Erythropoietic miR-486-5p and miR-451a depletion from whole blood-derived small RNA sequencing libraries *

Simonas Juzenas[1], Carl Mårten Lindqvist[2], Go Ito[1], Yewgenia Dolshanskaya[1], Jonas Halfvarson[2], Andre Franke[1] and Georg Hemmrich-Stanisak[1]

[1]*Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, DE 24105 Kiel, Germany;*
[2]*School of Medical Sciences, Faculty of Medicine and Health, Örebro University, SE 70182 Örebro, Sweden.*

Abstract

Erythroid-specific miR-451a and miR-486-5p are two dominant microRNAs (miRNAs) in human peripheral blood. In small RNA sequencing libraries their overabundance reduces diversity as well as complexity and consequently cause negative effects such as missing detectability and inaccurate quantification of low abundant miRNAs. Here we present a cost-effective and easy to implement hybridization-based method to deplete these two erythropoietic miRNAs from blood-derived RNA samples. By utilization of blocking oligonucleotides, this method provides a highly efficient and specific depletion of miR-486-5p and miR-451a, which leads to a considerable increase of measured expression as well as detectability of low abundant miRNA species. The blocking oligos are compatible with 5' ligation-dependent small RNA library preparation protocols, including commercially available kits, such as Illumina TruSeq and Perkin Elmer NEXTflex.

Introduction

Small RNA sequencing (smRNA-Seq) is a widely used application, enabling discovery and quantification of small RNAs, including microRNAs (miRNAs), which regulate gene expression and are emerging as important disease biomarkers (1, 2). Many of the current miRNA-based biomarker studies use plasma, serum, exosomes or peripheral blood as a source material. The biggest advantage of peripheral blood over the other blood-derived biological materials comes from very low technical variability during sample handling and processing, which makes this material very attractive for the use in a clinical setting (3). Moreover, peripheral blood contains complete convolved information about miRNA expression of all cellular and non-cellular blood compounds. On the other hand, some of these compounds in whole blood are causing more problems than benefits. For example, in addition to globin mRNAs, red blood cells also contain highly abundant miRNAs including miR-486-5p and miR-451a (4, 5). These two conserved, non-canonical miRNA molecules represent dominant miRNAs, which are specifically upregulated in the erythroid lineage (6). Loss of *mir-486* and *mir-451a* genes in mice leads to erythroid defects, showing their importance in erythrocyte development (6), and explaining their high abundance in erythroid cells as well as in whole blood.

In RNA-Seq workflows, overabundance of transcripts reduces the complexity of PCR-amplified libraries and exhausts sequencing space, which leads to unwanted effects such as inaccurate detectability and quantification of low abundant transcripts. Low level of detectability is a very

*This is a pre-print version of manuscript for bioRxiv; **Corresponding authors**: sjuzenas@ikmb.uni-kiel.de and g.hemmrich-stanisak@ikmb.uni-kiel.de.

important issue in the biomarker field, since it is believed that disease-derived miRNAs may circulate in blood at low amounts (7).

Concerns about depletion of unwanted miRNAs in small RNA libraries are not new, and at least two approaches have been described previously (8, 9). A hybridization-based approach relies on stem-loop shaped oligonucleotide with a 12-base 5' overhang which is the reverse complement to the first 12 bases of the 5' end of the target miRNA's canonical sequence (8). Another method is the CRISPR/Cas9-based approach, where Cas9 is complexed with single guide RNAs that target undesirable miRNA sequences for cleavage *in vitro* (9). Both of these methods have been shown to effectively reduce target miRNA sequences in small RNA libraries; however, both of them have their own limitations. The hybridization-based approach has a high chance of unspecific hybridization with non-target miRNAs due to the short complementarity region (12 bases) of the stem-loop oligonucleotide. The CRISPR/Cas9-based method includes additional steps and reagents which increase hands-on-time and costs of the procedure.

Here we describe a cost-effective and easy to implement hybridization-based method for erythropoietic miR-486-5p and miR-451a depletion from small RNA libraries prior to sequencing. The method relies on linear oligonucleotides covering the longest stable complementary region of target miRNA sequence variants to prevent them from 5' adapter ligation and therefore is compatible with 5' ligation-dependent small RNA library preparation protocols.

Methods

Blood miRNA catalog data analysis

Compositional analysis of miRNA species in blood compounds was performed using processed miRNA count data from our previously published study (10). The dataset contains miRNA expression counts of seven types of blood cells, exosomes, serum and whole blood, which were generated by using TruSeq (Illumina) small RNA library preparation protocol. The miRNA read counts for each sample in the dataset were down-sampled to a constant number of reads using random subsampling rrarefy() function implemented in the R package vegan (11). For visualization, relative abundance for each miRNA (i) in each sample (j) was estimated using the following formula:

$$relative\ abundance_{(i,\ j)}\ (\%) = \frac{miRNA\ counts_{(i,\ j)}}{total\ miRNA\ counts_j} * 100$$

Shannon diversity index was calculated for each sample based on down-sampled miRNA counts by employing the diversity() function from the R package vegan (11). A lower-tailed t-test was performed to evaluate the differences of the index values between whole blood and every other blood compound using t.test() function from the base R package.

Blocking oligo design

Erythropoietic miR-486-5p and miR-451a blockers are linear single-strand DNA oligonucleotides with 3' C3 spacer (propyl group) modification (IDT). The oligos are designed to prevent 5' adapter ligation via blocking the access of T4 RNA ligase to the 5' end of target miRNAs during smRNA-seq library preparation. Sequences of two blocking oligos were composed based on our previously published whole blood smRNA-seq data (10). Briefly, all sequences that were mapped to precursors of miR-486-5p and miR-451a were pooled in order to obtain all possible unique sequence variants for each miRNA. The obtained unique sequences for each target miRNA were then flattened into a single consensus sequence in order to retrieve the most frequent nucleotide found at each position

in a sequence alignment. The consensus sequences were then used to generate reverse complement oligonucleotides to bind target miRNAs by Watson–Crick base pairing. The C3 spacer modification was attached to the 3' ends of oligonucleotides to avoid self-ligation. The resulting oligonucleotides are provided in **Figure 2A**.

RNA extraction

Whole blood samples from 10 healthy volunteers (n = 5 females and n = 5 males) aged from 25 to 37 were collected into PAX gene RNA blood tubes (Qiagen). Total RNA samples were isolated using QIAcube automation with the PAXgene Blood RNA Kit (Qiagen) in accordance to manufacturer's instructions.

