Bioinformatics, 34(10), 2018, 1621–1628
doi: 10.1093/bioinformatics/btx814
Advance Access Publication Date: 21 December 2017
Original Paper

OXFORD

Genome analysis

A high-resolution map of the human small non-coding transcriptome

Tobias Fehlmann[1,†], Christina Backes[1,*,†], Julia Alles[2], Ulrike Fischer[2], Martin Hart[2], Fabian Kern[1], Hilde Langseth[3], Trine Rounge[3], Sinan Ugur Umu[3], Mustafa Kahraman[1,4], Thomas Laufer[4], Jan Haas[5,6,7], Cord Staehler[1], Nicole Ludwig[2], Matthias Hübenthal[8], Benjamin Meder[5,6,7], Andre Franke[8], Hans-Peter Lenhof[9], Eckart Meese[2] and Andreas Keller[1,*]

[1]Chair for Clinical Bioinformatics, Saarland University, 66123 Saarbrücken, Germany, [2]Department of Human Genetics, Saarland University, 66421 Homburg, Germany, [3]Cancer Registry of Norway, Institute of Population-based Cancer Research, N-0304 Oslo, Norway, [4]Hummingbird Diagnostics GmbH, 69120 Heidelberg, Germany, [5]Department of Internal Medicine III, University Hospital Heidelberg, 69120 Heidelberg, Germany, [6]German Center for Cardiovascular Research (DZHK), 69120 Heidelberg, Germany, [7]Klaus Tschira Institute for Integrative Computational Cardiology, 69120 Heidelberg, Germany, [8]Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, 24105 Kiel, Germany and [9]Center for Bioinformatics, Saarland University, 66123 Saarbrücken, Germany

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Inanc Birol

Received on September 22, 2017; revised on December 8, 2017; editorial decision on December 14, 2017; accepted on December 20, 2017

Abstract

Motivation: Although the amount of small non-coding RNA-sequencing data is continuously increasing, it is still unclear to which extent small RNAs are represented in the human genome.
Results: In this study we analyzed 303 billion sequencing reads from nearly 25 000 datasets to answer this question. We determined that 0.8% of the human genome are reliably covered by 874 123 regions with an average length of 31 nt. On the basis of these regions, we found that among the known small non-coding RNA classes, microRNAs were the most prevalent. In subsequent steps, we characterized variations of miRNAs and performed a staged validation of 11 877 candidate miRNAs. Of these, many were actually expressed and significantly dysregulated in lung cancer. Selected candidates were finally validated by northern blots. Although isolated miRNAs could still be present in the human genome, our presented set likely contains the largest fraction of human miRNAs.
Contact: c.backes@mx.uni-saarland.de or andreas.keller@ccb.uni-saarland.de
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The part of the human genome which is transcribed into RNAs but does not encode for proteins contains many elements with regulatory function that are central for physiological and pathophysiological processes. Generally, non-coding RNAs can be divided into long- (lncRNAs) and small non-coding RNAs (sncRNAs). Among the many different categories in the latter class are tRNAs, snoRNAs, miRNAs, snRNAs, piRNAs and others (Kowalczyk *et al.*, 2012). Altogether, over 85 000 human transcripts

not coding for proteins have been annotated so far (Harrow *et al.*, 2006). To discover further non-coding transcripts and to detect correlations to pathologies various specimens have been deeply sequenced and evaluated by a heterogeneous set of bioinformatics tools (Akhtar *et al.*, 2016). Many studies list large collections of putatively novel sequences that are frequently validated either to a limited extent or not at all (Backes *et al.*, 2016; Friedländer *et al.*, 2011; Londin *et al.*, 2015). The proposed candidates do not only consist of new representatives of the respective molecule class, but also consist of artifacts that are often also deposited in central databases. In a recent study, Vitsios *et al.* (2016) reanalyzed hundreds of small RNA-sequencing samples and revealed that many miRNAs in miRBase (Kozomara and Griffiths-Jones, 2014) are potentially mis-annotated. Another challenge arises from the substantial redundancies between the different studies. Since not all candidates are stored in a central repository, it is likely that a 'new' candidate has already been discovered in the same or very similar manner by others in previous studies. The heterogeneity and magnitude of small RNA-sequencing studies calls for a sophisticated meta-analysis of the available data. We have carried out such a meta-analysis and present our results in this paper including a high-resolution map of sncRNAs with a focus on miRNAs by an integrative analysis of thousands of small RNA-sequencing datasets.

2 Materials and methods

2.1 Sample collection

Our sample collection stems from three different sources. First, we downloaded sequencing data likely to contain small RNAs from the sequence read archive (SRA) (Kodama *et al.*, 2012) using the following query:

> ("small rna" OR srna OR mirna OR microrna) AND "Homo sapiens"[orgn:__txid9606]

This query resulted in 18 367 SRA Runs, of which we kept only those sequenced using an Illumina platform, with single-end reads, public access and assay type annotated as miRNA-, ncRNA- or RNA-seq. This resulted in 10 233 Runs. Since multiple runs can be performed for the same experiment, we merged them, leading to 8985 samples. We determined for these the presence or absence of 5' barcodes and 3' adapter, as described in the section below.

Second, we collected 10 999 miRNA-seq samples from the cancer genome atlas project (TCGA) (http://cancergenome.nih.gov/) (accessed on April 7, 2017). Since the raw files are only available as mapped BAM files, we transformed them back into FASTQ format for subsequent analysis.

As third data source, we used the collapsed reads users uploaded in our tool miRMaster (Fehlmann *et al.*, 2017), if they gave us consent for usage of their data in an aggregated manner, which summed up to 4570 samples. These different data sources provided in total 24 554 samples.

More information about the pre-processing of the samples can be found in the Supplementary Material.

