved goodness-of-fit and benefits from relaxing the normalization requirement without any loss of information. Moreover, in the differential expression analysis, the analysis being conducted and the hypotheses being tested are not affected by data transformation, unlike the case with popular transformations such as log-transformation. In addition, the use of the discrete stable distribution allows us to explore various types of differences, such as differences in the tail index and parameter λ. However, in this paper, conclusions were primarily based on differences in estimates for the parameter λ, with the constraint that estimates for the tail index α are small (indicating similar tails). While differences in the α estimates are interesting, interpreting them can be more challenging because they represent the degree of tail heaviness, which may be difficult to

grasp in a practical sense. Nonetheless, differences in α estimates can provide valuable information about the relative likelihood of extreme values in the different distributions.

In conclusion, the analysis performed in this paper shows that independently of the source (e.g., blood, tissue, etc.) or type (isoform or miRNA expression) of miRNA-seq raw read counts, discrete heavy-tailed distributions, such as the proposed discrete stable distribution, is a promising approach to modelling of miRNA-seq raw read counts.

Acknowledgment: The results published for TCGA data are based upon public data generated by the TCGA Research Network: https://www.cancer.gov/tcga.
Research ethics: TCGA: Info about TCGA study research ethics is available in https://www.cancer.gov/ccg/research/genome-sequencing/tcga/history/ethics-policies. NOWAC: All participants in NOWAC study gave written informed consent and the study was approved by the Regional Committee for Medical and Health Research Ethics and the Norwegian Data Inspectorate. For more info see https://uit.no/research/nowac.
Author contributions: AK and MT conceived the research idea. AK conducted the methodological developments and the statistical analyses. THN was responsible for the acquisition of NOWAC data and the biological interpretation of the results. AK and MT wrote the manuscript, with inputs from THN. All authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Competing interests: The authors state no conflict of interest.
Research funding: EU Horizon 2020; Marie Skłodowska-Curie Grant number 801133; Norwegian Research Council, Grant numbers 262111 and 237718.
Data availability: TCGA: Data are publically available at https://www.cancer.gov/tcgaNOWAC: Data cannot be shared publicly because of local and national ethical and security policies. For more info see https://uit.no/research/nowac.

