

RESEARCH

Open Access



cPAS-based sequencing on the BGISEQ-500 to explore small non-coding RNAs

Tobias Fehlmann¹, Stefanie Reinheimer³, Chunyu Geng^{2*}, Xiaoshan Su², Snezana Drmanac^{2,4}, Andrei Alexeev^{2,4}, Chunyan Zhang², Christina Backes¹, Nicole Ludwig³, Martin Hart³, Dan An², Zhenzhen Zhu², Chongjun Xu^{2,4}, Ao Chen², Ming Ni², Jian Liu², Yuxiang Li², Matthew Poulter², Yongping Li², Cord Stähler¹, Radoje Drmanac^{2,4}, Xun Xu^{2*}, Eckart Meese³ and Andreas Keller^{1*}

Abstract

Background: We present the first sequencing data using the combinatorial probe-anchor synthesis (cPAS)-based *BGISEQ-500* sequencer. Applying cPAS, we investigated the repertoire of human small non-coding RNAs and compared it to other techniques.

Results: Starting with repeated measurements of different specimens including solid tissues (brain and heart) and blood, we generated a median of 30.1 million reads per sample. 24.1 million mapped to the human genome and 23.3 million to the *miRBase*. Among six technical replicates of brain samples, we observed a median correlation of 0.98. Comparing *BGISEQ-500* to HiSeq, we calculated a correlation of 0.75. The comparability to microarrays was similar for both *BGISEQ-500* and HiSeq with the first one showing a correlation of 0.58 and the latter one correlation of 0.6. As for a potential bias in the detected expression distribution in blood cells, 98.6% of HiSeq reads versus 93.1% of *BGISEQ-500* reads match to the 10 miRNAs with highest read count. After using *miRDeep2* and employing stringent selection criteria for predicting new miRNAs, we detected 74 high-likely candidates in the cPAS sequencing reads prevalent in solid tissues and 36 candidates prevalent in blood.

Conclusions: While there is apparently no ideal platform for all challenges of miRNome analyses, cPAS shows high technical reproducibility and supplements the hitherto available platforms.

Keywords: Next-generation sequencing, miRNA, Biomarker discovery, *BGISEQ*

Background

Currently, high-throughput analytical techniques are massively applied to further the understanding of the non-coding transcriptome [1]. Still, the full complexity of non-coding RNAs is only partially understood. One class of well-studied non-coding RNAs comprises small oligonucleotides, so-called miRNAs [2, 3].

Among the techniques most commonly used for miRNA profiling are microarrays, RT-qPCR, and next-generation sequencing (NGS), also referred to as high-throughput sequencing (HTS). An excellent review on the different platforms and a cross-platform comparison has been recently published [4]. A detailed examination

of technologies, however, frequently reveals a bias. One reason for the respective bias is the ligation step, as, e.g., reported by Hafner and co-workers [5]. For example, the quantification of miRNAs differs between NGS and microarrays as it is dependent on base composition [6]. Especially, the guanine and uracil content of a miRNA seems to influence the abundance depending on the platform used. A substantial strength of NGS is the ability to support the completion of the non-coding transcriptome. Unlike microarrays and RT-qPCR, NGS allows the discovery of novel miRNA candidates. To this end, different algorithms have been implemented, with *miRDeep* being one of the most popular ones [7]. A substantial part of small RNA sequencing data has been obtained using HiSeq and MiSeq platforms (Illumina) based on stepwise sequencing by polymerase on DNA microarrays prepared by bridge PCR [8], as well as the

* Correspondence: gengchunyu@genomics.cn; xuxun@genomics.cn; andreas.keller@ccb.uni-saarland.de

²BGI-Shenzhen, Shenzhen, China

¹Clinical Bioinformatics, Saarland University, 66125 Saarbrücken, Germany

Full list of author information is available at the end of the article



IonTorrent systems from Thermo Fisher Scientific using a different type of polymerase-based stepwise sequencing on micro-bead arrays generated by emulsion PCR, the first method proposed for making microarrays for massively parallel sequencing [9]. Another approach is the ligase-based stepwise sequencing also using micro-bead arrays, applied for example by ThermoFisher Scientific's SOLiD sequencing platform, and which has also been used to analyze and present novel miRNAs [10].

In the current study, we applied the new combinatorial probe-anchor synthesis (cPAS)-based BGISEQ-500 sequencing platform that combines DNA nanoball (DNB) nanoarrays [11] with stepwise sequencing using polymerase. An important advantage of this technique compared to the previously mentioned sequencing systems is in that no PCR is applied in preparing sequencing arrays. Applying cPAS, we investigated the human non-coding transcriptome. We first evaluated the reproducibility of sequencing on standardized brain and heart samples, then compared the performance to Agilent's microarray technique and finally evaluated blood samples. Using the web-based miRNA analysis pipeline *miRmaster* and the tool *novoMiRank* [12], we finally predicted 135 new high-likely miRNA candidates specific for tissue and 35 new miRNA candidates specific for blood samples.

Methods

Samples

In this study, we examined the performance of three sample types using three techniques for high-throughput miRNA measurements (Illumina's HiSeq sequencer, Agilent's miRBase microarrays, and BGI's BGISEQ-500 sequencing system, see details below). The three specimens were standardized HBRR sample ordered from Ambion (catalog number AM6051) and UHRR sample ordered from Agilent (catalog number 740000). UHRR and HBRR samples were measured in two and six replicates, respectively. As third sample type, we used *PAXGene* blood tubes. Here, two healthy volunteers' blood samples were collected and miRNAs were extracted using PAXgene Blood RNA Kit (Qiagen) according to manufacturer's protocol. The study has been approved by the local ethics committee.

Next-generation sequencing using BGISEQ-500

We prepared the libraries starting with 1 μ g total RNA for each sample. Firstly, we isolated the microRNAs (miRNA) by 15% urea-PAGE gel electrophoresis and cut the gel from 18 to 30 nt, which corresponds to mature miRNAs and other regulatory small RNA molecules. After gel purification, we ligated the adenylated 3' adapter to the miRNA fragment. Secondly, we used the RT primer with barcode to anneal the 3' adenylated adapter in order to combine the redundant unligated 3'

adenylated adapter. Then, we ligated the 5' adapter and did reverse transcript (RT) reaction. After cDNA first strand synthesis, we amplified the product by 15 cycles. We then carried out the second size selection operation and selected 103–115 bp fragments from the gel. This step was conducted in order to purify the PCR product and remove any nonspecific products. After gel purification, we quantified the PCR yield by Qubit (Invitrogen, Cat No. Q33216) and pooled samples together to make a single strand DNA circle (ssDNA circle), which gave the final miRNA library.