2.2 Isoform analysis

For miRNAs different types of isoforms have been described in the literature (Guo and Chen, 2014). Most common are isoforms that only differ in the length of the sequence, but non-template additions at the 5' or 3' end have also been detected (Guo *et al.*, 2014). Variants within the miRNA sequence could also potentially stem from either sequencing errors or 'normal' genetic variability. To capture these different modification types, we performed our analysis as follows: For each human miRNA annotated in the miRBase (Kozomara and Griffiths-Jones, 2014) we mapped the reads to the respective precursor, while allowing up to two non-template additions to the 5' and 3' ends and up to one mismatch in between. We ensured that all precursors have at least 15 bases flanks on both ends so that no isoforms are missed. We then defined the mature 3' and 5' form covered by the largest fraction of reads per million mapped to miRNA (RPMMM) normalized reads as the canonical form. Mappings were allowed in a window of 10 bases up- and downstream of the annotated miRNAs. To avoid a bias through potentially mis-annotated miRNAs, we required that at least 80% of normalized reads mapped with at most one mismatch in an up- and downstream window of 2 and respectively 5 bases to the annotated miRNAs, or in-between. If only one miRNA was annotated, the other was derived from the read with the highest RPMMM, shorter than 25 bases and with at least 3 bases distance to the annotated miRNA. We determined the fraction of reads mapping to potential isoforms from the determined canonical forms, while allowing up to 5 bases variability at the 5' end and 7 bases at the 3' end. Isoforms that were longer than 25 bases or shorter than 17 bases were discarded. If a potential isoform was covered by at least 2% of all RPMMM normalized reads the isoform was considered to be present.

2.3 Single nucleotide variants in mature miRNAs

Single nucleotide variants were detected for each previously determined canonical miRNA form (see above) by mapping the reads against the respective precursor while allowing up to one mismatch. We allowed a variability of two nucleotides at the 5' end and five nucleotides at the 3' end.

2.4 Prediction of novel miRNAs

For the prediction of novel miRNAs, we used our web-based tool miRMaster (Fehlmann *et al.*, 2017). For each of the 18 035 samples we performed the prediction of novel miRNAs. Afterwards, the predicted precursors were merged and all reads aligned to the potential new precursors. From these mappings the candidate miRNAs were derived. To exclude candidates overlapping with already known ncRNAs, we mapped the predicted miRNAs to the human non-coding RNAs of Ensembl (release 85) (Yates *et al.*, 2016) and to NONCODE 2016 (Zhao *et al.*, 2016) using BLAST+ (Camacho *et al.*, 2009). We considered a candidate as not novel when the candidate miRNA sequence had an overlap of at least 90% with the aligned sequences while allowing at most one mismatch. All precursors containing mapping (i.e. non-novel) miRNAs were then discarded. To further asses the likelihood that miRNAs are true positives and to rank miRNA candidates we subsequently applied the novoMiRank tool (Backes *et al.*, 2016). This tool compares features of new miRNAs to a set of high-confidence miRNAs from early miRBase versions and ranks those miRNA candidates highest that match best to the known high-confidence miRNAs.

2.5 Validation of novel miRNAs

It is generally known that high-throughput approaches can lead to artifacts. This also holds for miRNA analysis where a substantial bias depending on the underlying measurement approach is known (Backes *et al.*, 2016). Since even different sequencing approaches can have different bias (Fehlmann *et al.*, 2016), we decided to validate new miRNA candidates by another technique. Since we wanted to avoid PCR bias, we decided in favor of amplification free

microarrays. Specifically, we collected a set of miRNAs from miRBase, a set of miRNA candidates from two other studies (Backes *et al.*, 2016; Londin *et al.*, 2015) and those that achieved a high rank in the above mentioned novoMiRank analysis. Altogether, 11 877 sequences were used for the custom microarray analysis. Previous studies (e.g. Ludwig *et al.*, 2016) suggested reasonable results with 20 technical replicates per miRNA, thus 237 540 features were required. Since the used Agilent microarrays do not provide sufficient feature numbers per array, the features were split across five arrays. Each of these arrays contains a fifth of the features, so that we get the full feature set when combining the different expression data from five arrays for one sample. A set of RNA samples (plasma and PAXGene blood pools, a reference sample and brain, kidney, liver, testis and heart tissue) was hybridized with the microarrays as previously described for standard miRBase v21 microarrays (Hecksteden *et al.*, 2016).

From all (candidate) miRNAs found expressed in the blood we generated a new custom microarray entitled 'all human blood miRNA array'. This array contains 2305 new and known miRNAs that are found in human blood. The microarray is manufactured by Agilent, handling is facilitated by standard Agilent equipment and the microarray is available for research use from Hummingbird Diagnostics. This array was hybridized with 53 patients, 25 controls and 28 small-cell lung carcinoma patients (SCLC). Microarray data were evaluated according to manufacturer's instructions and data were processed using R. *P*-values were determined using a two-tailed *t*-test and the Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995) was used to adjust for multiple testing.

3 Results

3.1 Dataset collection and processing

The important first step in the meta-analysis is the collection, quality control, and curation of available small RNA-sequencing datasets. As major sources for sequencing data we included the SRA (Kodama *et al.*, 2012), the cancer genome atlas project (http://cancergenome.nih.gov/) (Cancer Genome Atlas Research, 2008), and data that have been analyzed through our miRMaster workbench (Fehlmann *et al.*, 2017). The initial sample collection covered 24 554 individual NGS samples and 303 billion sequencing reads covering together 13 009 billion bases (corresponding roughly to the number of nucleic acids in 4200 human genomes). In a stringent quality control step, we removed samples having <1 million reads, samples without known sequencing adapters, samples not mapping to the human genome or those mapping with over 1% to coding exons. Further, samples covering >1% of the human genome were excluded, since high-quality small ncRNA-seq datasets usually cover a much smaller fraction of the human genome. More details about the quality filtering process can be found in the Methods section. Following this stringent QC process, 27% of all samples were removed for quality issues, leaving 18 035 samples containing 162 billion reads. The list of samples from SRA and TCGA is provided in Supplementary Table S1 and online at https://mircarta.cs.uni-saarland.de/data_sources/. As a matter of fact, a substantial percentage of all small RNA-sequencing reads is identical—collapsing the set of all reads leaves 1.9 billion unique reads. Mapping them to the human genome and excluding those that match to more than five unique positions in order to avoid unspecific reads, we found hits for 602 million unique reads in the human genome. A detailed breakdown of the numbers in the different steps is presented in Figure 1A. Next, the 18 035 samples were annotated with respect to their tissue of origin to create a map representing all datasets. This was achieved by applying t-distributed stochastic neighbor embedding. The respective map—color coded by the tissue of origin—highlights that in the majority of cases tissues of the same type build dense clusters (see Fig. 1B).