Appendix A: TCGA-BRCA: exploratory data analysis

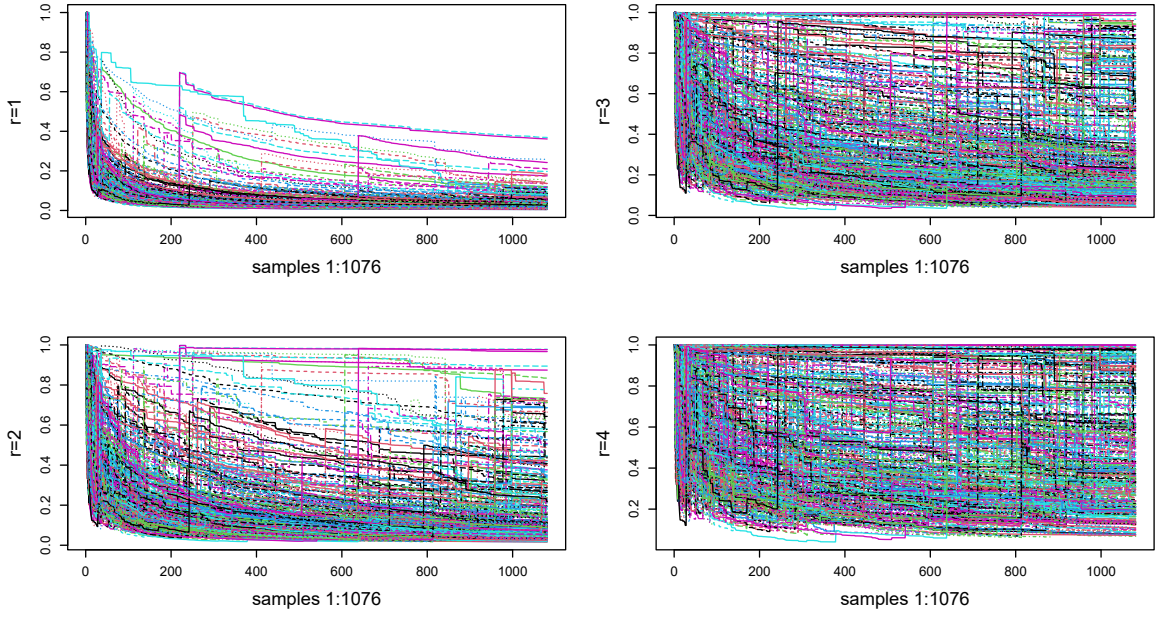

Figure A.1-F.1: TCGA-BRCA: for $p = 400$ miRNAs, the ratio of maximum to sum of order $r = 1, 2, 3, 4$ for data from all samples ($n = 1076$).

Appendix B: TCGA-BRCA: miR-200 family

Table A.2-T.1: TCGA-BRCA: summary of miR-200 family members.

Type	Min.	1st qu.	Median	Mean	3rd qu.	Max.
			Summary of miR200a			
Normal-like	99	1038	1955	2525	2824	13,483
LumA	68	1580	2688	3822	4767	56,886
LumB	19	1246	2041	2991	3380	19,371
Basal	6	1306	2321	3262	4120	21,310
Her2	33	1283	2163	2680	3452	13,458
			Summary of miR200b			
Normal-like	57.0	951.8	1938.0	2377.8	2602.5	10,401
LumA	59	1143	2034	3042	3905	35,239
LumB	24.0	885.2	1438.0	2177.5	2400.8	15,458
Basal	7.0	909.2	1745.5	2598.2	3340.5	20,688
Her2	17	792	1365	1896	2433	11,426
			Summary of miR200c			
Normal-like	1246	13,397	26,235	34,708	38,226	137,471
LumA	2732	19,000	31,958	44,683	58,266	324,329
LumB	4481	17,654	27,240	37,545	41,597	283,239
Basal	4589	15,679	27,344	41,514	52,562	241,924
Her2	4776	14,840	25,993	31,826	36,457	160,062
			Summary of miR141			
Normal-like	161	2262	3476	4576	5258	22,643
LumA	269	3082	5333	6674	8580	36,871
LumB	458	3042	4572	6281	7561	36,743
Basal	485	2920	4440	6831	9347	37,275
Her2	640	2499	4273	5510	6640	26,727
			Summary of miR429			
Normal-like	8.0	135.0	259.0	310.1	384.0	1379
LumA	10.0	170.5	307.0	461.4	566.8	6277
LumB	4.0	145.0	253.0	382.4	444.2	2410
Basal	0.0	167.0	314.5	508.3	662.5	4122
Her2	1.0	132.0	289.0	434.1	535.0	3072

Table A.2-T.2: TCGA-BRCA: Moody's median tests in miR-200 family. Results are based on multiple comparisons, where groups labeled with the same letter were found to originate from populations with the same median.

Type	Median	Groups[a]
	Median test for miR200a	
Normal-like	1955.0	c
LumA	2687.5	a
LumB	2041.0	bc
Basal	2321.0	b
Her2	2163.0	bc
	Median test for miR200b	
Normal-like	1938.0	ab
LumA	2034.5	a
LumB	1438.0	b
Basal	1745.5	ab
Her2	1365.0	b
	Median test for miR200c	
Normal-like	26,235.0	b
LumA	31,958.0	a
LumB	27,239.5	b
Basal	27,343.5	b
Her2	25,993.0	b
	Median test for miR141	
Normal-like	3476.0	b
LumA	5333.0	a
LumB	4571.5	b
Basal	4440.0	b
Her2	4273.0	b
	Median test for miR429	
Normal-like	259.0	ab
LumA	307.0	a
LumB	253.0	b
Basal	314.5	a
Her2	289.0	ab

[a]The medians of groups labeled with the same letter were found to be not significantly different at a significance level of 0.05. For example, for miR429, normal-like is labeled 'ab' and subgroups LumA and LumB are labeled 'a' and 'b', respectively, indicating they are not significantly different from normal-like, while they differ from each other.