DNA nanoballs (DNBs) were generated with the ssDNA circle by rolling circle replication (RCR) to enlarge the fluorescent signals at the sequencing process as previously described [11]. The DNBs were loaded into the patterned nanoarrays and single-end read of 50 bp were read through on the BGISEQ-500 platform for the following data analysis study. For this step, the BGISEQ-500 platform combines the DNA nanoball-based nanoarrays [11] and stepwise sequencing using polymerase, as previously published [13–15]. The new modified sequencing approach provides several advantages, including among others high throughput and quality of patterned DNB nanoarrays prepared by linear DNA amplification (RCR) instead of random arrays by exponential amplification (PCR) as, e.g., used by Illumina's HiSeq and longer reads of polymerase-based cycle sequencing compared to the previously described combinatorial probe-anchor ligation (cPAL) chemistry on DNB nanoarrays [11]. The usage of linear DNA amplification instead of exponential DNA amplification to make sequencing arrays results in lower error accumulation and sequencing bias.

Next-generation sequencing using HiSeq

Samples have been sequenced using Illumina HiSeq sequencing according to manufacturer's instructions and as previously described [16, 17].

Agilent microarray measurements

For detection of known miRNAs, we used the SurePrint G3 8 \times 60k miRNA microarray (miRBase version 21, Agilent Technologies) containing probes for all miRNAs from miRBase version 21 in conjunction with the miRNA Complete Labeling and Hyb Kit (Cat. No. 5190-0456) according to the manufacturer's recommendations. In brief, 100 ng total RNA including miRNAs was dephosphorylated with calf intestine phosphatase. After denaturation, Cy3-pCp was ligated to all RNA fragments. Labeled RNA was then hybridized to an individual 8 \times 60k miRNA microarray. After washing, array slides were scanned using the Agilent Microarray Scanner G2565BA with 3- μ m resolution in double-pass mode. Signals were retrieved using Agilent AGW Feature Extraction software (version 10.10.11).

Data availability

The new sequencing data using BGISEQ-500 data are available in the Additional file of this manuscript (Additional file 1: Table S3).

Bioinformatics analysis

The raw reads were collapsed and used as input for the web-based tool miRMaster, allowing for integrated analysis of NGS miRNA data. On the server side, mapping to the human genome was carried out using *Bowtie* [18] (one mismatch allowed). miRNAs were quantified similar to the popular *miRDeep2* [19] algorithm. The prediction of novel miRNAs was performed using an extended feature set built up on *novoMiRank* [12]. For classification, an *AdaBoost* model using decision trees was applied. Novel miRNAs were cross-checked against other RNA resources, including the *miRBase* [20], *NONCODE2016* [21], and *Ensembl* non-coding RNAs. The assessment of the quality of new miRNAs was carried out using the *novoMiRank* algorithm. A downstream analysis of results including cluster analysis was performed using R. For target prediction, we applied TargetScan 7.1 (http://www.targetscan.org/vert_71/) and predicted for all new miRNAs the targets. With the predictions, we extracted the context ++ scores and used them for prioritizing the targets, miRNA-target interactions with context++ scores below 1 were considered as high-likelihood targets. Target networks were constructed using an offline version of MiR-TargetLink [22] and visualized in Cytoscape. miRNA target pathway analysis has been carried out using GeneTrail2 [23]. For the GeneTrail2 analysis, all available categories were analyzed, the minimal category size was set to 4 and all *p* values were adjusted using Benjamini-Hochberg adjustment.

Results

Raw data analysis

We sequenced six brain, two heart, and two blood samples using the BGISEQ-500 system. The resulting reads were mapped to the human genome allowing one mismatch per read. The 10 samples had a median of 30.1 million reads. Of these, 24.1 million reads mapped to the human genome and 23.3 million reads to miRNAs annotated in the human miRBase version 21. The remaining 0.7 million reads per sample contain potentially new miRNAs.

Technical reproducibility of the BGISEQ-500 and comparison to microarrays

To assess the technical reproducibility of the sequencing platform, we evaluated the six technical replicates of the human brain sample (see correlation matrix in Fig. 1). The median correlation between the six replicates was 0.98, and the 25 and 75% quantile were 0.98 and 0.99, respectively. These data suggest an overall high

correlation for technical replicates on the BGISEQ-500 platform.

Comparing the BGISEQ-500 data to the measurements of the brain sample with microarrays (miRBase version 21) that have also been carried out as six technical replicates (median correlation of the microarrays was 0.999), we observed a log correlation of 0.48. A direct comparison is presented in the scatter plot in Fig. 2a. This plot highlights many miRNAs that can be measured at a comparable level on both platforms. However, a subset of the small non-coding RNAs is shifted towards higher expression on the array platform. The same behavior can be observed in the cluster heat map in Fig. 2b. This heat map graphically represents the 50 miRNAs with most different detection between both techniques. To compare rather the ranks of miRNAs instead of the absolute read counts, the replicated brain samples on both platforms were jointly quantile normalized. Three miRNAs, in particular, showed highly significant deviations (multiple testing adjusted *p* values below 10^{-20}). Hsa-miR-8069 was almost not detected in the BGISEQ-500 but had 0.9 million normalized intensity counts on the array platform, hsa-miR-4454 had 51.6 normalized reads on the BGISEQ-500 versus 1.9 million normalized counts on the microarrays, and hsa-miR-7977 had 343.2 normalized reads on the BGISEQ-500 versus 1.3 million normalized counts on the microarrays. This means that the three miRNAs were orders of magnitudes more abundant on microarrays as compared to the sequencing system. The secondary structures of the three precursors are presented in Additional file 2: Figure S1. These results match well to previously published platform comparisons between NGS and microarrays [6]. Here, several miRNAs such as hsa-miR-941 (not detected in any array experiment, not detected in RT-qPCR, average read count of ~1000 reads using Illumina HiSeq sequencing) had expression levels differing several orders of magnitude between the miRBase microarrays and using HiSeq sequencing.

The full list of miRNAs with raw and adjusted *p* values in *t* test and Wilcoxon-Mann-Whitney test comparing BGISEQ-500 and microarrays is presented in Additional file 3: Table S1. Overall, the results are well in-line with those obtained between HiSeq NGS and the same microarray platform [6]. Reasons that explain differences between arrays and NGS include different sensitivity levels of the platforms, cross-hybridization of miRNAs with similar sequences on the microarrays or bias in library preparation. Further, effects of the normalization can lead to variations in miRNA quantification.