3.2 Coverage of the human genome

Given the large collection of small RNA-sequencing reads, a natural first question to ask is how densely the human genome is covered by the sncRNA-sequencing reads. Mapping the 602 million unique reads without allowing mismatches we calculate a total (1-fold) coverage of 64%. Increasing the coverage threshold leads to a rapid drop of the covered part of the genome. For example, we observe a decrease from 13 to 6% when considering the 10- and 20-fold coverage of the human genome by unique reads. Detailed analysis immediately highlights that large fractions of the human genome are covered only by single samples (16.3%) or few samples (41.1% are represented by at most five of the 18 035 samples). Although such regions covered by very few samples could still contain true non-coding RNAs, we excluded these regions because our aim was to present a collection of reliable regions present in a reasonable number of samples containing potentially new sncRNAs. We thus calculated the fraction of the genome covered by at least n samples and at least m reads. The diagonal of the matrix, the percentage covered by at least n reads in at least n samples, is presented in Figure 2A. For $n = m = 20$, the fraction already decreases to 7.5%, for $n = m =$ 50–3% and for $n = m =$ 100–1.5%. Heatmaps for the complete matrix are shown in Supplementary Figure S1. For determining regions that are reliably covered by small sequencing reads, we set the lower coverage threshold to 180, corresponding to 1% of the 18 035 samples. We denote this part of the genome that consists of 0.8% of the genome covered by 25 million bases as 'high confidence' or 'reliable' regions. The length distribution of the roughly 900 000 different regions is presented in Figure 2B as histogram. On top of the histogram, the length distribution of known and annotated non-coding RNAs is presented. The peak of the regions is at a length of 22 nucleotides, fitting best to annotated human miRNAs. Especially in this range, many reliable regions fall into not-annotated regions of the human genome. However, also longer stretches exceeding 150 bases were detected among the reliable regions. These usually match to lncRNAs or even genes. The high confidence regions that represent one of the central results of our work are provided as GFF3 file in Supplementary Material S1.

3.3 The most prevalent RNA species

In a next step, we asked how many percent of the considered RNA classes are covered by reliable regions on each chromosome (Supplementary Fig. S2). The reliable regions contain on average per chromosome 81% of miRNAs that are annotated in miRBase v21. Best concordance was computed for Chromosome 14 (91%). Since the data may not only contain miRNA-sequencing reads, but potentially also other non-coding RNAs or mRNA contaminations, we calculated the fraction of these classes covered by reliable regions. Only 0.5% of all piRNAs and 0.9% of all lncRNAs matched to reliable regions. For coding exons, the number however increased to 20%. Although this distribution in principle shows that our quality filtering successfully extracted samples primarily containing miRNA data, we also observed many regions corresponding to protein-coding genes. These can potentially also contain non-coding elements; however, contamination due to fragmented mRNA is more likely. In the following, we focus on miRNAs as molecule class and highlight

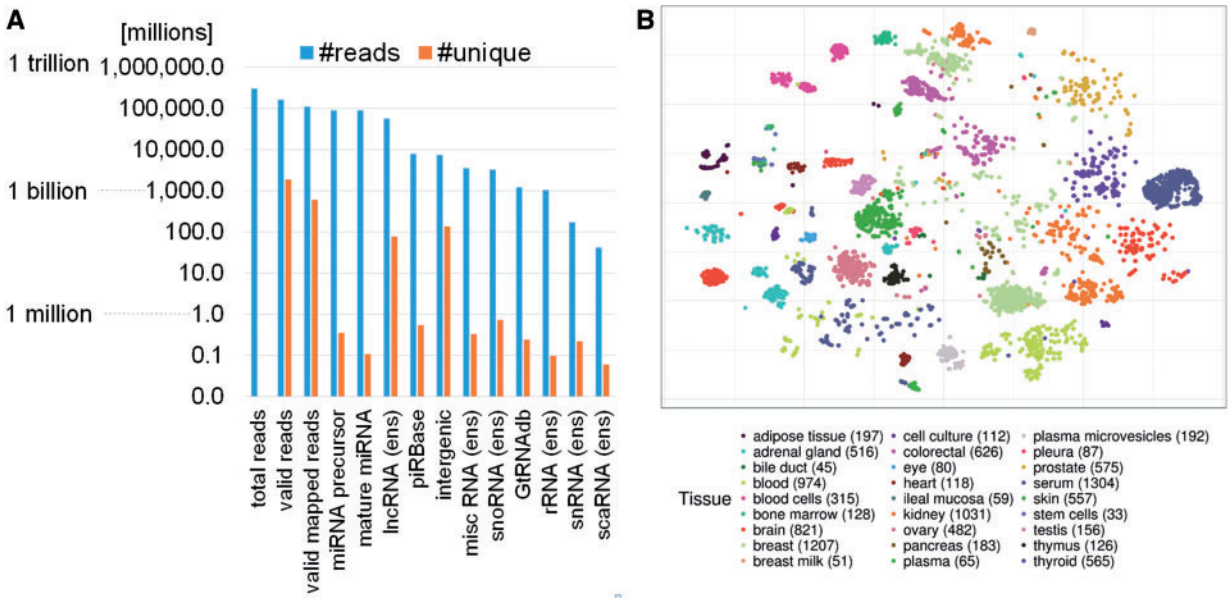

Fig. 1. Overview of the sample collection. Panel **(A)** shows a detailed breakdown of the reads during our quality filtering process into different categories. We determined for the remaining 18 035 samples to which different genomic annotations they map (using zero mismatches and allowing at least five mappings in the genome per read). Panel **(B)** shows a t-SNE map for samples with known tissue of origin. We computed the pairwise distances of these samples using the software Mash and plotted a visual representation with the Rtsne package. For almost all tissues distinct clusters can be discovered. Some clusters contain outlying samples from other tissues that could partially be due to wrong annotations in the original datasets