Table A.2-T.3: TCGA-BRCA: the log2 transform based *t*-test in miR-200 family.

Group1	Group2	*p*
	log2 transform *t*-test for miR200a	
Normal-like	Basal	0.04
Normal-like	Her2	0.84
Normal-like	LumA	0.00
Normal-like	LumB	0.40
	log2 transform *t*-test for miR200b	
Normal-like	Basal	0.37
Normal-like	Her2	0.16
Normal-like	LumA	0.05
Normal-like	LumB	0.69
	log2 transform *t*-test for miR200c	
Normal-like	Basal	0.07
Normal-like	Her2	0.73
Normal-like	LumA	0.01
Normal-like	LumB	0.19
	log2 transform *t*-test for miR141	
Normal-like	Basal	0.00
Normal-like	Her2	0.03
Normal-like	LumA	0.00
Normal-like	LumB	0.00
	log2 transform *t*-test for miR429	
Normal-like	Basal	0.01
Normal-like	Her2	0.52
Normal-like	LumA	0.02
Normal-like	LumB	0.27

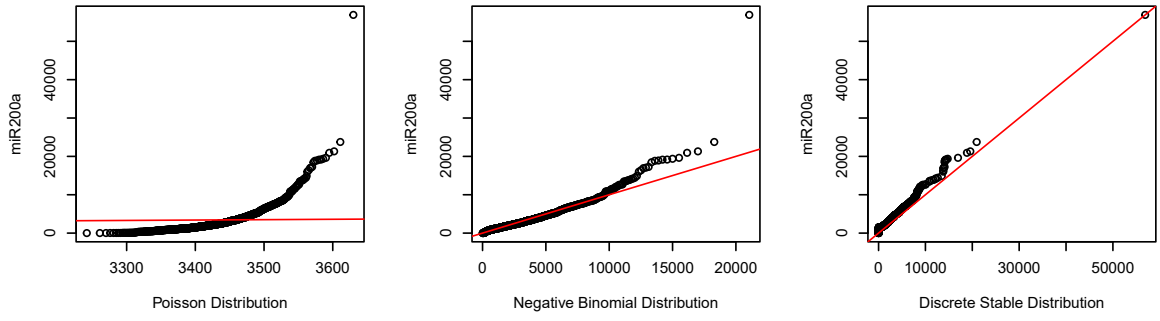

Figure A.2-F.1: TCGA-BRCA: QQ-plots for miR-200 family member miR200a.

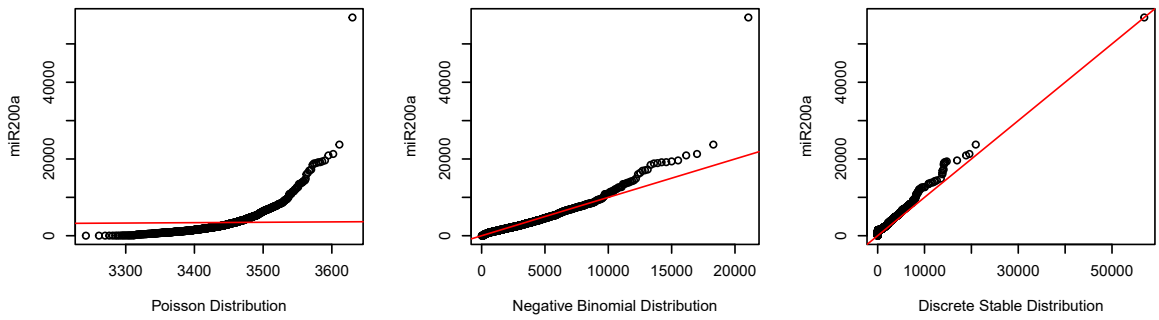

Figure A.2-F.2: TCGA-BRCA: QQ-plots for miR-200 family member miR200b.