Biological replicates of blood samples and comparison to other platforms

One of the most promising applications in small RNA analysis is biomarker profiling in body fluids. We

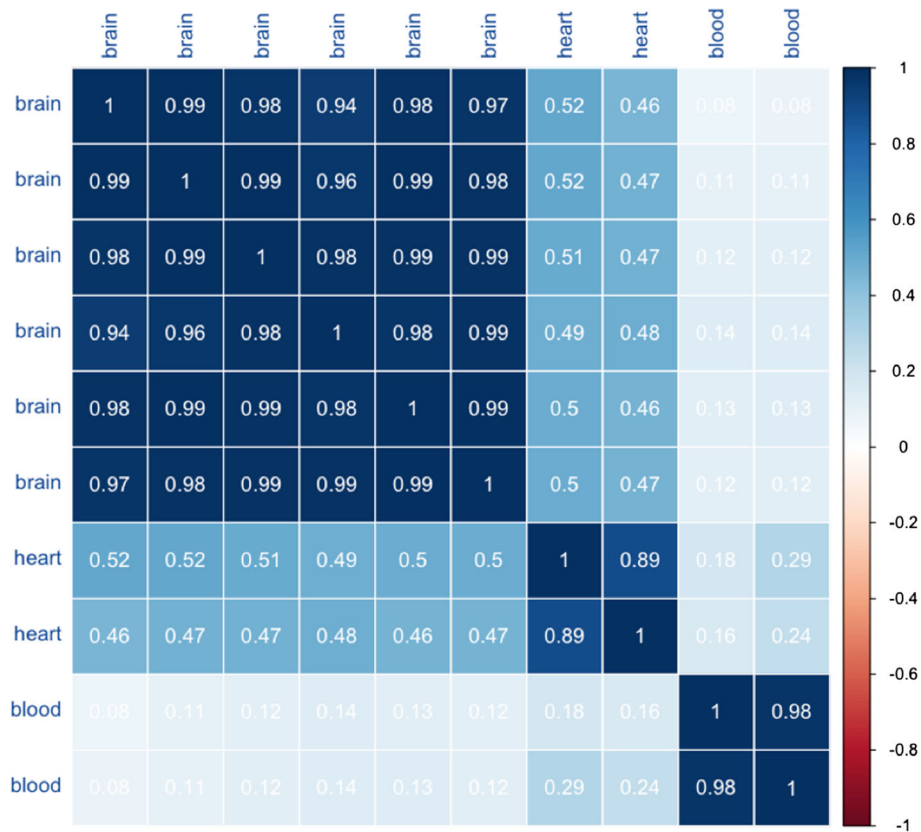


Fig. 1 Correlation matrix of the brain (six technical replicates), heart (two technical replicates), and blood (two biological replicates) sequenced by the BGISEQ-500 system

previously analyzed over 2000 blood samples on Agilent microarrays [17, 24, 25] and about 1000 samples using HiSeq sequencing [26, 27] and compared both platforms [6]. We correlated two newly sequenced blood samples using the BGISEQ-500 system to the data generated by HiSeq and Agilent microarrays. When interpreting the results, it is important to keep in mind that the microarrays and HiSeq data are from the same samples [6] while the newly sequenced blood drawings are from other individuals and thus biological but no technical replicates. To minimize a potential bias between the platforms with respect to different miRNA sets, we first reduced the marker set to the 2525 human miRNAs that were profiled on all platforms and next to the subset of 658 miRNAs that were discovered in all three platforms. For each, platform data were normalized using quantile normalization. Due to the wide dynamic range of miRNAs in blood samples, which is approximately 10^7 , we present the three pairwise comparisons (BGISEQ-500 to microarrays, BGISEQ-500 to HiSeq, and HiSeq to microarrays) on a log scale. The scatter plots are presented in Fig. 3. The highest correlation was observed for BGISEQ-500 to Illumina (0.75, Fig. 3a). Even the correlation between microarrays and HiSeq was below this

value (0.6, Fig. 3c). Especially since technical replicates have been measured for these platforms, the increased correlation of sequencing platforms is remarkable. The comparison of BGISEQ-500 and microarrays revealed correlation values in the same range as for the brain samples (0.58, Fig. 3b). The 3D scatter plot in Fig. 3d compares the expression of the three platforms directly to each other. The coloring of the miRNAs has been carried out with respect to the GC content.

Expression distribution of miRNAs

As mentioned, miRNA expression is highly variable and can scatter across many orders of magnitude. We thus compared the distribution of the sequencing reads in blood samples on the HiSeq to the BGISEQ-500. Blood samples, including blood cells (especially red blood cells) are known to be enriched for few miRNAs that are highly expressed. The diagram in Fig. 4 (panel A) highlights that 90.8% of all blood sequencing reads from the HiSeq match to one single miRNA: hsa-miR-486-5p. The second most abundant miRNA miR-92a-3p takes further 5.5%, and already the third most abundant marker miR-451a has below 1% of all reads. In sum, 98.6% of all reads match to the top 10 miRNAs. For the