interesting findings about potential isoforms, variations and novel miRNA predictions.

3.4 Variability in mature miRNAs

Frequently, miRNAs show variability in their length, start and end position, which can influence their regulatory function. In addition to the canonical form, for each mature miRNA several dozens of so-called isomiRs can exist. Importantly, the mature miRNAs as annotated in the miRBase (considered as the canonical form) represented in <43% of cases the most expressed mature miRNA across our 18 035 samples. In particular, we observed shifts at the 5' end in 23% of cases, thereby influencing the expected miRNA seed region. This effect is due to the fact that many canonical miRNAs in the miRBase are derived only from few samples usually only in one tissue type or cell fraction while we aim to identify the overall most abundant mature form of a miRNA. We thus set the mature form with highest read count to be the canonical form for each miRNA and calculated the variation frequency for respective isoforms. To reduce the influence of potentially mis-annotated precursors on our analysis we considered only precursors that passed a basic signature check (details are provided in Section 2), leaving 1415 precursors out of the original 1881 extracted from miRBase v21. The determined canonical forms are presented in Supplementary Table S2. As presented in Figure 3A, we observed a substantial variability in the length distribution of miRNAs. Comparing the 3' and the 5' mature forms of all miRNAs suggests that the 3' forms have a slightly increased variability. Most significantly, for both, the 3' and 5' mature forms the frequency of variations at the 3' end were significantly higher as compared to the frequency of variation at the 5' end. A factor that may compromise the considerations is the depth of coverage per miRNA. For high-abundant miRNAs, the likelihood to discover an isoform is higher as compared to low abundant miRNAs. Therefore, we used a threshold relative to the expression of the miRNA and considered an isoform as detected if at least 2% of all RPMMMs matched to the variant. We observed 15 437 isoforms for 2258 miRNA/precursor pairs, thus corresponding to seven isoforms on average. An example of a miRNA with only two isoforms is hsa-miR-153-5p (Fig. 3B). The canonical form of this miRNA was covered by 92% of all RPMMM, while the most abundant isoform—two bases longer at the 3' end—was found in 3% of reads. Although this miRNA is an example with a clearly most abundant mature form, the canonical form as annotated in miRBase v21 was one base shorter at the 5' end and two bases longer at the 3' end. An example of a miRNA with many detected isoforms, in total 14 isoforms, is hsa-miR-330-3p (Fig. 3C). The corresponding canonical form is represented only by 19% of all RPMMM, while its second and third most expressed isoforms are represented by 14–15% of all RPMMM. Again, the annotation in miRBase does not match the canonical form in our study—it is one base longer at the 3' end. A complete list of the 15 437 isoforms for human miRNAs for which we detected a canonical form with the relative RPMMM frequencies and absolute number of reads mapping to this isoform is available in Supplementary Table S3.

3.5 Single nucleotide variants in mature miRNAs

Variability in miRNAs is not limited to the length of the mature miRNAs. Single point variants in miRNAs, especially in the seed region, can significantly influence miRNA-target gene interactions. Therefore, we investigated the mutation frequency of all miRNAs considered in the previous sub-section. As shown in Figure 4A, mutations at the 3' end are the most frequent ones, similar to the isoforms that are also mostly observed at the 3' end. The most frequent variations are adenylation and uridylations. This is in agreement with previous findings where post-transcriptional adenylation by GLD-2 (Katoh *et al.*, 2009) and uridylation by TUTases (Heo *et al.*, 2012) were reported. Interestingly, also a wide variability of modifications at the 5' end was observed. However, the relative frequency of the different alterations was usually low. One example with a high frequency is hsa-miR-376a-2-5p for which we observed the sequence with an A->G mutation at the third base in 48% of the