Small RNA-seq and erythropoietic miRNA blocking

Small RNA libraries were prepared using standard and modified TruSeq Small RNA Library Prep Kit v02 (Illumina) and gel-free NEXTFLEX Small RNA Seq Kit v3 (Perkin Elmer) protocols. The modified versions of the protocols include an additional step where blocking oligos for miR-486-5p and miR-451a are included. In case of the TruSeq protocol, 1 μl of 20 μM of the blocking oligo mix was introduced and annealed immediately after the 3' adapter ligation (ramp from 65 to 45 °C – 0.1 per sec), whereas for the NEXTflex protocol, 1 μl of 10 μM of the blocking oligo mix was introduced and annealed directly to the total RNA sample prior to library preparation (detailed protocols are provided in **Supplementary Methods**). In order to achieve the best performance of the TruSeq and NEXTflex library preparation methods, starting RNA amounts were selected to be in a range of manufacturer's provided recommendations. Concisely, for each TruSeq (standard and modified) library preparation, 1 μg of total RNA was used as starting material, whereas for each NEXTflex (standard and modified) library preparation, 100 ng of total RNA was used as input. Subsequently, for each sample standard and modified libraries were generated and randomized in a supervised fashion (blocked and unblocked paired samples on the same lane to minimize batch-effect) and pooled with 10 samples per lane. Sequencing was performed on an Illumina HiSeq 4000 (1 x 50 bp SR, v3) platform.

Small RNA-seq data processing and mapping

Obtained demultiplexed raw sequencing reads (fastq) were processed by cutadapt v1.9 (12) which was used to trim adapter sequences and low quality bases (>Q20), and to discard sequences shorter than 18 nucleotides in length, with the following parameters for TruSeq data: "cutadapt -a TGGAATTCTCGGGTGCCAAGG -m 18 -q 20 –discard-untrimmed"; and with the following parameters for NEXTflex data: "cutadapt -u 4 -a NNNNTGGAATTCTCGGGTGCCAAGG -m 18 -q 20 –discard-untrimmed". The processed reads were then mapped to miRNA sequences from miRBase v22 (13) using mirAligner (14) with default parameters (1 mismatch, 3 nt in the 3' or 5' trimming variants and the 3 nt in 3' - addition variants). The R package isomiRs v1.10.1 (15) with default parameters was used to generate the count matrix of miRNA counts per library. Samples with fewer than one million mapped reads were excluded from further analysis. The sequencing depth of mapped reads to miRNA reference sequences is shown in **Supplementary Figure 1**. Raw sequencing reads and quantified read-count data have been deposited at NCBI Gene Expression Omnibus (GEO) (16) under the accession number GSE138318.

Blocking efficiency estimation

In order to estimate the efficiency of blocking oligos for miR-486-5p and miR-451a, read counts were normalized to counts per million (CPM) by employing the cpm() function from the R package edgeR (17). The blocking efficiency for each target miRNA (i) in each paired sample (j) was then calculated using the following formula:

$$blocking\ efficiency_{(i,\ j)}\ (\%) = \left(1 - \frac{target\ miRNA\ CPM_{(i,\ j)}^{(blocking\ protocol)}}{target\ miRNA\ CPM_{(i,\ j)}^{(standard\ protocol)}}\right) * 100$$

A one-sided Wilcoxon rank sum test was used to compare blocking efficiencies between NEXTflex and TruSeq protocols by employing wilcox.test() function from the base R package.

Detectability determination

In order to compare miRNA detectability with or without the use of the blocking oligos, the obtained read counts were normalized to CPM as described previously and only the miRNAs which had at least 1 CPM in at least 75% of the libraries per protocol were considered as detected with confidence. The overlaps of detected miRNAs among different protocols were calculated and visualized using the R package ggupset (18). The simulation of random down-sampling of miRNA counts was performed by employing drarefy() function from the R package vegan (11), which returns probabilities for each miRNA to be detected in a random subsample. For each sample, miRNA counts were subsampled to seven different levels (5M, 4M, 3M, 2M, 1M, 0.5M and 0.1M). In the down-sampling experiment, a miRNA was considered to be detected when the probability was above 0.9.

Blocking effect on quantitative performance estimation

The analysis of blocking oligo effect on non-targeted miRNA quantitative estimates was based on samples for which both blocked (modified) and unblocked (standard) libraries were generated. Spearman and Pearson correlation analyses between blocked and unblocked paired samples were performed on the log-transformed (with pseudo-count 1) CPM values of detected miRNAs. Spearman's rank and Pearson's correlation coefficients were calculated using the cor() function (implemented in the R base package) with all observations except those of miR-486-5p and miR-451a. The R package DESeq2 (19) with default parameters and paired sample design was used to estimate the differential expression of miRNAs between blocked and unblocked small RNA libraries. The P-values resulting from Wald tests were corrected for multiple testing according to Bonferroni method (20). The miRNAs with a corrected P-value < 0.01 and $|\log2FC| > 1$ were considered to be significantly differentially expressed. The RNA-cofold algorithm from the Vienna RNA package (21) was employed to evaluate nonspecific hybridization between blocking oligonucleotides and down-regulated non-targeted miRNAs. All the data was visualized using the R package ggplot2 (22).

Results

Erythropoietic miR-486-5p and miR-451a domination reduces the detectability of low abundant miRNAs in whole blood samples

Human blood is a liquid, composite biological tissue consisting of multiple cells (erythrocytes, monocytes, neutrophils, lymphocytes, thrombocytes, etc.) and cell-derived components (platelets, exosomes, etc.) suspended in a medium known as plasma (23). Besides blood-cell expressed miRNAs, blood also contains circulating miRNAs that are detected in serum, plasma or exosomes (24). Therefore, the expectation is that the small RNA-sequenced whole blood sample should contain convolved information about the miRNA composition of all cellular and non-cellular blood components.

To test this hypothesis, we performed a compositional analysis of miRNA species in blood compounds and whole blood samples using blood miRNA catalog data (10). Here, for each sample, we calculated the Shannon diversity index, which combines the information about miRNA richness (number of detected different miRNA species) and evenness (proportion of each miRNA) within a given sample (25). Unexpectedly, whole blood samples showed the lowest average miRNA diversity (corrected P-value range: $9.94\text{x}10^{-110}$ – $1.42\text{x}10^{-07}$; **Figure 1**) as well as the lowest average level of richness (data not shown) values among separately sequenced blood compounds, meaning that more miRNAs were detected in every single blood compound than in whole blood itself.