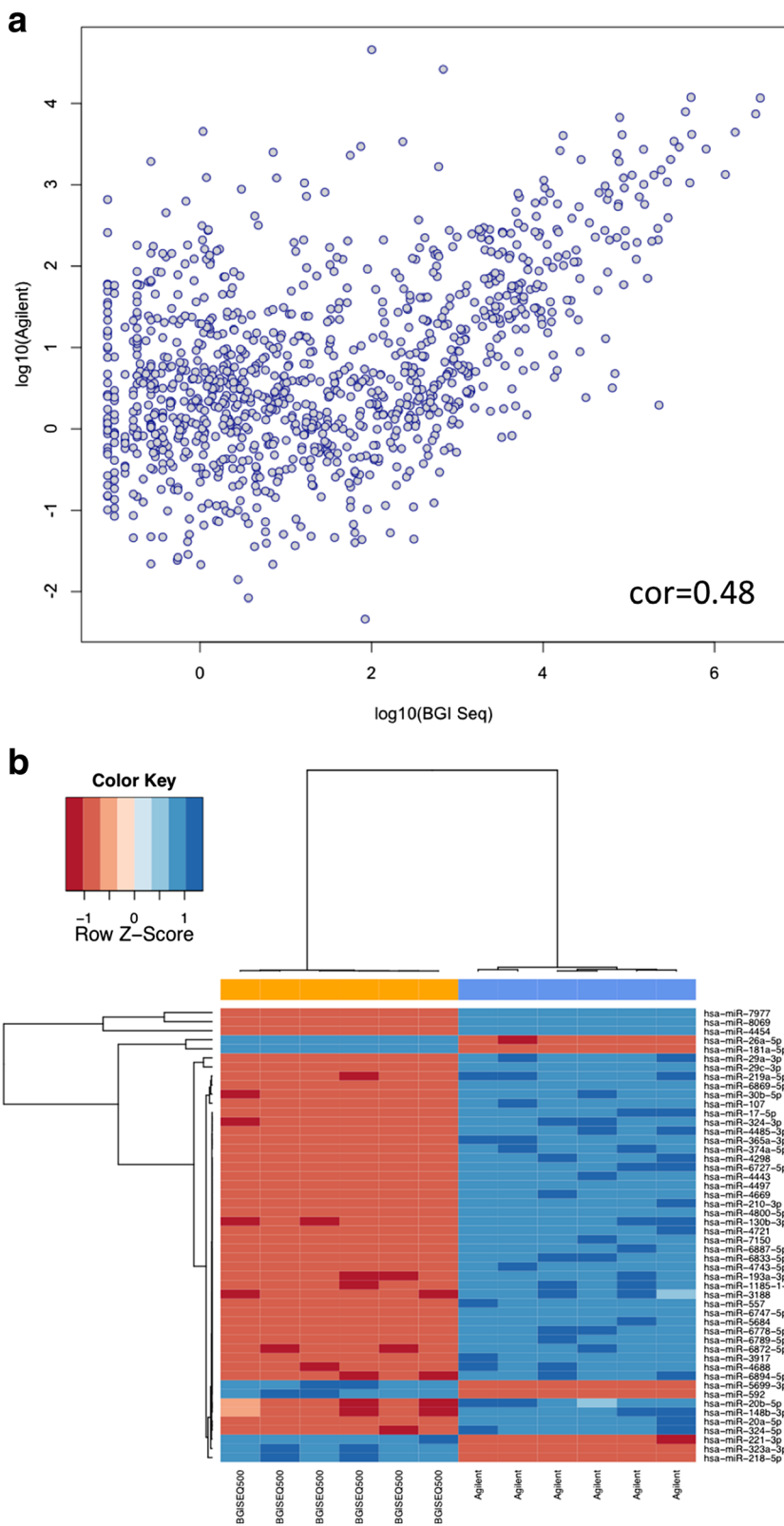
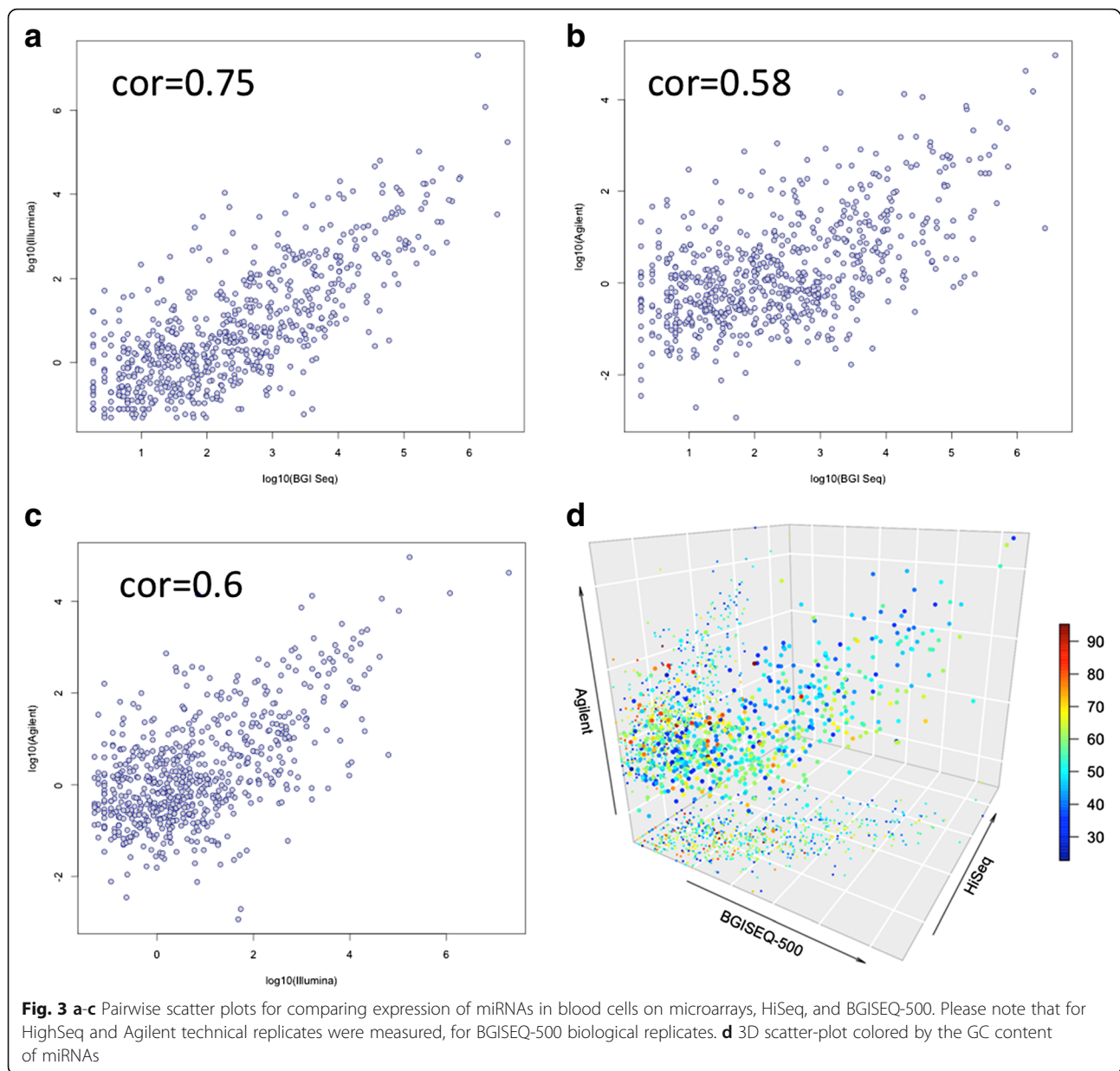