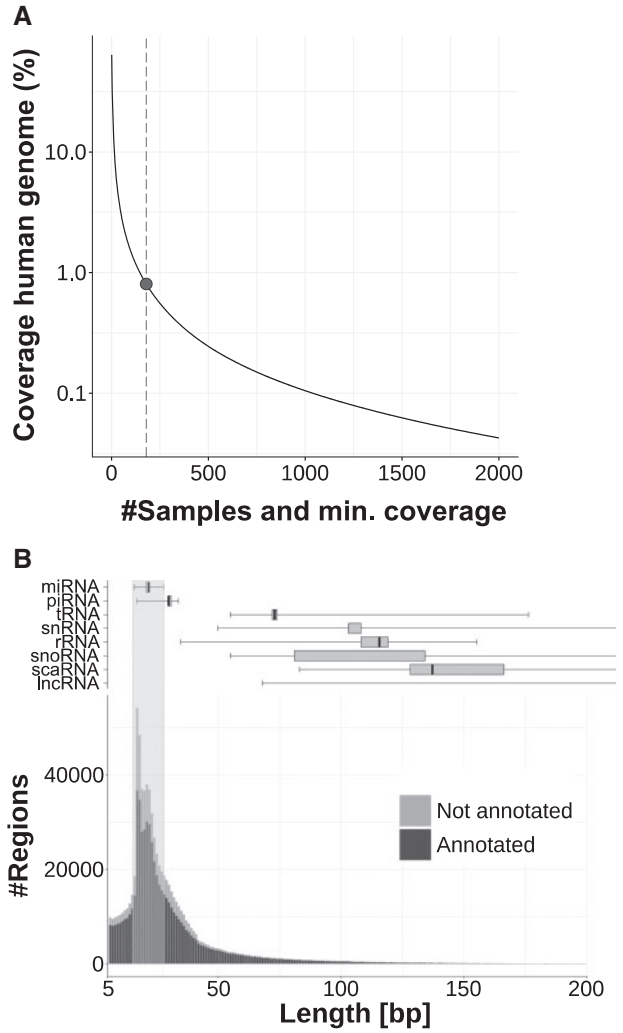

Fig. 2. Panel **(A)** Distribution of coverage/samples against percentage of human genome covered. With 1% of all samples, we cover almost 1% of the human genome (represented by the dot). Panel **(B)** Length distribution of the high confidence regions. There is a clear peak at 18–22 nt, which falls in the known length distribution of mature miRNAs (shaded region). In the peak region we see an enriched fraction of regions where no annotation is known yet. We also added the length distribution of selected other RNA entities on top of the plot for completeness. However, only piRNAs and miRNAs have lengths distributions that match the observed peak

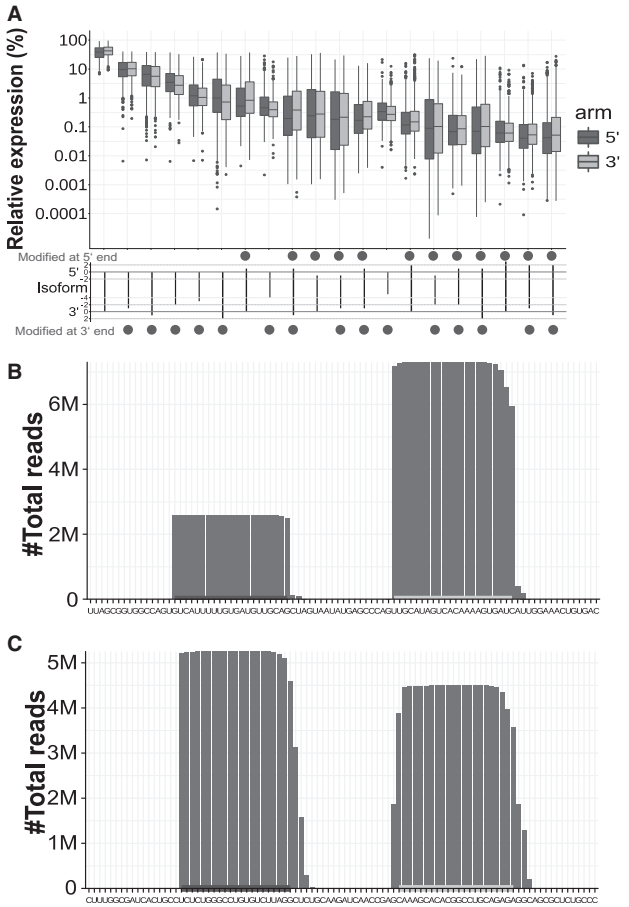

Fig. 3. Panel **(A)** shows the frequency of canonical and iso-miRNAs. For each human miRNA we calculated the canonical form as the form with highest read stack. Then we likewise calculated the isoforms. The analyses were performed for 3′ and 5′ mature miRNAs separately. Below the box plots, the respective isoform is presented schematically. The horizontal black lines mark the canonical form. The dots above the isoform flag those with 5′ modifications, those below the isoforms flag those with 3′ modifications. The isoforms are sorted with respect to decreasing median frequency and only the top 20 are shown. Higher abundant miRNA isoforms are dominated by modifications at the 3′ end. Panels **(B)** and **(C)** represent examples for miRNAs with few isoforms (panel B) and many isoforms (panel C)

normalized reads of this miRNA. Interestingly, no other sequence in the human genome matched the mutated sequence (with up to one mismatch), minimizing the possibility of a false detection. This mutation could be the result of an adenosine to inosine RNA editing, since the inosine is reported as guanine during sequencing. We also detected the same A->G mutation for the other arm, hsa-miR-376a-3p, at the sixth base. Here, the variant was observed even in 67% of the total reads. This variability is not limited to a certain cell type or tissue: the miRNA was detected with at least 10 reads per sample in a total of 5744 samples. The high frequency of these modifications and their expression across a large number of samples suggest that the RNA editing of this miRNA is important in a wide range of diseases or tissue contexts, confirming and extending previous findings by other researchers (Kawahara *et al.*, 2007; Zheng *et al.*, 2016). When comparing the relative frequencies on the level of mature arms, we can see that the most abundant modifications at the 3' end seem to be slightly more frequent for the 3' arm than for the 5' arm. For the other modifications there seems to be no observable bias. Figure 4B shows the mean relative expression of the observed mutations per position for all known miRNAs of miRBase v21. A list containing all detected mismatches including their absolute number of reads and relative RPMMM frequency is provided in Supplementary Table S4.