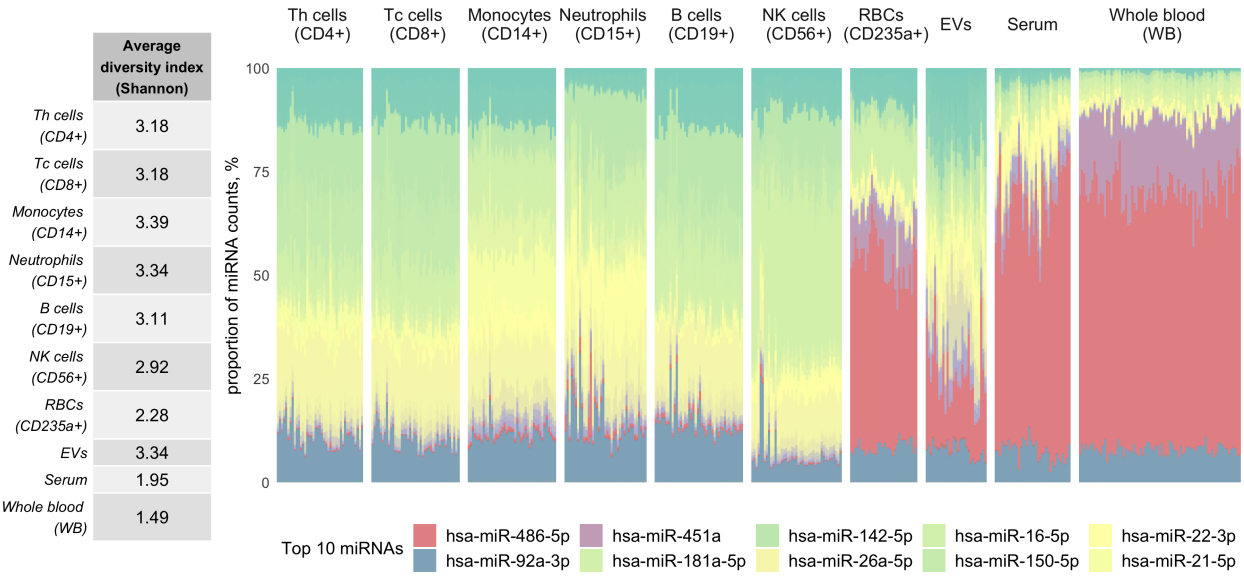

Figure 1: Erythropoietic miR-486-5p and miR-451a domination reduces diversity of whole blood-derived small RNA libraries. The left side panel represents averaged Shannon diversity index of every blood compound. Shannon index value for each sample was calculated on down-sampled miRNA count data, and then the values were compared between whole blood and every other compound using pairwise t-test. The significantly lowest miRNA diversity is observed in whole blood (PAXgene) samples. In the right side panel, a stacked bar chart presents relative abundance values of miRNAs in different blood compounds, where every bar represents a sample and every color represents a fraction of single miRNA within a given sample. The graph reveals high dominance of erythropoietic miR-486-5p and miR-451a in whole blood, red blood cell (RBC), serum and exosome samples. The ten most abundant miRNAs in all of the blood compounds are shown in the legend.

To understand what is causing low miRNA diversity in the whole blood samples, we looked at relative miRNA abundances and observed high domination of miR-486-5p and miR-451a molecules

in the samples, where they both comprise about ~80% of total mapped reads (**Figure 1**). In addition to whole blood, we observed that these miRNAs are also dominant in erythrocytes, serum and partly in exosomes. Interestingly, miR-486-5p and miR-451a were previously shown to be involved in erythroid development (6), which partly explains why these miRNAs are so abundant in erythrocytes. Since erythrocytes and reticulocytes together comprise more than 90% of blood cells, these miRNAs are even more abundant in whole blood. A high abundance of these two erythropoietic miRNAs in exosomes and, especially, in serum might be explained by contamination due to hemolysis, which occurs during sample handling.

Taken together, these observations suggested that the domination of erythropoietic miR-486-5p and miR-451a in the whole blood small RNA libraries exhausts sequencing space and reduces the detectability of other low abundant miRNAs coming from less copious cells or non-cellular blood compounds.

Erythropoietic miRNA blocking in 5' ligation-dependent small RNA libraries

To deplete erythropoietic miR-486-5p and miR-451a, we designed linear oligonucleotides covering the longest stable complementarity of target miRNA sequence variants and bearing terminal modifications to prevent these oligonucleotides from participating in ligation reactions or being extended by polymerase. To design the oligonucleotides, we have obtained all sequence variants (isomiRs) of mature miR-486-5p and miR-451a molecules from our previously published whole blood smRNA-seq dataset (10). To identify the consensus sequences for each target miRNA, we calculated nucleotide frequencies at each position in sequence alignments of the precursor miRNAs and flattened them into one sequence (**Figure 2A**). The consensus sequences revealed higher variability within 3' than within 5' end of target miRNA sequences. This occurs due to non-templated nucleotide additions, which, beside cleavage-directed 5' and 3' end modifications, introduce additional variation within 3' ends of miRNA sequences (26). Therefore, because of this miRNA feature, we decided to block 5' end of target miRNAs and to prevent them from 5' adapter ligation. For each target miRNA, we generated reverse complement oligonucleotides of the consensus sequence containing the most stable nucleotides starting from the 5' ends of the molecules. Finally, on the 3' ends of the oligonucleotides, we added C3 spacer (propyl group) modification, which prevents the blocking oligonucleotides from self-ligation and extension. Complementary DNA oligos bearing this modification have been previously shown to prevent 5' adapter ligation to Drosophila 2S rRNA in smRNA-seq libraries (27).

In order to test our blocking oligos targeting miR-486-5p and miR-451a, we chose two 5' adapter ligation-dependent methods, the commercially-available TruSeq and NEXTflex protocols. TruSeq is one of the most commonly used methods for small RNA library preparation employing adapters with invariant sequences, while NEXTflex utilizes adapters containing four degenerate nucleotides at the ligation ends as a strategy to reduce ligation bias, which is a well-known issue in small RNA library preparation (28).

The standard TruSeq protocol can be summarized in three core steps: 3' adapter ligation, 5' adapter ligation and reverse transcription (RT) accompanied by amplification of the cDNA library. Since the blocking oligos were designed to target 5' ends of complementary miRNA sequences, in the modified TruSeq protocol, we have introduced these oligos prior to 5' adapter ligation step (**Figure 2B**), where the complementary segments of targeted miRNA sequences and the blocking oligos take part in the formation of double-stranded RNA:DNA hybrids. These blunt-ended or slight 3' DNA overhang-having double-stranded hybrids are not suitable substrates for T4 RNA ligase to join a single-stranded adapter to the 5' end of RNA strand in the hybrid. The hybrids without adapter sequences cannot be amplified and are therefore removed from the final library.