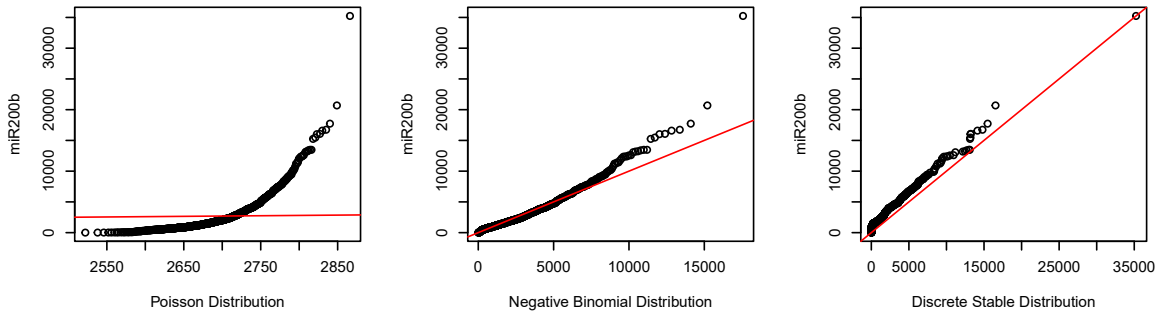

Figure A.2-F.3: TCGA-BRCA: QQ-plots for miR-200 family member miR200c.

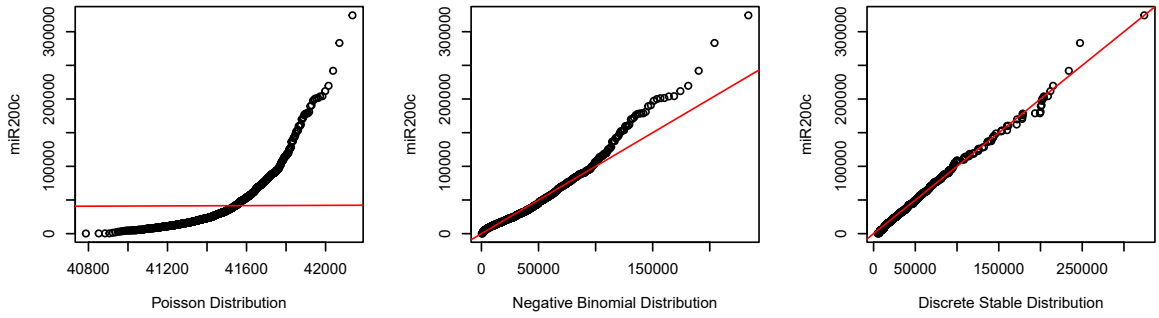

Figure A.2-F.4: TCGA-BRCA: QQ-plots for miR-200 family member miR141.

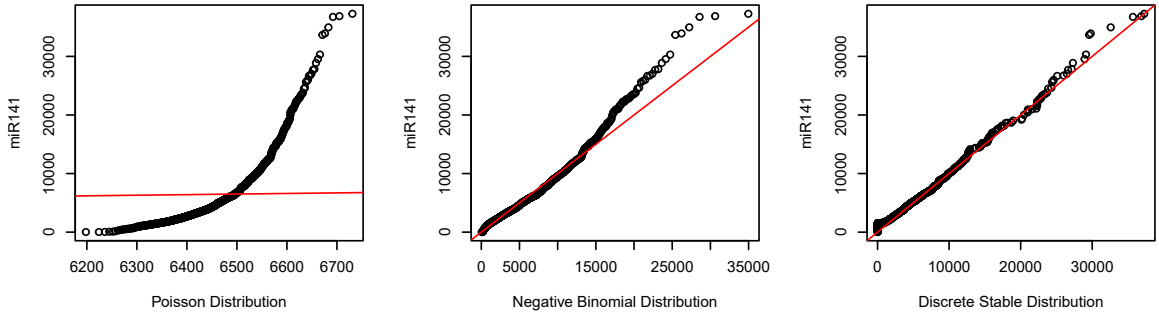

Figure A.2-F.5: TCGA-BRCA: QQ-plots for miR-200 family member miR429.