3.6 Discovery of new miRNAs

As described earlier and highlighted in Figures 1A and 2B, not all reads match to known RNA resources. Respective regions that are covered but are not annotated yet contain potentially novel miRNAs. We applied the algorithms implemented in miRMaster (Fehlmann *et al.*, 2017) to predict new miRNA candidates. From its basic principles the approach is similar to miRDeep2 (Friedländer *et al.*, 2011). The very large number of samples however required a completely re-implemented and optimized version in C++ to facilitate the joint analysis of the billions of reads in reasonable computing time. From all sequencing reads, we predicted a total of 135 290 new miRNA candidates. It is expected that these contain many false

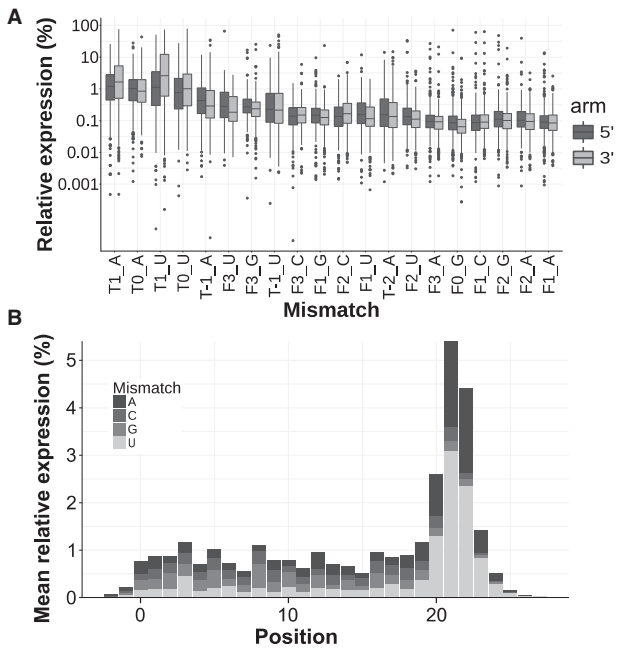

Fig. 4. Panel **(A)** details the frequency of the different variants in the 3′ and 5′ mature form of the miRNAs. The variants are shown relative to the 3′ (T) or 5′ (F) end of the miRNAs. The variants are sorted according to decreasing median frequency and only the top 20 are shown. Panel **(B)** shows the frequency of different single nucleotide variants across the position in the miRNA

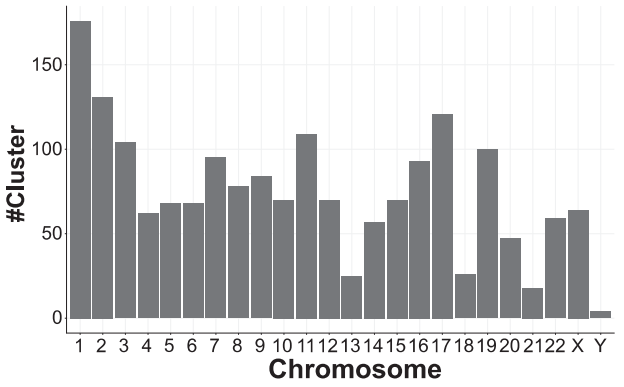

Fig. 5. Frequency of known and predicted precursor clusters across the different human chromosomes. A cluster was defined as a region containing at least two precursors with a distance of at most 10 kb between the middle positions of at least two members

positives. Thus, in a first step to reduce the number of potential precursors and to increase the precision (the fraction of true positive prediction on all positive predictions), we matched the predicted precursors to the previously annotated high confidence regions. This analysis yielded 17 400 miRNA candidates overlapping with our reliable regions that are not equal or similar to known ncRNAs. These candidates are available as GFF3 in Supplementary Material S2 and are stored in our new online repository miRCarta (http://www.ccb.uni-saarland.de/mircarta) (Backes *et al.*, 2017). MiRCarta (v1.0) is a comprehensive database that allows for browsing and filtering the miRNA candidates from this article, the candidates from the publications of Backes *et al.* (2016) and Londin *et al.* (2015), as well as the latest miRBase data. In addition, it visualizes the expression data from the 18 035 samples as pileup plots for the stem-loops to facilitate the assessment if the expression profile belongs to a true positive finding. To have a consistent naming of the novel candidates and known miRNAs we implemented a new scheme. This scheme, which is detailed in the Methods section, is also suited to unify future findings and is valid across species. In brief, identical mature miRNAs have the same identifier independent of the species (e.g. m-17) and the precursors are named by an organism tag (similar to miRBase) followed by the 5' and 3' mature miRNAs they consist of (e.g. hsa-17-22 contains the 5' mature miRNA m-17 and the 3' mature miRNA m-22). Although this new naming scheme is the primary identifier in miRCarta, the well-known miRNA name and miRBase ID is contained for all currently available miRNAs.

3.7 Genomic miRNA clusters

Frequently, miRNA genes are not isolated and uniformly distributed across the genome but accumulate in clusters. Thus, we searched for such clusters in the human genome for the set of known and predicted miRNAs. A cluster was defined as a region containing at least two precursors with a distance of at most 10 kb between the middle positions of at least two members. Altogether, 1802 clusters with an average of 2.35 precursors were observed. The chromosomes with the largest number were the largest chromosomes: Chromosome 1 (176 clusters) and Chromosome 2 (131 clusters). In contrast, Chromosome 21 (18 clusters) and Chromosome 13 (25 clusters) had significantly fewer representatives (Fig. 5). This figure shows that the clusters are not uniformly distributed across the genome and only partially reflect the chromosome sizes. Although Chromosome 4 with ∼190 million bases contains 62 clusters (1 precursor per ∼3 million bases), and Chromosome 13 (∼110 million bases) contains 25 precursors (1 precursor per ∼4.4 million bases), especially Chromosomes 17 (∼80 million bases, 121 clusters, 1 cluster per ∼0.66 million bases) and 19 are enriched for miRNA clusters (∼60 million bases, 100 clusters, 1 cluster per ∼0.8 million bases). The full list of clusters with regions, and the number of known as well as predicted miRNAs is presented in Supplementary Table S5.