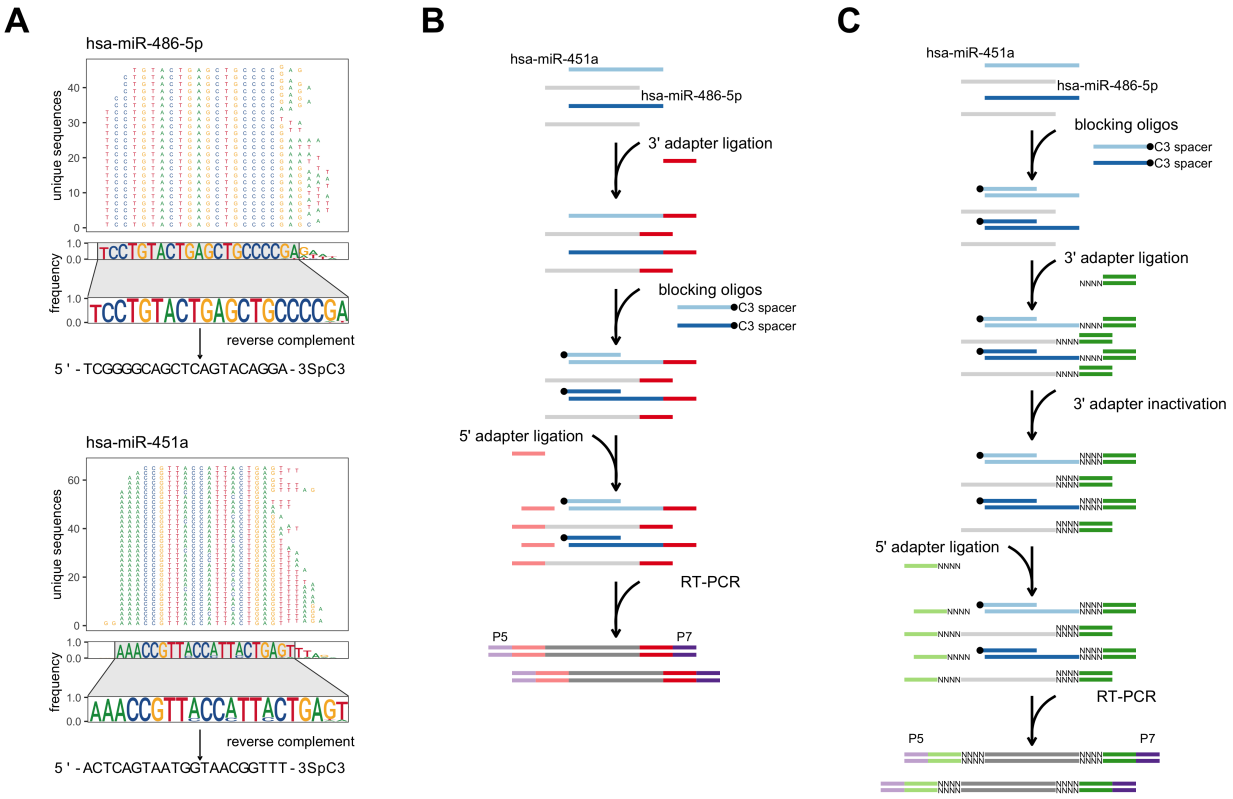

Figure 2: Blocking oligo design and application workflows using modified TruSeq and NEXTflex small RNA library preparation protocols. (**A**) Design principle of miR-486-5p and miR-451a blocking oligonucleotides. Briefly, unique pooled-sample sequences which mapped to precursors of miR-486-5p and miR-451a were used to retrieve the most frequent nucleotide found at each position in a sequence alignment. The most stable consensus sequences were used to generate reverse complement DNA oligonucleotides of targeted miRNAs. The C3 spacer (propyl group) modification was added to the 3' ends of the synthetic oligonucleotides to avoid self-ligation. Whole blood smRNA-seq data used for oligo design was obtained from GSE100467; (**B**) A schematic representation of modified Illumina's TruSeq small RNA library preparation protocol. The modified protocol involves an additional step, where synthetic blocking oligonucleotides are introduced right before the 5' adapter ligation reaction. In this step, the blocking oligonucleotides are annealed to target miRNAs, which results in double-stranded RNA:DNA hybrid formation. These blunt-ended or slight 3' DNA overhang-having double-stranded hybrids are not suitable substrates for T4 RNA ligase-mediated addition of adapter oligonucleotide to the 5' end of RNA strand in the hybrid. As a consequence, blocked RNA:DNA hybrids without 5' adapter sequences cannot be amplified and therefore are depleted from final small RNA library; (**C**) A schematic workflow of modified Perkin Elmer's NEXTflex small RNA library preparation protocol. In comparison to TruSeq, the standard NEXTflex protocol includes an extra step called 3' adapter inactivation, where end-filling is performed to fill the gaps of random nucleotides bearing 5' overhang portions of 3' adapter duplexes. Because of this step, in order to avoid denaturation of the 3' adapter duplexes, blocking oligos for miR-486-5p and miR-451a were introduced directly to total RNA sample.

The standard NEXTflex protocol, besides having randomized adapter ends, also includes an additional step called 3' adapter inactivation. In this step, end-filling is performed to fill the gaps of single-stranded 5' overhang portions of pre-annealed 3' adapters (29). As in the case of "blocked" RNA:DNA hybrids, the 5' end-filled and blunt-ended adapter-oligonucleotide duplexes are poor substrates for T4 RNA ligase, which leads to reduced adapter dimer formation during the 5' adapter ligation step. Due to the 3' adapter inactivation step, we could not introduce our blocking oligos prior to 5' ligation, because in order to anneal the blocking oligos to target miRNAs, the temperature has to be increased to at least 70 °C which could denature 3' adapter-oligonucleotide

duplexes with random nucleotide ends and lead to reduced library output. Therefore, in the modified NEXTflex protocol, we introduced and annealed our blocking oligos directly to total RNA samples prior to library preparation (**Figure 2C**).

Blocking oligonucleotides with high efficiency reduce miR-486-5p and miR-451a sequences in small RNA libraries

In order to test the efficiency of our blocking oligos targeting the erythropoietic miRNAs, for each replicate sample, we generated unblocked (standard) and blocked (modified) libraries using both TruSeq and NEXTflex protocols. With this design, we generated 8 paired libraries for TruSeq and 5 paired libraries for NEXTflex using whole blood total RNA as an input.
As expected, within the libraries that we generated by unmodified protocols, miR-486-5p was the most abundant miRNA in each library independent of the kit. On average, we obtained around ~954K and ~435K counts per million (CPM) of miR-486-5p using unblocked TruSeq and NEXTflex protocols, respectively. When using blocking oligos, we were able to suppress the average CPM values of miR-486-5p down to ~21K and to ~5K in TruSeq and NEXTflex protocols, respectively (**Figure 3A**). In contrast to miR-486-5p, we observed high variability of miR-451a expression within whole blood RNA samples which were prepared using unblocked NEXTflex protocol. The CPM values of this miRNA were also higher in NEXTflex than in TruSeq unblocked libraries. On average, we obtained ~6K and ~66K CPM of miR-451a using unblocked TruSeq and NEXTflex protocols, respectively. When using blocking oligos for miR-486-5p, the average CPM values were suppressed down to ~150 and to ~426 CPM in TruSeq and NEXTflex protocols, respectively (**Figure 3A**).
By calculating blocking efficiencies, we observed that the blocking performance of linear oligonucleotides was slightly better in the NEXTflex than in the TruSeq protocol (P-value = 0.012). The blocking efficiency of the oligonucleotides for miR-486-5p ranged from 90.2% to 99.0% with an average of 97.7% in TruSeq, and from 98.6% to 99.4% with an average of 98.9% in NEXTflex protocol. In case of miR-451a, the blocking efficiency ranged from 89.5% to 99.1% with an average of 96.2% in TruSeq, and from 97.5% to 99.6% with an average of 98.8% in NEXTflex protocol (**Figure 3B**).