Appendix C: NOWAC-LUCA: exploratory data analysis

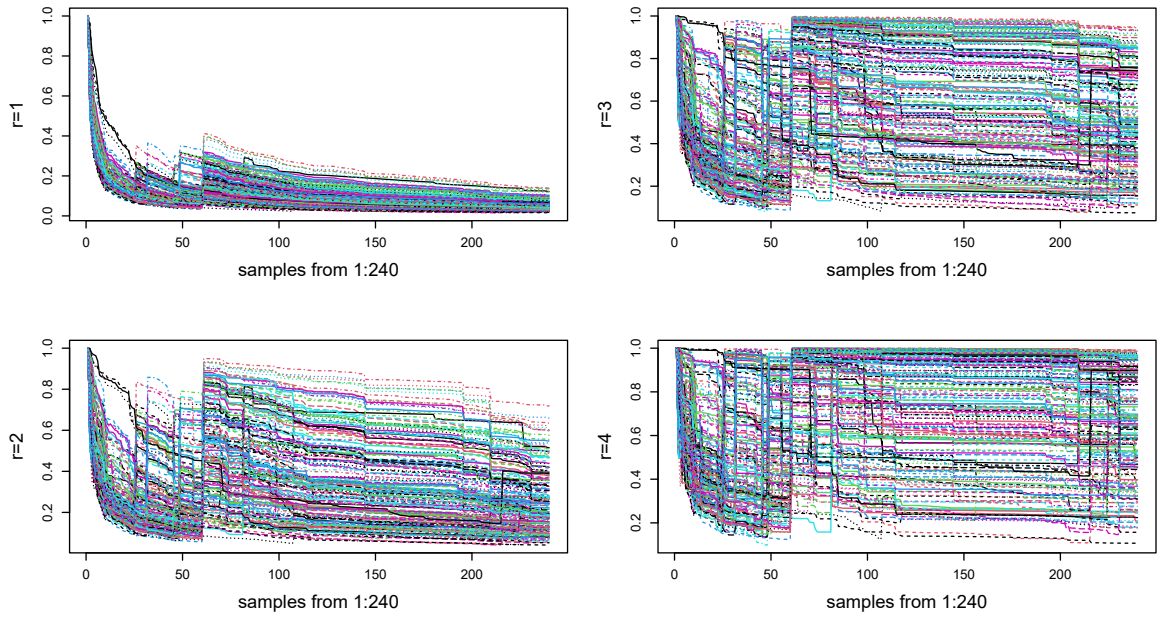

Figure A.3-F.1: NOWAC-LUCA: for $p = 198$ miRNAs, the ratio of maximum to sum of order $r = 1, 2, 3, 4$ for data from all samples ($n = 240$).

References

Allen, G.I. and Liu, Z. (2013). A local Poisson graphical model for inferring networks from sequencing data. *IEEE Trans. NanoBioscience* 12: 189—198.

Ara, T. (2020). Brunnermunzel: (permuted) Brunner-Munzel test, Available at: https://CRAN.R-project.org/package=brunnermunzel .Rpackageversion1.4.1.

Baccini, A., Barabesi, L., and Stracqualursi, L. (2016). Random variate generation and connected computational issues for the Poisson—Tweedie distribution. *Comput. Stat.* 31: 729—748.

Cavallari, I., Ciccarese, F., Sharova, E., Urso, L., Raimondi, V., Silic-Benussi, M., D'Agostino, D.M., and Ciminale, V. (2021). The mir-200 family of microRNAs: Fine tuners of epithelial-mesenchymal transition and circulating cancer biomarkers. Cancers 13: 5874.

Choi, H., Gim, J., Won, S., Kim, Y.J., Kwon, S., and Park, C. (2017). Network analysis for count data with excess zeros. BMC Bioinf. 18: 93.

Christoph, G. and Schreiber, K. (1998). Discrete stable random variables. *Stat. Prob. Lett.* 37: 243—247.

Delignette-Muller, M.L. and Dutang, C. (2015). fitdistrplus: an R package for fitting distributions. *J. Stat. Software* 64: 1—34.

Devroye, L. (1993). A triptych of discrete distributions related to the stable law. *Stat. Prob. Lett.* 18: 349—351.

Doray, L.G., Jiang, S.M., and Luong, A. (2009). Some simple method of estimation for the parameters of the discrete stable distribution with the probability generating function. *Commun. Stat. Simulat. Comput.* 38: 2004—2017.