3.8 Validation of new miRNAs

As described earlier miRNA candidates can only be considered real miRNAs once they have been experimentally validated. Among the core criteria for validating a miRNA is to report expression by using a hybridization-based technique. We consider the detection of processed mature forms from cloned precursors via Northern Blotting as gold standard for experimental validation. However, due to the large number of candidates we performed a first validation step via a custom microarray. Since the predicted miRNAs could be influenced by technological bias, e.g. the library preparation or sequencing, we selected a subset of 11 877 miRNAs containing the annotated miRNAs from miRBase v21 and further candidates, with the focus on blood miRNA candidates. Using these candidates, we built a custom microarray and hybridized the arrays with plasma and PAXGene blood pools, a reference sample, brain, kidney, liver, testis and heart tissue samples. By this amplification free procedure, we measured signals for 1146 (44%) known miRNAs and 3151 (34%) miRNA candidates. The results of the microarray analysis are presented as heat map in Figure 6A. The pileup plot and secondary structure of one novel predicted precursor is shown exemplarily in Figure 6B. Another example of a predicted precursor is shown in Supplementary Figure S3. This precursor is already annotated in two other species (oan-mir-2985 and tgu-mir-2985-1), but not yet as human miRNA precursor.

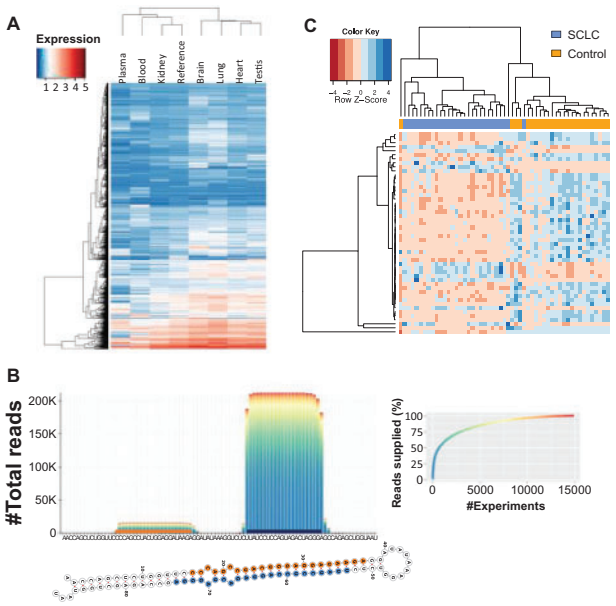

Fig. 6. Panel **(A)** describes the first-pass validation of human (candidate) miRNAs by microarrays. Cluster heat map of all 4297 mature (candidate) miRNAs that showed signals in high-quality RNA samples based on amplification free hybridization. Panel **(B)** presents a representative example with the mapping distribution as presented in Figure 3. In addition to the pileup plot, the secondary structure is shown. Panel **(C)** shows the result of the human blood miRNA array hybridized with lung cancer (blue) and control samples (orange). Results are presented as heat map resulting from hierarchical clustering with dendrograms on top (clustering of samples) and at the left side (clustering of miRNAs)

Among the specimens with the most complex miRNomes were blood samples. Since they can be gathered minimally invasive, blood samples represent also an important source of new biomarkers for various human pathologies. Many studies have shown that known miRNAs can be differentially expressed comparing disease and control samples. In an additional validation step, we therefore asked whether we can also observe expression differences for novel blood miRNA candidates. To this end, we built a microarray containing 2305 (candidate) miRNAs from the first validation that were expressed in blood. The resulting human blood miRNA microarray can be used to facilitate the discovery and validation of circulating miRNAs for many human pathologies. Among the diseases with the largest effect size and the highest reproducibility in miRNA biomarkers is lung cancer. With the new array, we hybridized 53 individuals, 25 controls and 28 SCLC patients. Following adjustment for multiple testing, 695 miRNAs had an adjusted *t*-test *P*-value below 0.05 and were considered statistically significant. Although the six most significant features are known from the miRBase, already the seventh marker was a new candidate miRNA (raw and adjusted *P*-value $10{-}^{9}$ and $10{-}^{6}$). Altogether, 457 candidate miRNAs that are not included in the miRBase were among the 695 significant markers for SCLC. As the cluster heat map in Figure 6C details, the most significant markers were predominantly down-regulated in SCLC patients. These results suggest that new miRNAs are not only detectable by hybridization-based techniques but also bear a substantial diagnostic information content.

Of course, the hybridization on arrays does not replace a thorough analysis of the expression and biogenesis of miRNAs followed by Northern Blotting. We thus cloned the precursors of miRNAs from different miRBase versions and new candidates and tried to detect the processed/mature forms of miRNAs on Northern Blots. For 59 of 103 tested candidates that have previously not been reported by Northern Blots, bands in the expected size of around 22 nucleotides were observed (manuscript in preparation). This set is to our knowledge the largest collection of miRNAs that have jointly been validated using respective experimental methods. The precursor presented in Figure 6B was among the ones for which both miRNAs was validated.

3.9 Evolutionary conservation of new miRNAs

MiRNAs are frequently highly conserved. To verify if we can also observe this for our novel candidates, we mapped the sequences of the candidates without mismatches against the genomes of 148 organisms that are also contained in miRCarta and counted a hit for an organism if we could find the sequence at least one time in its genome. We find about 85% of our candidates in at least one other organism than *Homo sapiens*. A novel candidate occurs on average in 4.5 organisms. The sequences of the top five miRNA candidates (sorted by the sum of hits in different organisms) can even be found on average in 37 species. The miRNA candidate with the most hits (m-7214) was detected in 58/148 organisms. These 58 organisms are from various taxonomy classes such as *Mammalia*, *Insecta*, *Amphibia* etc. For the remaining top five candidates, the variety in species is smaller such that we can find the lowest common ancestor *Craniata* for these taxa at subphylum level. Furthermore, some of our candidates are already annotated in miRBase for other organisms than human. For example, m-3155 corresponds to the known miRNAs mmu-miR-3085-3p and rno-miR-3085. Of course, these findings only illustrate that identical sequences are contained in other organisms, not that they necessarily function as miRNAs.