Blocking oligonucleotides increase the detectability of low abundant miRNAs in blood-derived RNA samples

To test whether the depletion of erythropoietic miRNAs increases the information in blood-derived small RNA libraries, we compared miRNA detectability in blocked and unblocked libraries prepared by using TruSeq and NEXTflex methods. In the analysis, we considered a miRNA as detected if its CPM value was higher than 1 in at least 75% of the libraries prepared by exactly the same protocol. We detected the highest number of miRNAs in NEXTflex blocked libraries (n = 606), followed by NEXTflex unblocked (n = 521), TruSeq blocked (n = 337) and TruSeq unblocked (n = 186) libraries (**Figure 3C**). Interestingly, when looking at the overlapping and uniquely detected miRNAs, we not only observed uniquely detected miRNAs in the blocked protocols, but we were also able to detect several (n = 14) unique miRNAs in the unblocked NEXTflex protocol (**Figure 3D**). As expected, we found that most of the unique miRNAs were detected in the blocked NEXTflex (n = 82), followed by blocked TruSeq (n = 27) protocols.
To evaluate if the blocking effect on miRNA detectability is stable across libraries at varying sequencing depths, we have performed down-sampling of total mapped reads. We found that the usage of the blocking oligos already at 1 million subsampled reads increases miRNA detectability on average by 33.2% in the TruSeq and by 11.4% in the NEXTflex protocols. The increase of detectability stays more or less stable for both TruSeq (mean: 33.2%; range: 33–34%) and NEXTflex

(mean: 11.4%; range: 10–14%) protocols, even when the subsample size is steadily increased to 5 millions of reads (**Figure 3E**).

Overall, these results clearly show that blocking oligos for miR-486-5p and miR-451a, independent of library size, increase detectability of other miRNA species in whole blood small RNA libraries.

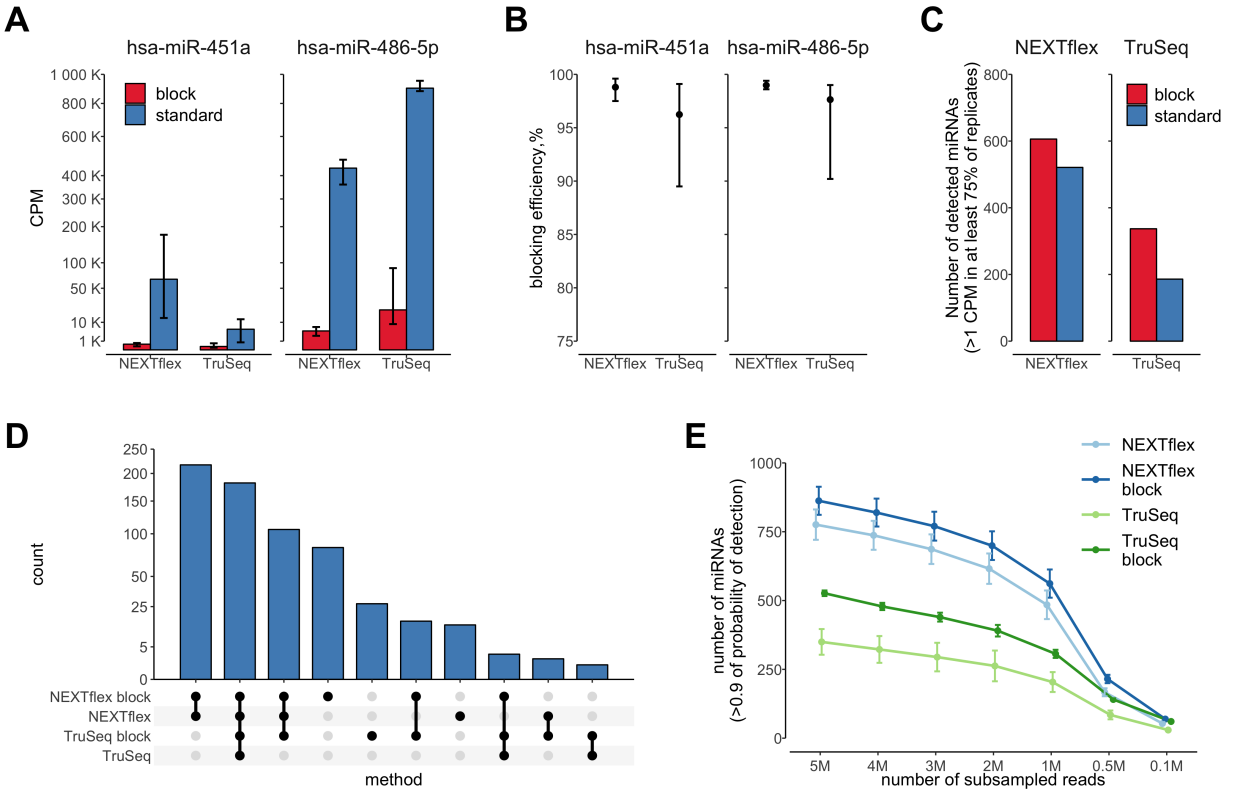

Figure 3: Blocking oligonucleotides efficiently suppress miR-486-5p and miR-451a, and increase detectability of other miRNA species in small RNA libraries. (**A**) A bar chart represents quantitative estimates of miR-486-5p and miR-451a in blocked (red) and unblocked (blue) libraries prepared by NEXTflex and TruSeq protocols. The y-axis depicts counts per million (CPM) and it is scaled by square root. The error bars indicate min and max values obtained from replicates. The graph illustrates a high degree suppression of miR-451a and miR-486-5p sequences in blocked NEXTflex and TruSeq libraries; (**B**) A dot plot represents blocking efficiencies (y-axis) of miR-486-5p and miR-451a in NEXTflex and TruSeq libraries. The data points depict mean values, whereas the error bars indicate min and max values obtained from replicates. The overall blocking efficiency is observed to be slightly better in the NEXTflex than in the TruSeq small RNA libraries (Wilcoxon P-value = 0.012); (**C**) A bar chart shows number of detected miRNA species (y-axis) in blocked (red) and unblocked (blue) libraries prepared by NEXTflex and TruSeq protocols. The detectability of miRNAs is increased in both blocked NEXTflex and TruSeq libraries; (**D**) An upset plot representing intersection of uniquely detected miRNA species amongst the set of the four protocols. The highest numbers of uniquely detected miRNAs are found in blocked libraries; (**E**) A line chart illustrates number of detected miRNAs (y-axis) in subsamples (x-axis) of down-sampled libraries prepared by different protocols. The data points represent mean values, whereas the error bars depict standard errors of the mean. The graph displays a steady increase of detected miRNAs over the increasing size of subsampled miRNA counts.