Embrechts, P., Klüppelberg, C., and Mikosch, T. (2013). Modelling extremal events: for insurance and finance. In: *Stochastic modelling and applied probability*. Springer Berlin Heidelberg, Available at: https://books.google.no/books?id=BXOI2pICfJUC.

Faraway, J., Marsaglia, G., Marsaglia, J., and Baddeley, A. (2021). Goftest: classical goodness-of-fit tests for univariate distributions, Available at: https://CRAN.R-project.org/package=goftest.Rpackageversion1.2-3.

Felipe de Mendiburu (2023). Agricolae: statistical procedures for agricultural research, Available at: https://CRAN.R-project.org/ package=agricolae.Rpackageversion1.3-7.

Fontana, A., Barbano, R., Dama, E., Pasculli, B., Rendina, M., Morritti, M.G., Melocchi, V., Castelvetere, M., Valori, V.M., Ravaioli, S., et al. (2021). Combined analysis of mir-200 family and its significance for breast cancer. *Sci. Rep.* 11: 2980.

Grabchak, M. (2018). Domains of attraction for positive and discrete tempered stable distributions. *J. Appl. Prob.* 55: 30—42.

Grabchak, M. (2022). Discrete tempered stable distributions. *Methodol. Comput. Appl. Probab.* 24: 1877—1890.
Joanes, D.N. and Gill, C.A. (1998). Comparing measures of sample skewness and kurtosis. *J. R. Stat. Soc. Ser. D Statistician* 47: 183—189.
Kalecky, K., Modisette, R., Pena, S., Cho, Y.R., and Taube, J. (2020). Integrative analysis of breast cancer profiles in TCGA by TNBC subgrouping reveals novel microRNA-specific clusters, including mir-17-92a, distinguishing basal-like 1 and basal-like 2 TNBC subtypes. *BMC Cancer* 20: 141.
Klebanov, L.B. and Slámová, L. (2013). Integer valued stable random variables. *Stat. Prob. Lett.* 83: 1513—1519.
Krutto, A. (2018). Empirical cumulant function based parameter estimation in stable laws. *Acta Commentationes Univ. Tartuensis Math.* 22: 311—338.
Krutto, A. (2023). dstabledist: the discrete stable distribution functions, Available at: https://CRAN.R-project.org/package=dstabledist .Rpackageversion0.1.0.
Kume, K., Iwama, H., Deguchi, K., Ikeda, K., Takata, T., Kokudo, Y., Kamada, M., Fujikawa, K., Hirose, K., Masugata, H., et al. (2017). Serum microRNA expression profiling in patients with multiple system atrophy. *Mol. Med. Rep.* 17: 852—860.
Lember, J. and Krutto, A. (2022). Estimating the logarithm of characteristic function and stability parameter for symmetric stable laws. Methodol. Comput. Appl. Probab. 24: 2149—2167.
Li, Y., Rahman, T., Ma, T., Tang, L., and Tseng, G.C. (2023). A sparse negative binomial mixture model for clustering RNA-seq count data. Biostatistics 24: 68—84.
Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with deseq2. Genome Biol. 15: 550.
Lund, E., Dumeaux, V., Braaten, T., Hjartåker, A., Engeset, D., Skeie, G., and Kumle, M. (2007). Cohort profile: the Norwegian women and cancer study—NOWAC—kvinner og kreft. *Int. J. Epidemiol.* 37: 36—41.
Marcheselli, M., Baccini, A., and Barabesi, L. (2008). Parameter estimation for the discrete stable family. *Commun. Stat. Theor. Methods* 37: 815—830.
Misra, N. and Kuruoglu, E.E. (2016). Stable graphical models. *J. Mach. Learn. Res.* 17: 5862—5897.
Mounir, M., Lucchetta, M., Silva, T.C., Olsen, C., Bontempi, G., Chen, X., Noushmehr, H., Colaprico, A., and Papaleo, E. (2019). New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEx. PLOS Comput. Biol. 15: e1006701.
Nøst, T.H., Skogholt, A.H., Urbarova, I., Mjelle, R., Paulsen, E.E., Dønnem, T., Andersen, S., Markaki, M., Røe, O.D., Johansson, M., et al. (2023). Increased levels of microRNA-320 in blood serum and plasma is associated with imminent and advanced lung cancer. *Mol. Oncol.* 17: 312—327.