Fig. 2 a Log average expression of common miRNAs for the brain RNA on BGISEQ-500 and on Agilent microarrays (six technical replicates each). **b** Heat map with dendrogram for the 50 most differently detected miRNAs in the brain RNA between Agilent and BGISEQ-500 (six technical replicates each)



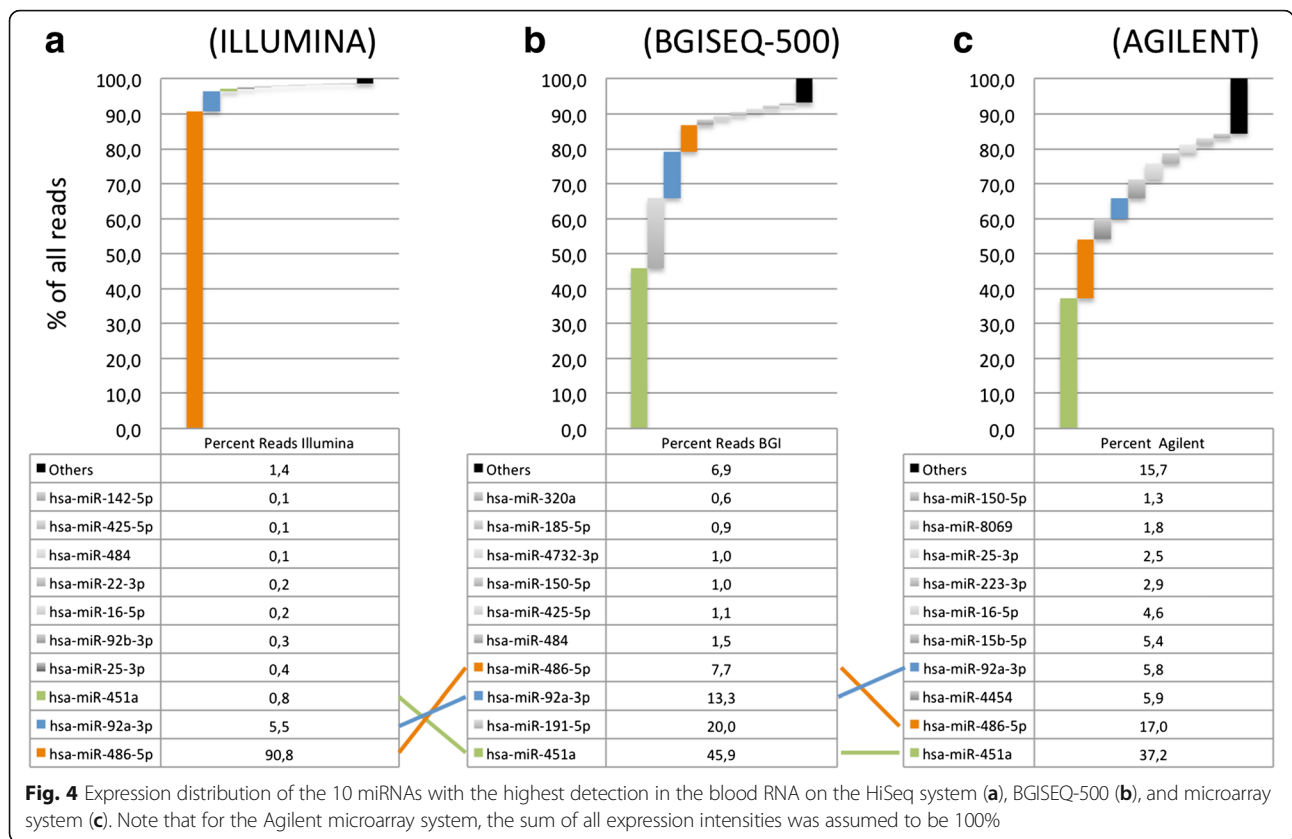
BGISEQ-500 (panel B), 45.9% of reads match to miR-451a, further 20% map to miR-191-5p and 13.3% map to miR-92a-3p. The most abundant miRNA in HiSeq, miR-486-5p, is detected in 7.7% of all reads. 93.1% of all sequenced reads match to the top 10 miRNAs.

Comparison of the distribution and abundance of miRNAs on the microarray platform is difficult since microarrays show a saturation effect. This means that for two miRNAs expressed in a range above the saturation, no difference can be observed. We nonetheless performed the same analysis as presented above, assuming that the sum of all expression counts equals to 100%. In this analysis, miR-451a which is found in 0.8% of HiSeq reads and 45.9% of BGISEQ-500 reads is the highest expressed

in microarrays (37.2% of all expression counts), followed by 17% of miR-486-5p.

Prediction of novel miRNAs

Predicting new miRNAs from NGS data is a challenging task since many false positive miRNA candidates are observed. We implemented our own prediction tool for miRNAs from NGS data and filtered the candidates stringently to reduce the false discovery rate. Without any filtering steps, our initial predictor trimmed for maximizing the ROC AUC returned 25,086 candidates across all samples. The exclusion of the candidates with low abundance (less than 10 total reads) reduced the number of candidates to around 10% (2354 candidates).