Blocking oligonucleotides have positive and some negative effects on non-targeted miRNA quantitative estimates

To evaluate the effect of miR-486-5p and miR-451a blocking on non-targeted miRNA species, we compared quantitative estimates of blocked and unblocked paired small RNA libraries which were

generated using TruSeq and NEXTflex protocols.

By looking at the distribution of average log-transformed CPM values of miRNAs, we observed that erythropoietic miRNA depletion resulted in a noticeable shift towards higher values in the density curves of libraries prepared by both TruSeq and NEXTflex protocols, which means that the CPM values of not only lowly but also of highly abundant miRNAs were increased proportionally. This global shift of log-transformed CPM values was more pronounced in the blocked TruSeq than in the blocked NEXTflex protocols (**Figure 4A**). We also observed consistent results when we looked at the paired blocked and unblocked libraries of each sample separately, where we saw an increase in the number of detectable miRNAs as well as a global increase of measured miRNA expression in the blocked libraries of both TruSeq and NEXTflex methods (**Figure 4B**). For both of the library preparation methods, on average, we observed a high concordance of miRNA expression (log-transformed CPM) estimates between paired blocked and unblocked libraries which was 0.94 (range: 0.80–0.97) for TruSeq, and 0.93 (range: 0.91–0.96) for NEXTflex protocols, in terms of Pearson correlation coefficient (**Supplementary Figure 2**).

To obtain further insights whether the blocking oligos may have an impact on measured expression of non-targeted miRNAs, for TruSeq and NEXTflex methods separately, we performed differential expression analysis between blocked and unblocked libraries using DESeq2 with paired sample design. As expected, we saw a global increase of log-transformed fold change values, which was generally higher for low abundant miRNA species in both comparisons of blocked versus unblocked paired libraries (**Figure 4C**; **Supplementary Tables 1 and 2**). We also identified significantly up-regulated high abundant miRNAs, especially in the libraries prepared by TruSeq protocol, which might be explained by 5' ligation bias, because, in the absence of the highly abundant targeted-miRNAs, non-target miRNAs might have different ligation efficiency resulting in a non-proportional increase. In addition to miR-486-5p and miR-451a, unexpectedly, we also observed 14 significantly down-regulated (corrected P-value < 0.01; absolute value of log2 fold change > 1) miRNAs in TruSeq, and 5 miRNAs in NEXTflex libraries. To see whether there is a systemic problem with blocking oligos and down-regulated miRNAs, we looked at the miRNAs which were commonly down-regulated in both methods, and besides miR-486-5p and miR-451a, found two such molecules: miR-339-3p and miR-93-3p (**Figure 4C**). By using RNA cofold analysis we showed that these two miRNAs may interact with blocking oligos of miR-486-5p and form secondary structures which might inhibit or reduce 5' ligation of the non-targeted miRNAs (**Figure 4D**).

Overall, in addition to a number of positive effects such as increased miRNA detectability, global increase of expression values and little effect on non-targeted miRNA species, negative effects of the blocking oligos such as non-specific hybridization may also appear. This should be taken into consideration when analyzing the data.