Prodanov, D. (2019). Regularized integral representations of the reciprocal gamma function. Fractal Fractional 3: 1.
Purdom, E. and Holmes, S.P. (2005). Error distribution for gene expression data. *Stat. Appl. Genet. Mol. Biol.* 4: 1—35.
Qian, L. and Zhu, F. (2023). A flexible model for time series of counts with overdispersion or underdispersion, zero-inflation and heavy-tailedness. *Commun. Math. Stat.*.
Qian, L., Li, Q., and Zhu, F. (2020). Modelling heavy-tailedness in count time series. Appl. Math. Model. 82: 766—784.
R Core Team (2020). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, Available at: https://www.R-project.org/.
Rémillard, B. and Theodorescu, R. (2000). Inference based on the empirical probability generating function for mixtures of Poisson distributions. *Stat. Risk Model.* 18: 349—366.
Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43: e47.
Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26: 139—140.
Robust Analysis Inc (2017). *STABLE 5.3 R version for windows*. Robust Analysis Inc., Washington, DC, USA, Available at: http://www .robustanalysis.com.
Sinclair, D. and Hooker, G. (2019). Sparse inverse covariance estimation for high-throughput microRNA sequencing data in the Poisson log-normal graphical model. J. Stat. Comput. Simulat. 89: 3105—3117.
Slámová, L. and Klebanov, L.B. (2014). Approximated maximum likelihood estimation of parameters of discrete stable family. *Kybernetika* 50: 1065—1076.
Soltani, A.R., Shirvani, A., and Alqallaf, F. (2009). A class of discrete distributions induced by stable laws. *Stat. Prob. Lett.* 79: 1608—1614.
Staudte, R.G. and Sheather, S.J. (2011). *Robust estimation and testing. Wiley series in probability and statistics*. Wiley, Available at: https:// books.google.no/books?id=9ut_NMzC1l4C.
Stephens, M.A. (1986). Tests based on edf statistics. In: D'Agostino, R.B. and Stephens, M.A. (Eds.). *Goodness-of-fit techniques, volume 68 of statistics, textbooks and monographs*, chapter 4. Marcel Dekker, New York.
Steutel, F.W. and van Harn, K. (1979). Discrete analogues of self-decomposability and stability. *Ann. Probab.* 7: 893—899.
Steutel, F.W. and van Harn, K. (2003). *Infinite divisibility of probability distributions on the real line*. Chapman & Hall/CRC Pure and Applied Mathematics. Taylor & Francis, Available at: https://books.google.no/books?id=sPnSmAEACAAJ.
Wang, L., Feng, Z., Wang, X., Wang, X., and Zhang, X. (2009). DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics* 26: 136—138.

Wen, B., Zhu, R., Jin, H., and Zhao, K. (2021). Differential expression and role of miR-200 family in multiple tumors. *Anal. Biochem.* 626: 114243.

Wilcox, R.R. (2022). *Chapter 5 — comparing two groups*, 5th ed. Academic Press, pp. 153—251, Available at: https://www.sciencedirect.com/science/article/pii/B9780128200988000117.

Yang, E., Ravikumar, P., Allen, G.I., and Liu, Z. (2015). Graphical models via univariate exponential family distributions. *J. Mach. Learn. Res.* 16: 3813—3847.

Ye, F., Tang, H., Liu, Q., Xie, X., Wu, M., Liu, X., Chen, B., and Xie, X. (2014). mir-200b as a prognostic factor in breast cancer targets multiple members of rab family. *J. Trans. Med.* 12: 17.

Zhao, Y., Wong, L., and Goh, W.W.B. (2020). How to do quantile normalization correctly for gene expression data analyses. *Sci. Rep.* 10: 15534.

Žitnik, M. and Zupan, B. (2015). Gene network inference by fusing data from diverse distributions. *Bioinformatics* 31: i230—i239.