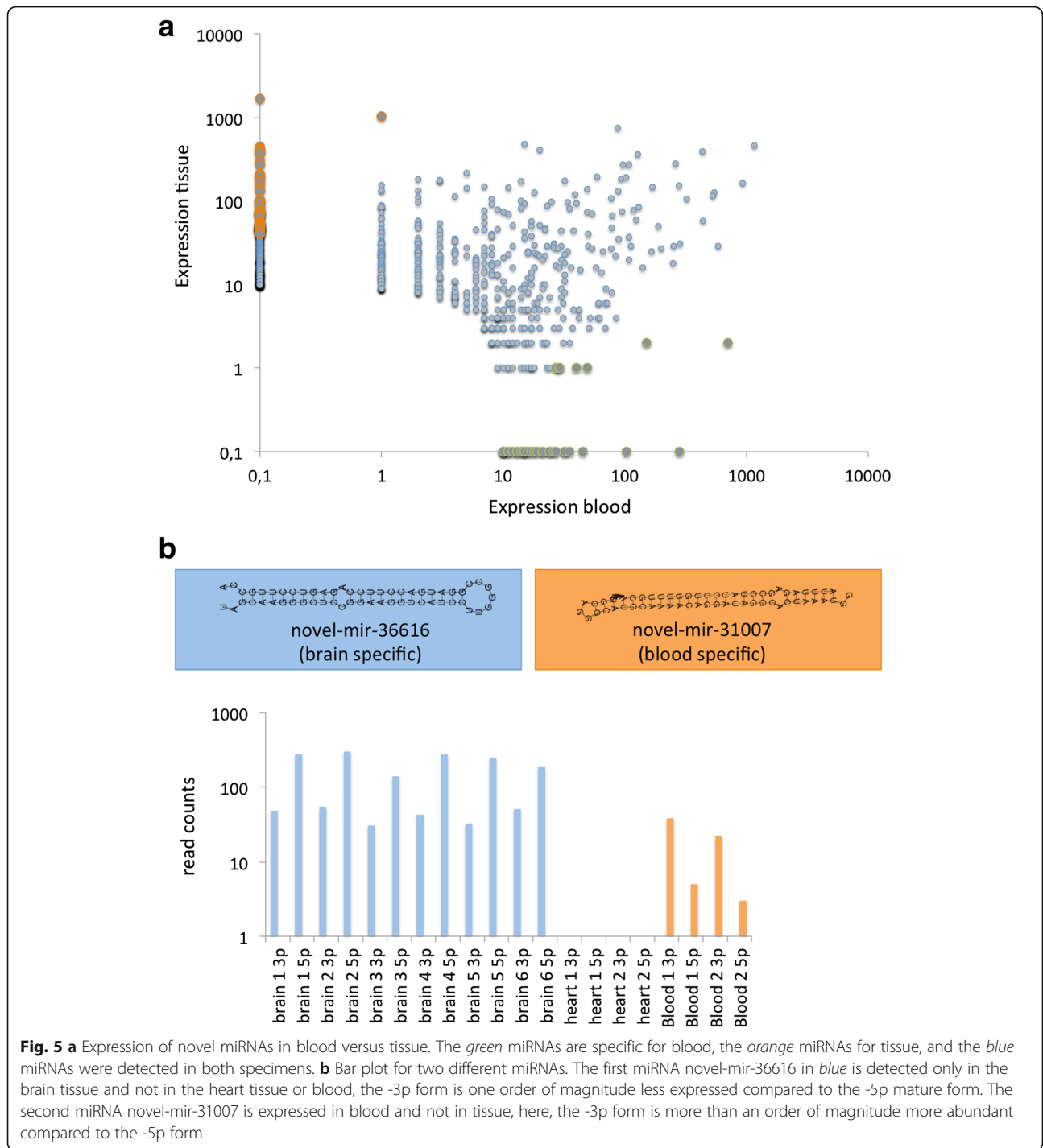


Further analysis with *novoMiRank* (cutoff 1.5) filtered out more miRNAs, leaving 1553. The miRNAs were flagged by *novoMiRank* because of a high deviation from miRNAs in the first *miRBase* versions, including deviating length, free energy, or nucleic acid composition of miRNAs. Matching the remaining candidates to other RNA resource in a blacklisting step finally presented 926 miRNA candidates (Additional file 4: Table S2). Still, it is likely that this set contains many false positives. Additionally, low-throughput experimental validation of almost 1000 miRNA candidates, e.g., by Northern Blot is a very labor-extensive approach. We thus additionally compared the frequency of reads mapping to the blood versus tissue samples. As detailed in Fig. 5a, we observe a substantial variability between blood and tissue for the 926 miRNA candidates (correlation 0.18). Defining a miRNA as tissue/blood specific if it occurs with a factor of 100-fold higher in one of both sample types (normalized for the total number of samples) highlighted 74 new miRNA candidates specific for tissue and 36 new miRNA candidates specific for blood samples. Figure 5b shows bar plots for two miRNA precursors, the most tissue specific novel-mir-36616 (blue), only present in the brain samples, and the blood specific novel-mir-31007. The first miRNA, which is observed exclusively in the brain samples and not in the heart, reveals a significantly

less expressed 3' mature form as compared to the 5' mature form. The second miRNA is exclusively observed in blood samples. Here, the 5' mature form is lower expressed compared to the 3' form. The boxes above the bar plots show the secondary structures of both miRNA candidates.

miRNA target analysis

For all 926 miRNAs, we predicted targets using TargetScan. To rank miRNA-target interactions, we used the context++ score (distribution of the context++ score across all predictions is provided in Additional file 5: Figure S2). Thereby, we observed an accumulation of high-likelihood targets for tissue-specific miRNAs. Of the 926 miRNAs, the tissue specific had an average 42.8 targets, the neither for blood nor for tissue-specific miRNAs 40.7 targets while for blood-specific miRNAs, only 34.5 targets were predicted. The complex miRNA-target network is presented in Additional file 6: Figure S3. It contains 6014 nodes (5088 genes and 926 miRNAs). Network characteristics such as degree distribution and shortest path length are presented in Additional file 7: Figure S4. The genes with largest numbers of predicted miRNAs targeting the gene were CYB561D1 (229 miRNAs), FBXL12 (174 miRNAs), PML (162 miRNAs), and VNN3 (154 miRNAs). The distribution of miRNAs in



the different group is presented as Venn diagram in Additional file 8: Figure S5). Among the predicted target genes that were found only for candidate miRNAs being blood specific was, e.g., HMOX1, heme oxygenase 1, mediating the first step of the heme catabolism by cleaving heme to build biliverdin or HPX, coding for hemopexin. The complex nature of the in silico calculated miRNA-target network requires further analyses to

understand whether target genes accumulate in specific biochemical categories such as KEGG pathways or gene ontologies. We thus applied GeneTrail2 separately to the set of genes targeted by blood specific miRNAs, targeted by tissue specific miRNAs and by all other miRNAs. As the background sets, all genes predicted to be targeted by at least a single miRNA were selected and the functionality to compare different enrichment analyses by

GeneTrail2 has been used. Enriched pathways seem to be largely relevant for either blood or tissue miRNAs, as Additional file 9: Figure S6 highlights. Tissue specific miRNAs had target genes enriched for DNA damage response, the apoptosis, or RNA polymerase II regulatory region DNA binding while blood miRNAs target genes were, e.g., enriched for TP53 network. Interestingly, tissue miRNA target genes also clustered on specific genomic locations (e.g., 19p12 and 19.q13) while blood miRNA targets did not show such an enrichment. In contrast, blood miRNA targets were enriched for disease phenotypes such as carotid artery diseases. In sum, the enrichment analysis highlights very distinct patterns for blood and tissue miRNA targets. Of course, not only the new miRNAs themselves but also the predicted targets deserve detailed experimental validation.