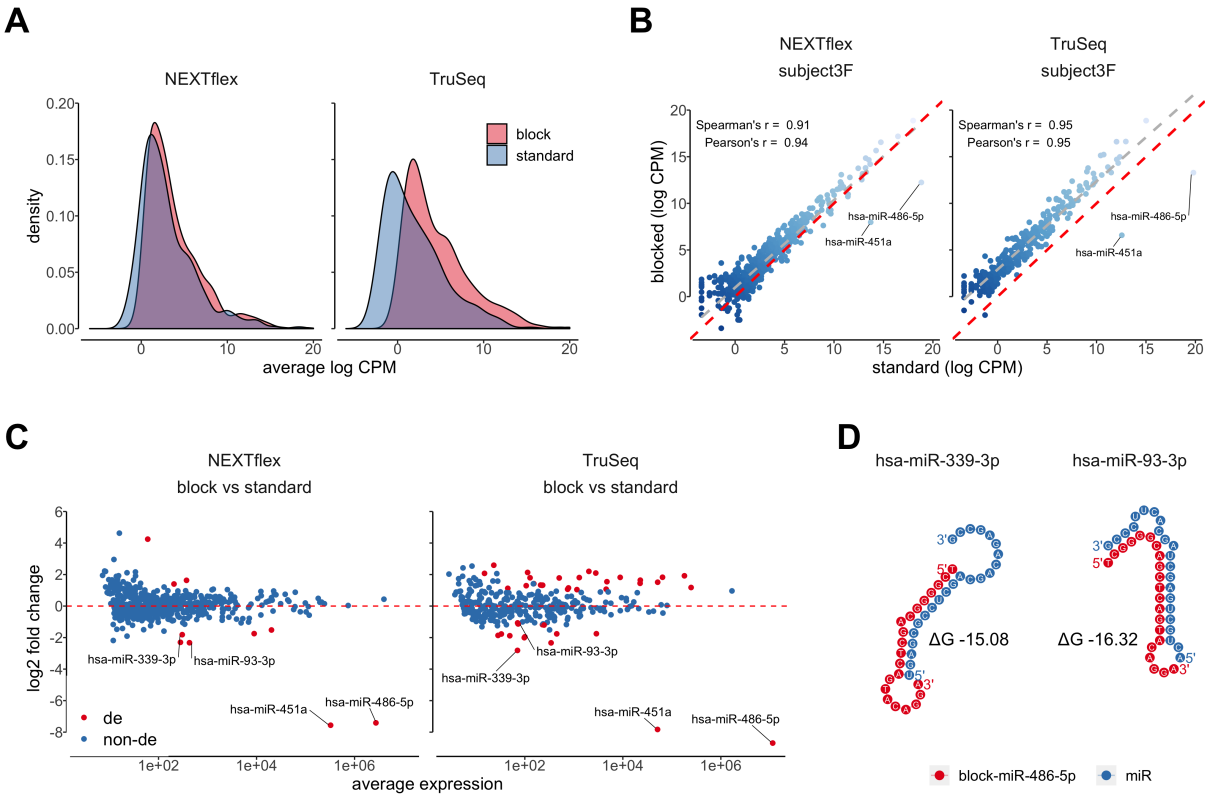

Figure 4: Blocking oligonucleotides for miR-486-5p and miR-451a have positive and some negative effects on measured expression of non-targeted miRNAs. (**A**) A density chart shows distributions of averaged log2-transformed CPM values of blocked (red) and unblocked (blue) libraries prepared by NEXTflex and TruSeq protocols. The expression values of each miRNA were averaged per protocol (x-axis). The graph illustrates a proportional global shift of log2-transformed CPM values towards higher expression estimates in both blocked NEXTflex and TruSeq libraries; (**B**) A scatter plot represents correlation of log2-transformed CPM values of paired blocked (y-axis) and unblocked (x-axis) exemplary libraries generated by NEXTflex and TruSeq protocols. The red dashed line divides panel in two equal parts, whereas the grey dashed line displays a linear regression curve. The chart illustrates a high concordance of miRNA expression values between blocked and unblocked paired libraries; (**C**) An MA plot shows paired-differential expression analysis results of blocked versus unblocked libraries, where log2 fold changes are presented on the y-axis and averaged normalized counts on the x-axis. The red colored dots indicate significantly differentially expressed miRNAs (corrected P-value < 0.01; absolute value of log2 fold change > 1). The miRNAs which were found to be down-regulated in both NEXTflex and TruSeq libraries are labeled with miRNA names; (**D**) A dot chart represents predicted non-specific interactions between miR-486-5p blocking oligonucleotide (red) and two non-targeted miRNA molecules (blue). These interactions might cause decreased expression of miR-339-3p and miR-93-3p in blocked small RNA libraries.

Discussion

The erythroid-specific, highly dominant and less likely informative miR-486-5p and miR-451a transcripts are muting detection of lowly abundant miRNAs in whole blood-derived small RNA libraries. To overcome this problem, we have developed a cost-effective and easy-to-use hybridization-based method to deplete these erythropoietic miRNAs from small RNA libraries. This method, besides custom oligos, does not require any other additional reagents, and is easily compatible and adjustable with 5' ligation-based small RNA library preparation methods such as Illumina's TruSeq and Perkin Elmer's NEXTflex protocols.

We demonstrate the high overall blocking efficiency of our oligonucleotides, whereas, the average efficiency was slightly better in NEXTflex (98,9%) than in TruSeq (96,9%) libraries. As a consequence of erythropoietic miRNA blocking, the measured expression as well as detectability of low abundant miRNA species was considerably increased. This increase was more pronounced in the TruSeq than in the NEXTflex libraries, probably due to higher domination of miR-486-5p in the unblocked TruSeq libraries. It seems that this particular miRNA is highly preferred by TruSeq method (~90% of total mapped reads) and once it is depleted, much more space is freed up for other miRNA species. Even though the relative increase of detectability is higher in blocked TruSeq libraries, the nominal detectability is much higher in the blood RNA-derived NEXTflex libraries, which bear more miRNA species than TruSeq libraries even without the use of blocking oligos. This is probably due to utilization of degenerate adapters which reduce ligation bias in small RNA libraries.

We also demonstrate that our method does not reduce the reproducibility of the quantitative estimates of non-targeted miRNAs and, moreover, does not remove or significantly disturb individual-specific biological variation. In terms of Spearman and Pearson correlation coefficients, the reproducibility of blocked and unblocked libraries are very similar in both TruSeq and NEXTflex libraries. Despite good performance of our hybridization-based method, we also detected some off-target effects, which were observed independent of library preparation method, suggesting a systemic effect of blocking oligos on at least two non-targeted miRNAs. On the other hand, since it is a systemic effect, in studies such as differential expression between cases and controls this effect should even out.

In the current version of the protocol, we have designed and optimized the blocking oligonucleotides and their hybridization conditions specifically for miR-486-5p and miR-451a; however, the oligo design principles can be adapted to target any other miRNA molecule. Of note, each tissue or cell type might contain different isomiR composition (30), and therefore, we would recommend to always ensure that the designed oligo covers all nucleotides at the 5' end of the target sequence. We demonstrated compatibility of this method with TruSeq and NEXTflex protocols; however, theoretically the blocking oligos should be also compatible with other 5' ligation-based methods which employ T4 RNA ligase (such as NEBNext, QIAseq, CleanTag, etc.) to attach an adapter oligonucleotide to the 5' end of small RNA molecules.

References

[1] David P. Bartel. Metazoan MicroRNAs. *Cell*, 173(1):20–51, mar 2018. ISSN 00928674. doi:10.1016/j.cell.2018.03.006.

[2] Alexander Link and Juozas Kupcinskas. MicroRNAs as non-invasive diagnostic biomarkers for gastric cancer: Current insights and future perspectives. *World Journal of Gastroenterology*, 24(30):3313–3329, aug 2018. ISSN 1007-9327. doi:10.3748/wjg.v24.i30.3313.

[3] Kristina Vartanian, Rachel Slottke, Timothy Johnstone, Amanda Casale, Stephen R Planck, Dongseok Choi, Justine R Smith, James T Rosenbaum, and Christina A Harrington. Gene expression profiling of whole blood: Comparison of target preparation methods for accurate and reproducible microarray analysis. *BMC Genomics*, 10(1):2, jan 2009. ISSN 1471-2164. doi:10.1186/1471-2164-10-2.

[4] Heesun Shin, Casey P. Shannon, Nick Fishbane, Jian Ruan, Mi Zhou, Robert Balshaw, Janet E. Wilson-McManus, Raymond T. Ng, Bruce M. McManus, Scott J. Tebbutt, and for the PROOF Centre of Excellence Team. Variation in RNA-Seq Transcriptome Profiles of Peripheral Whole Blood from Healthy Individuals with and without Globin Depletion. *PLoS ONE*, 9(3):e91041, mar 2014. ISSN 1932-6203. doi:10.1371/journal.pone.0091041.

[5] C. C. Pritchard, E. Kroh, B. Wood, J. D. Arroyo, K. J. Dougherty, M. M. Miyaji, J. F. Tait, and M. Tewari. Blood Cell Origin of Circulating MicroRNAs: A Cautionary Note for Cancer Biomarker Studies. *Cancer Prevention Research*, 5(3):492–497, mar 2012. ISSN 1940-6207. doi:10.1158/1940-6207.CAPR-11-0370.

[6] David Jee, Jr-Shiuan Yang, Sun-Mi Park, D'Juan T Farmer, Jiayu Wen, Timothy Chou, Arthur Chow, Michael T McManus, Michael G Kharas, and Eric C Lai. Dual Strategies for Argonaute2-Mediated Biogenesis of Erythroid miRNAs Underlie Conserved Requirements for Slicing in Mammals. *Molecular cell*, 69(2):265–278.e6, jan 2018. ISSN 1097-4164. doi:10.1016/j.molcel.2017.12.027.