4 Discussion and conclusions

Since the advent of next-generation sequencing, thousands of small RNA-sequencing datasets have been created and also been partially deposited in public databases. However, depending on which pipelines were used for the evaluation of the datasets in the different study setups, these are not directly comparable to each other. To make maximal use of the available datasets, a consistent analysis of the different samples with the same pipeline is necessary. The aim of our study was to remove redundancies due to different study setups and to provide a reliable map of high-quality small RNA annotations, particularly for miRNAs.

Starting with a collection of 24 554 human small RNA NGS samples, we performed stringent quality controls, which left us with 18 035 usable samples for down-stream analysis.

Depending on the coverage thresholds, up to 64% of the genome are covered (1-fold coverage). However, this number does not reflect the true complexity of the human non-coding transcriptome but is affected by different sources of bias. By defining those regions that are covered by at least 1% (180) of all samples as reliable, we still obtained about 900 000 such regions with variable lengths. Although they contain true positive sncRNAs, we still expect a reasonable number of false positive hits. Since solid low-throughput validation of such large sets of potential non-coding miRNAs is not feasible, we performed a large first pass validation by using microarrays. We were able to detect expression signals for 34% of our novel candidates in high-quality samples, minimizing the risk of false positives e.g. by degradation of mRNAs. Further, we found that some of these miRNAs have considerable potential as biomarkers in SCLC. Still, these are not necessarily functional miRNAs but remain

candidates until a detailed validation has been carried out. Thus, we performed such a validation for selected candidates using Northern Blotting.

With our sncRNA study, we performed to our knowledge the most complete analysis of human sncRNAs with a focus on miRNAs. The set of reported reliable regions, which is covering 0.8% of the human genome, likely contains a very substantial fraction of all small non-coding elements in the human genome.

Funding

The different aspects of this work have been supported by the following sources: The generation of the own datasets was supported by Siemens Healthineers. The *in silico* analysis was supported by the EU FP7 project BestAgeing and the Michael J. Fox Foundation. The validation of new miRNAs by microarrays was supported by Hummingbird Diagnostics. The validation of new miRNAs by Northern Blots was supported by the Michael J. Fox Foundation.

Conflict of Interest: none declared.

References

Akhtar,M.M. *et al*. (2016) Bioinformatic tools for microRNA dissection. *Nucleic Acids Res*., **44**, 24–44.

Backes,C. *et al*. (2017) miRCarta: a central repository for collecting miRNA candidates. *Nucleic Acids Res*., doi: 10.1093/nar/gkx851.

Backes,C. *et al*. (2016) Prioritizing and selecting likely novel miRNAs from NGS data. *Nucleic Acids Res*., **44**, e53.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. Ser. B Methodol*., **57**, 289–300.

Camacho,C. *et al*. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Cancer Genome Atlas Research, N. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061–1068.

Fehlmann,T. *et al*. (2017) Web-based NGS data analysis using miRMaster: a large-scale meta-analysis of human miRNAs. *Nucleic Acids Res*., **45**, 8731–8744.

Fehlmann,T. *et al*. (2016) cPAS-based sequencing on the BGISEQ-500 to explore small non-coding RNAs. *Clin. Epigenet*., **8**, 123.

Friedländer,M.R. *et al*. (2011) miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res*., **40**, 37–52.

Guo,L. and Chen,F. (2014) A challenge for miRNA: multiple isomiRs in miRNAomics. *Gene*, **544**, 1–7.

Guo,L. *et al*. (2014) A genome-wide screen for non-template nucleotides and isomiR repertoires in miRNAs indicates dynamic and versatile microRNAome. *Mol. Biol. Rep*., **41**, 6649–6658.

Harrow,J. *et al*. (2006) GENCODE: producing a reference annotation for ENCODE. *Genome Biol*., **7** (Suppl 1), S4 1–S9.

Hecksteden,A. *et al*. (2016) miRNAs and sports: tracking training status and potentially confounding diagnoses. *J. Transl. Med*., **14**, 219.

Heo,I. *et al*. (2012) Mono-uridylation of pre-microRNA as a key step in the biogenesis of group II let-7 microRNAs. *Cell*, **151**, 521–532.

Katoh,T. *et al*. (2009) Selective stabilization of mammalian microRNAs by 3' adenylation mediated by the cytoplasmic poly(A) polymerase GLD-2. *Genes Dev*., **23**, 433–438.

Kawahara,Y. *et al*. (2007) Redirection of silencing targets by adenosine-to-inosine editing of miRNAs. *Science*, **315**, 1137–1140.

Kodama,Y. *et al*. (2012) The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res*., **40**, D54–D56.

Kowalczyk,M.S. *et al*. (2012) Molecular biology: RNA discrimination. *Nature*, **482**, 310–311.

Kozomara,A. and Griffiths-Jones,S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res*., **42**, D68–D73.

Londin,E. *et al*. (2015) Analysis of 13 cell types reveals evidence for the expression of numerous novel primate- and tissue-specific microRNAs. *Proc. Natl. Acad. Sci. USA*, **112**, E1106–E1115.

Ludwig,N. *et al*. (2016) Distribution of miRNA expression across human tissues. *Nucleic Acids Res*., **44**, 3865–3877.

Vitsios,D.M. *et al*. (2016) Large-scale analysis of microRNA expression, epi-transcriptomic features and biogenesis. *Nucleic Acids Res*., **45**, 1079–1090.

Yates,A. *et al*. (2016) Ensembl 2016. *Nucleic Acids Res*., **44**, D710–D716.

Zhao,Y. *et al*. (2016) NONCODE 2016: an informative and valuable data source of long non-coding RNAs. *Nucleic Acids Res*., **44**, D203–D208.

Zheng,Y. *et al*. (2016) Accurate detection for a wide range of mutation and editing sites of microRNAs from small RNA high-throughput sequencing profiles. *Nucleic Acids Res*., **44**, e123.