Discussion

The advent of next-generation sequencing reduced the costs of sequencing while simultaneously increasing the speed of throughput [28]. Today, the costs for small RNA seq are almost equal to and even lower than miRNA microarrays, although small RNA-seq provides the additional possibility for detecting novel small RNA entities.

In the present study, we investigated two current sequencing approaches supporting massively parallel sequencing, which is of high relevance in small RNA research because of the high dynamic range of these molecules: DNA nanoball [11]-based sequencing by BGISEQ-500 and PCR cluster [8]-based sequencing by HiSeq. An important difference between these techniques is in that the first approach uses linear DNA amplification, and the second uses exponential DNA amplification to make sequencing arrays. The latter approach may in turn lead to amplification errors and some specific biases. Besides this fundamental difference, both approaches have their additional advantages and disadvantages. Specifically for the BGISEQ-500, the library preparation currently takes around three working days, the sequencing itself needs one or at maximum two working days. Each flowcell of the BGISEQ-500 has two lanes. On each of these lanes, 32 Gb data can be generated using single-end reads of length 50 bases. The cost of the reagent and material is around 200 USD for 20 million reads ensuring high-quality data at a reasonable cost.

Recently, we published a manuscript about bias in NGS and microarray analysis for miRNAs [6], highlighting that the expression of miRNAs on different platforms varies by, for example, the nucleic acid composition. In the validation by RT-qPCR, we focused on miRNAs discordant between the high-throughput platforms. Thereby, we observed cases where the RT-qPCR results were concordant with Illumina HiSeq, with

microarrays or with none of the techniques. Therefore, we were especially interested how the BGISEQ-500 platform compares to the HiSeq platform and microarrays with the content from the *miRBase* for small RNA analysis.

Three miRNAs had high divergence between arrays and BGISEQ-500, among them hsa-miR-4454, which was high abundant in arrays but almost not detectable in BGISEQ-500. According to the miRBase, only 28% of users believe that this miRNA is real. Although such votes have only limited value, they at least indicate that this miRNA may be influenced by technological bias.

For high-throughput sequencing, the library preparation and the kits used play a crucial role for the quality of the sequencing results. Others and we noticed an overly abundance of the miRNA miR-486-5p when using the TruSeq kit (Illumina, San Diego), which seems to be independent of the source of the analyzed material [6, 29, 30]. Using the BGISEQ-500 platform, we observed lower read counts for this miRNA. However, in some cases, the miRNA abundance of BGISEQ-500 matches to the HiSeq sequencing results while microarrays show a different expression level, and in other cases, the BGISEQ-500 deviates from the other platforms and in several cases, all three techniques provide substantially divergent results. The more even distribution of reads of the BGISEQ-500 compared to the HiSeq results facilitates the discovery of new miRNAs, which are expected to be significantly less expressed as compared to the already known miRNAs, especially from early miR-Base versions.

With respect to many miRNA currently annotated in miRBase and the rapidly growing number of new miRNAs, it is essential not only to have tools for filtering likely false-positives such as the NovoMiRank tool but also to carry out validation of miRNAs using other molecular biology approaches such as cloning and Northern blotting.

Focusing on the performance of the BGISEQ-500, we found a high technical reproducibility of sequencing results, which was however slightly below the technical reproducibility of microarrays. This fact can have different reasons, e.g., the different limit of detection of microarrays. In contrast to sequencing, microarrays have a saturation effect. With respect to the total number of discovered known miRNAs, performance of the BGISEQ-500 was comparable both to the Illumina and the microarray platform.

Conclusions

In sum, none of the mentioned platforms seems to provide the “ultimate solution” in miRNA analysis. All have their advantages and disadvantages and show some bias for the detection of certain sequence types.

Additional files

Additional file 1: Table S3. miRNA read count of the BGISEQ-500. (XLSX 250 kb)

Additional file 2: Figure S1. Predicted secondary structures for selected miRNAs. (PNG 241 kb)

Additional file 3: Table S1. Comparison of BGISEQ-500 to Agilent. (XLSX 135 kb)

Additional file 4: Table S2. List of novel miRNA candidates. (XLSX 6531 kb)

Additional file 5: Figure S2. Histogram of the decade logarithm of the context++ scores (multiplied by -1) of predicted targets for the candidate miRNAs. Since negative context++ scores are favorable, the miRNA targets on the right of the diagram are more likely true interactions. (PNG 78 kb)

Additional file 6: Figure S3. Full interaction network. Predicted miRNAs are represented in large nodes, colored by type (red: blood specific, blue: tissue specific, green: all others) and genes are represented by smaller gray nodes. (PNG 1033 kb)

Additional file 7: Figure S4. Core network characteristics as node degree distribution (*top*) and shortest path length (*bottom*). (PNG 129 kb)

Additional file 8: Figure S5. Venn diagram showing the distribution of predicted target genes for tissue-specific miRNA candidates, blood-specific miRNA candidates, and all other miRNA candidates. (PNG 156 kb)

Additional file 9: Figure S6. Comparison of the pathway enrichment analysis for the GeneTrail2 analysis with respect to the three target sets. *Red arrows* represent significant enrichments. (PNG 289 kb)

Acknowledgements

We acknowledge the support of BGI-Shenzhen and Complete Genomics.

Funding

The study has been funded by internal funds of Saarland University.

Availability of data and materials

Following publication expression data are available in the gene expression omnibus (GEO).

Authors' contributions

Setting up the assay were done by CG, XS, AA, SD, CZ, DA, JL, and RD. Generating miRNA data were done by SR, CZ, NL, MH, ZZ, CX, AC, and MN. Evaluation of data was done by TF, CB, NL, YL, and AK. Drafting and revision of the manuscript were done by EM, AK. Study design and set-up were done by YL, CS, XX, EM, and AK. All authors read and approved the final manuscript.

Competing interests

Authors with affiliations 1 and 2 are employed by BGI-Shenzhen, Shenzhen, China, and Complete Genomics (a BGI company), Mountain View, CA, USA.

Consent for publication

Not applicable.

Ethics approval and consent to participate

The study has been approved by the local ethics committee (Ärztchamber des Saarlandes).

Author details

¹Clinical Bioinformatics, Saarland University, 66125 Saarbrücken, Germany. ²BGI-Shenzhen, Shenzhen, China. ³Department of Human Genetics, Saarland University, Saarbrücken, Germany. ⁴Complete Genomics (a BGI company), Mountain View, CA, USA.