[7] Rimi Hamam, Dana Hamam, Khalid A Alsaleh, Moustapha Kassem, Waleed Zaher, Musaad Alfayez, Abdullah Aldahmash, and Nehad M Alajez. Circulating microRNAs in breast cancer: novel diagnostic and prognostic biomarkers. *Cell Death & Disease*, 8(9):e3045–e3045, sep 2017. ISSN 2041-4889. doi:10.1038/cddis.2017.440.

[8] Brian S. Roberts, Andrew A. Hardigan, Marie K. Kirby, Meredith B. Fitz-Gerald, C. Mel Wilcox, Robert P. Kimberly, and Richard M. Myers. Blocking of targeted microRNAs from next-generation sequencing libraries. *Nucleic Acids Research*, 43(21):gkv724, jul 2015. ISSN 0305-1048. doi:10.1093/nar/gkv724.

[9] Andrew A Hardigan, Brian S Roberts, Dianna E Moore, Ryne C Ramaker, Angela L Jones, and Richard M Myers. CRISPR/Cas9-targeted removal of unwanted sequences from small-RNA sequencing libraries. *Nucleic Acids Research*, jun 2019. ISSN 0305-1048. doi:10.1093/nar/gkz425.

[10] Simonas Juzenas, Geetha Venkatesh, Matthias Hübenthal, Marc P Hoeppner, Zhipei Gracie Du, Maren Paulsen, Philip Rosenstiel, Philipp Senger, Martin Hofmann-Apitius, Andreas Keller, Limas Kupcinskas, Andre Franke, and Georg Hemmrich-Stanisak. A comprehensive, cell specific microRNA catalogue of human peripheral blood. *Nucleic Acids Research*, 45(16): 9290–9301, 2017. ISSN 13624962. doi:10.1093/nar/gkx706.

[11] Jari Oksanen, Guillaume F. Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens, Eduard Szoecs, and Helene Wagne. vegan: Community Ecology Package, 2019. URL https://cran.r-project.org/package=vegan.

[12] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):pp. 10–12, 2011. ISSN 2226-6089. doi:10.14806/ej.17.1.200.

[13] Ana Kozomara, Maria Birgaoanu, and Sam Griffiths-Jones. miRBase: from microRNA sequences to function. *Nucleic Acids Research*, 47(D1):D155–D162, jan 2019. ISSN 0305-1048. doi:10.1093/nar/gky1141.

[14] L. Pantano, X. Estivill, and E. Marti. SeqBuster, a bioinformatic tool for the processing and analysis of small RNAs datasets, reveals ubiquitous miRNA modifications in human embryonic cells. *Nucleic Acids Research*, 38(5):e34–e34, mar 2010. ISSN 0305-1048. doi:10.1093/nar/gkp1127.

[15] Lorena Pantano and Georgia Escaramis. isomiRs: Analyze isomiRs and miRNAs from small RNA-seq. 2019. doi:10.18129/B9.bioc.isomiRs.

[16] Ron Edgar, Michael Domrachev, and Alex E Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–10, jan 2002. ISSN 1362-4962. doi:10.1093/nar/30.1.207.

[17] Davis J. McCarthy, Yunshun Chen, and Gordon K. Smyth. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, 40(10):4288–4297, may 2012. ISSN 1362-4962. doi:10.1093/nar/gks042.

[18] Constantin Ahlmann-Eltze. ggupset: Combination Matrix Axis for 'ggplot2' to Create 'UpSet' Plots, 2019. URL https://cran.r-project.org/web/packages/ggupset/index.html.

[19] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15(12):550, dec 2014. ISSN 1465-6914. doi:10.1186/s13059-014-0550-8.

[20] J M Bland and D G Altman. Multiple significance tests: the Bonferroni method. *BMJ (Clinical research ed.)*, 310(6973):170, jan 1995. ISSN 0959-8138. doi:10.1136/bmj.310.6973.170.

[21] Ronny Lorenz, Stephan H Bernhart, Christian Höner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6(1):26, dec 2011. ISSN 1748-7188. doi:10.1186/1748-7188-6-26.

[22] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009. ISBN 978-0-387-98140-6. URL http://ggplot2.org.

[23] Sean M Davidson, Ioanna Andreadou, Lucio Barile, Yochai Birnbaum, Hector A Cabrera-Fuentes, Michael V Cohen, James M Downey, Henrique Girao, Pasquale Pagliaro, Claudia Penna, John Pernow, Klaus T Preissner, and Péter Ferdinandy. Circulating blood cells and extracellular vesicles in acute cardioprotection. *Cardiovascular Research*, 115(7):1156–1166, jun 2019. ISSN 0008-6363. doi:10.1093/cvr/cvy314.

[24] Paula M. Godoy, Nirav R. Bhakta, Andrea J. Barczak, Hakan Cakmak, Susan Fisher, Tippi C. MacKenzie, Tushar Patel, Richard W. Price, James F. Smith, Prescott G. Woodruff, and David J. Erle. Large Differences in Small RNA Composition Between Human Biofluids. *Cell Reports*, 25(5):1346–1358, oct 2018. ISSN 2211-1247. doi:10.1016/J.CELREP.2018.10.014.

[25] Anne E. Magurran. *Measuring Biological Diversity*. John Wiley & Sons, 2013. ISBN 9781118687925.

[26] Luca F. R. Gebert and Ian J. MacRae. Regulation of microRNA function in animals. *Nature Reviews Molecular Cell Biology*, page 1, aug 2018. ISSN 1471-0072. doi:10.1038/s41580-018-0045-7.

[27] Michelle L Wickersheim and Justin P Blumenstiel. Terminator oligo blocking efficiently eliminates rRNA from Drosophila small RNA sequencing libraries. *BioTechniques*, 55(5):269–72, nov 2013. ISSN 1940-9818. doi:10.2144/000114102.

[28] Markus Hafner, Neil Renwick, Miguel Brown, Aleksandra Mihailović, Daniel Holoch, Carolina Lin, John T G Pena, Jeffrey D Nusbaum, Pavel Morozov, Janos Ludwig, Tolulope Ojo, Shujun Luo, Gary Schroth, and Thomas Tuschl. RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. *RNA (New York, N.Y.)*, 17(9):1697–712, sep 2011. ISSN 1469-9001. doi:10.1261/rna.2799511.

[29] Masoud Toloue, Adam R. Morris, and Kevin D. Allen. Methods and kits for reducing adapter-dimer formation. U.S. Patent 15/354,491, nov 2016.

[30] Matthew N McCall, Min-Sik Kim, Mohammed Adil, Arun H Patil, Yin Lu, Christopher J Mitchell, Pamela Leal-Rojas, Jinchong Xu, Manoj Kumar, Valina L Dawson, Ted M Dawson, Alexander S Baras, Avi Z Rosenberg, Dan E Arking, Kathleen H Burns, Akhilesh Pandey, and Marc K Halushka. Toward the human cellular microRNAome. *Genome research*, sep 2017. ISSN 1549-5469. doi:10.1101/gr.222067.117.