Received: 6 October 2016 Accepted: 4 November 2016

Published online: 21 November 2016

References

- Veneziano D, Nigita G, Ferro A. Computational approaches for the analysis of ncRNA through deep sequencing techniques. *Front Bioeng Biotechnol.* 2015;3:77.
- Lee RC, Feinbaum RL, Ambros V. The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell.* 1993;75(5):843–54.
- Ruvkun G. Molecular biology. Glimpses of a tiny RNA world. *Science.* 2001;294(5543):797–9.
- Mestdagh P, Hartmann N, Baeriswyl L, Andreasen D, Bernard N, Chen C, Cheo D, D'Andrade P, DeMayo M, Dennis L, et al. Evaluation of quantitative miRNA expression platforms in the microRNA quality control (miRQC) study. *Nat Methods.* 2014;11(8):809–15.
- Hafner M, Renwick N, Brown M, Mihailovic A, Holoch D, Lin C, Pena JT, Nusbaum JD, Morozov P, Ludwig J, et al. RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. *RNA.* 2011;17(9):1697–712.
- Backes C, Sedaghat-Hamedani F, Frese K, Hart M, Ludwig N, Meder B, Meese E, Keller A. Bias in high-throughput analysis of miRNAs and implications for biomarker studies. *Anal Chem.* 2016;88(4):2088–95.
- Friedlander MR, Chen W, Adamidi C, Maaskola J, Einspanier R, Knespel S, Rajewsky N. Discovering microRNAs from deep sequencing data using miRDeep. *Nat Biotechnol.* 2008;26(4):407–15.
- Mayer P, Farinelli L, Kawashima EHUhwgcpUS. Method of nucleic acid amplification. In.: Google Patents; 2011
- Drmanc R, Crkvenjakov R. Prospects for a miniaturized, simplified and frugal human genome project. *Sci Yugosl.* 1990;16(1–2):97–107.
- Keller A, Backes C, Leidinger P, Kefer N, Boisguerin V, Barbacioru C, Vogel B, Matzas M, Huwer H, Katus HA, et al. Next-generation sequencing identifies novel microRNAs in peripheral blood of lung cancer patients. *Mol BioSyst.* 2011;7(12):3187–99.
- Drmanc R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, Yeung G, et al. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science.* 2010;327(5961):78–81.
- Backes C, Meder B, Hart M, Ludwig N, Leidinger P, Vogel B, Galata V, Roth P, Menegatti J, Grasser F, et al. Prioritizing and selecting likely novel miRNAs from NGS data. *Nucleic Acids Res.* 2016;44(6):e53.
- Canard B, Sarfati RS. DNA polymerase fluorescent substrates with reversible 3'-tags. *Gene.* 1994;148(1):1–6.
- Tsien RY, Ross P, Fahnestock M, Johnston AJUhwgcpCAAce. Dna sequencing. In.: Google Patents; 1991
- Church GM, Mitra RDUhwgcpEPace. Nucleotide compounds having a cleavable linker. In.: Google Patents; 2003
- Meder B, Backes C, Haas J, Leidinger P, Stahler C, Grossmann T, Vogel B, Frese K, Giannitsis E, Katus HA, et al. Influence of the confounding factors age and sex on microRNA profiles from peripheral blood. *Clin Chem.* 2014;60(9):1200–8.
- Leidinger P, Backes C, Deutscher S, Schmitt K, Mueller SC, Frese K, Haas J, Ruprecht K, Paul F, Stahler C, et al. A blood based 12-miRNA signature of Alzheimer disease patients. *Genome Biol.* 2013;14(7):R78.
- Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 2009;10(3):R25.
- Friedlander MR, Mackowiak SD, Li N, Chen W, Rajewsky N. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res.* 2012;40(1):37–52.
- Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.* 2006;34(Database issue):D140–4.
- Zhao Y, Li H, Fang S, Kang Y, Wu W, Hao Y, Li Z, Bu D, Sun N, Zhang MQ, et al. NONCODE 2016: an informative and valuable data source of long non-coding RNAs. *Nucleic Acids Res.* 2016;44(D1):D203–8.
- Hamberg M, Backes C, Fehlmann T, Hart M, Meder B, Meese E, Keller A. MiRTargetLink—miRNAs, genes and interaction networks. *Int J Mol Sci.* 2016;17(4):564.
- Stockel D, Kehl T, Trampert P, Schneider L, Backes C, Ludwig N, Gerasch A, Kaufmann M, Gessler M, Graf N, et al. Multi-omics enrichment analysis using the GeneTrail2 web service. *Bioinformatics.* 2016;32(10):1502–8.
- Keller A, Leidinger P, Vogel B, Backes C, ElSharawy A, Galata V, Mueller SC, Marquart S, Schrauder MG, Strick R, et al. miRNAs can be generally associated with human pathologies as exemplified for miR-144. *BMC Med.* 2014;12:224.

25. Keller A, Leidinger P, Bauer A, Elsharawy A, Haas J, Backes C, Wendschlag A, Giese N, Tjaden C, Ott K, et al. Toward the blood-borne miRNome of human diseases. *Nat Methods*. 2011;8(10):841–3.
26. Keller A, Backes C, Haas J, Leidinger P, Maetzler W, Deuschle C, Berg D, Ruschil C, Galata V, Ruprecht K, et al. Validating Alzheimer's disease micro RNAs using next-generation sequencing. *Alzheimers Dement*. 2016;12(5):565–76.
27. Backes C, Leidinger P, Altmann G, Wuerstle M, Meder B, Galata V, Mueller SC, Sickert D, Stahler C, Meese E, et al. Influence of next-generation sequencing and storage conditions on miRNA patterns generated from PAXgene blood. *Anal Chem*. 2015;87(17):8910–6.
28. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. *Trends Genet*. 2014;30(9):418–26.
29. Huang X, Yuan T, Tschannen M, Sun Z, Jacob H, Du M, Liang M, Dittmar RL, Liu Y, Liang M, et al. Characterization of human plasma-derived exosomal RNAs by deep sequencing. *BMC Genomics*. 2013;14:319.
30. Burgos KL, Javaherian A, Bompreszi R, Ghaffari L, Rhodes S, Courtright A, Tembe W, Kim S, Metpally R, Van Keuren-Jensen K. Identification of extracellular miRNA in human cerebrospinal fluid by next-generation sequencing. *RNA*. 2013;19(5):712–22.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

